

Kohei Arai
Rahul Bhatia *Editors*

Advances in Information and Communication

Proceedings of the 2019 Future of
Information and Communication
Conference (FICC), Volume 2

Lecture Notes in Networks and Systems

Volume 70

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Systems Research Institute,
Warsaw, Poland

e-mail: kacprzyk@ibspan.waw.pl

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Advisory Board

Fernando Gomide, Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil

e-mail: gomide@dca.fee.unicamp.br

Okyay Kaynak, Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Turkey

e-mail: okyay.kaynak@boun.edu.tr

Derong Liu, Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA and Institute of Automation, Chinese Academy of Sciences, Beijing, China

e-mail: derong@uic.edu

Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada and Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

e-mail: wpedrycz@ualberta.ca

Marios M. Polycarpou, KIOS Research Center for Intelligent Systems and Networks, Department of Electrical and Computer Engineering, University of Cyprus, Nicosia, Cyprus

e-mail: mpolycar@ucy.ac.cy

Imre J. Rudas, Óbuda University, Budapest Hungary

e-mail: rudas@uni-obuda.hu

Jun Wang, Department of Computer Science, City University of Hong Kong Kowloon, Hong Kong

e-mail: jwang.cs@cityu.edu.hk

More information about this series at <http://www.springer.com/series/15179>

Kohei Arai · Rahul Bhatia
Editors

Advances in Information and Communication

Proceedings of the 2019 Future of Information
and Communication Conference (FICC),
Volume 2



Springer

Editors

Kohei Arai
Faculty of Science and Engineering
Saga University
Saga, Japan

Rahul Bhatia
The Science and Information
(SAI) Organization
Bradford, UK

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-3-030-12384-0

ISBN 978-3-030-12385-7 (eBook)

<https://doi.org/10.1007/978-3-030-12385-7>

Library of Congress Control Number: 2018968383

© Springer Nature Switzerland AG 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

After the success of Future of Information and Communication Conference (FICC) 2018, FICC 2019 is held on March 14–15, 2014 in San Francisco, USA.

The Future of Information and Communication Conference (FICC), 2019 focuses on bringing together experts from both industry and academia, to exchange research findings in the frontier areas of Communication and Computing. This conference delivers programs of latest research contributions and future vision (inspired by the issues of the day) in the field and potential impact across industries. It features an innovative format for presenting new research, focussing on participation and conversation rather than passive listening.

FICC 2019 attracted a total of 462 submissions from many academic pioneering researchers, scientists, industrial engineers, and students from all around the world. These submissions underwent a double-blind peer review process. Of those 462 submissions, 160 submissions (including 15 poster papers) have been selected to be included in this proceedings. It covers several hot topics which include Ambient Intelligence, Intelligent Systems, Data Science, Machine Learning, Internet of Things, Networking, Security, and Privacy. This conference showcases paper presentations of new research, demos of new technologies, and poster presentations of late-breaking research results, along with inspiring keynote speakers and moderated challenge sessions for participants to explore and respond to big challenge questions about the role of technology in creating thriving, sustainable communities.

Many thanks goes to the Keynote Speakers for sharing their knowledge and expertise with us and to all the authors who have spent the time and effort to contribute significantly to this conference. We are also indebted to the organizing committee for their great efforts in ensuring the successful implementation of the conference. In particular, we would like to thank the technical committee for their constructive and enlightening reviews on the manuscripts in the limited timescale.

We hope that all the participants and the interested readers benefit scientifically from this book and find it stimulating in the process. See you in next SAI Conference, with the same amplitude, focus, and determination.

Saga, Japan

Regards,
Kohei Arai

Contents

Contextual Binding: A Deductive Apparatus in Artificial Neural Networks	1
Jim Q. Chen	
Intelligent Signal Classifier for Brain Epileptic EEG Based on Decision Tree, Multilayer Perceptron and Over-Sampling Approach	11
Jimmy Ming-Tai Wu, Meng-Hsiun Tsai, Chia-Te Hsu, Hsien-Chung Huang, and Hsiang-Chun Chen	
Korean-Optimized Word Representations for Out-of-Vocabulary Problems Caused by Misspelling Using Sub-character Information	25
Seonhghyun Kim, Jai-Eun Kim, Seokhyun Hawang, Berlocher Ivan, and Seung-Won Yang	
A Regressive Convolution Neural Network and Support Vector Regression Model for Electricity Consumption Forecasting	33
Youshan Zhang and Qi Li	
Facial Expression Recognition and Analysis of Interclass False Positives Using CNN	46
Junaid Baber, Maheen Bakhtyar, Kafil Uddin Ahmed, Waheed Noor, Varsha Devi, and Abdul Sammad	
CASCADENET: An LSTM Based Deep Learning Model for Automated ICD-10 Coding	55
Sheikh Shams Azam, Manoj Raju, Venkatesh Pagidimarri, and Vamsi Chandra Kasivajjala	
Automated Gland Segmentation Leading to Cancer Detection for Colorectal Biopsy Images	75
Syed Fawad Hussain Naqvi, Salahuddin Ayubi, Ammara Nasim, and Zeeshan Zafar	

A Two-Fold Machine Learning Approach for Efficient Day-Ahead Load Prediction at Hourly Granularity for NYC	84
Syed Shahbaaz Ahmed, Raghavendran Thiruvengadam, A. S. Shashank Karrthikeyaa, and Vineeth Vijayaraghavan	
A Classical-Quantum Hybrid Approach for Unsupervised Probabilistic Machine Learning	98
Prasanna Date, Catherine Schuman, Robert Patton, and Thomas Potok	
Comparing TensorFlow Deep Learning Performance and Experiences Using CPUs via Local PCs and Cloud Solutions	118
Robert Nardelli, Zachary Dall, and Sotiris Skevoulis	
An Efficient Segmentation Technique for Urdu Optical Character Recognizer (OCR)	131
Saud Ahmed Malik, Muazzam Maqsood, Farhan Aadil, and Muhammad Fahad Khan	
Adaptive Packet Routing on Communication Networks Based on Reinforcement Learning	142
Tanyaluk Deeka, Boriboon Deeka, and Surajate On-rit	
ScaffoldNet: Detecting and Classifying Biomedical Polymer-Based Scaffolds via a Convolutional Neural Network	152
Darlington Ahiale Akogo and Xavier-Lewis Palmer	
Transfer Learning for Cross-Domain Sequence Tagging Tasks	162
Meng Cao, Chaohe Zhang, Dancheng Li, Qingping Zheng, and Ling Luo	
Ensemble Models for Enhancement of an Arabic Speech Emotion Recognition System	174
Rached Zantout, Samira Klaylat, Lama Hamandi, and Ziad Osman	
Automating Vehicles by Deep Reinforcement Learning Using Task Separation with Hill Climbing	188
Mogens Graf Plessen	
Information Augmentation, Reduction and Compression for Interpreting Multi-layered Neural Networks	211
Ryotaro Kamimura	
Enhance Rating Algorithm for Restaurants	224
Jeshreen Balraj and Cassim Farook	
Reverse Engineering Creativity into Interpretable Neural Networks	235
Marilena Oita	

Developing a Deep Learning Model to Implement Rosenblatt's Experiential Memory Brain Model	248
Abu Kamruzzaman, Yousef Alhwaiti, and Charles C. Tappert	
The Influence of Media Types on Students' Learning Performance in a Role Playing Game on English Vocabulary Learning	262
Yuting Li and Jing Leng	
Edit Distance Kernelization of NP Theorem Proving For Polynomial-Time Machine Learning of Proof Heuristics	271
David Windridge and Florian Kammüller	
Performance Analysis of Artificial Neural Networks Training Algorithms and Activation Functions in Day-Ahead Base, Intermediate, and Peak Load Forecasting	284
Lemuel Clark P. Velasco, Noel R. Estoperez, Renbert Jay R. Jayson, Caezar Johnlery T. Sabijon, and Verlyn C. Sayles	
Quantum Deep Learning Neural Networks	299
Abu Kamruzzaman, Yousef Alhwaiti, Avery Leider, and Charles C. Tappert	
Hierarchical Modeling for Strategy-Based Multi-agent Multi-team Systems	312
D. Michael Franklin	
Bioinformatics Tools for PacBio Sequenced Amplicon Data Pre-processing and Target Sequence Extraction	326
Zeeshan Ahmed, Justin Pranulis, Saman Zeeshan, and Chew Yee Ngan	
SimTee: An Automated Environment for Simulation and Analysis of Requirements	341
Saad Zafar, Musharif Ahmed, Taskeen Fatima, and Zohra Aslam	
SuperDense Coding Step by Step	357
Lewis Westfall and Avery Leider	
Artificial Swarming Shown to Amplify Accuracy of Group Decisions in Subjective Judgment Tasks	373
Gregg Willcox, Louis Rosenberg, David Askay, Lynn Metcalf, Erick Harris, and Colin Domnauer	
High-Resolution Streaming Functionality in SAGE2 Screen Sharing	384
Kazuya Ishida, Daiki Asao, Arata Endo, Yoshiyuki Kido, Susumu Date, and Shinji Shimojo	
A Secure Scalable Life-Long Learning Based on Multiagent Framework Using Cloud Computing	400
Ghalib Ahmad Tahir, Sundus Abrar, and Loo Chu Kiong	

An Autonomic Model-Driven Architecture to Support Runtime Adaptation in Swarm Behavior	422
Mark Allison, Melvin Robinson, and Grant Rusin	
Review of Paradigm Shift in Patent Within Digital Environment and Possible Implications for Economic Development in Africa	438
Stephen Odirachukwu Mwim and Tana Pistorius	
*Thing: Improve Anything to Anything Collaboration	453
Giancarlo Corti, Luca Ambrosini, Roberto Guidi, and Nicola Rizzo	
eSense 2.0: Modeling Multi-agent Biomimetic Predation with Multi-layered Reinforcement Learning	466
D. Michael Franklin and Derek Martin	
System of Organizational Terms as a Methodological Concept in Replacing Human Managers with Robots	479
Olaf Flak	
Interruption Timing Prediction via Prosodic Task Boundary Model for Human-Machine Teaming	501
Nia Peters	
Computer Science Education: Online Content Modules and Professional Development for Secondary Teachers in West Tennessee—A Case Study	523
Lee Allen	
Enterprises and Future Disruptive Technological Innovations: Exploring Blockchain Ledger Description Framework (BLDF) for the Design and Development of Blockchain Use Cases	533
Hock Chuan Lim	
Human Superposition Allows for Large-Scale Quantum Computing	541
Bruce Levinson	
Student User Experience with the IBM QISKit Quantum Computing Interface	547
Stephan Barabasi, James Barrera, Prashant Bhalani, Preeti Dalvi, Ryan Dimiecik, Avery Leider, John Mondrosch, Karl Peterson, Nimish Sawant, and Charles C. Tappert	
Searching for Network Modules	564
Giovanni Rossi	
Routing Sets and Hint-Based Routing	586
Ivan Avramovic	

Comparison between Maximal Independent Sets and Maximal Cliques Models to Calculate the Capacity of Multihop Wireless Networks	603
Maher Heal and Jingpeng Li	
A Priority and QoS-Aware Scheduler for LTE Based Public Safety Networks	616
Mahir Ayhan and Hyeong-Ah Choi	
Efficient Mobile Base Station Placement for First Responders in Public Safety Networks	634
Chen Shen, Mira Yun, Amrinder Arora, and Hyeong-Ah Choi	
Solutions of Partition Function-Based TU Games for Cooperative Communication Networking	645
Giovanni Rossi	
Interoperable Convergence of Storage, Networking, and Computation	667
Micah Beck, Terry Moore, Piotr Luszczek, and Anthony Danalis	
QoS for SDN-Based Fat-Tree Networks	691
Haitham Ghalwash and Chun-Hsi Huang	
LBMM: A Load Balancing Based Task Scheduling Algorithm for Cloud	706
Yong Shi and Kai Qian	
Exploiting Telnet Security Flaws in the Internet of Things	713
James Klein and Kristen R. Walcott	
Probabilistic Full Disclosure Attack on IoT Network Authentication Protocol	728
Madiha Khalid, Umar Mujahid, Muhammad Najam-ul-Islam, and Binh Tran	
Multilevel Data Concealing Technique Using Steganography and Visual Cryptography	739
Chaitra Rangaswamaiah, Yu Bai, and Yoonsuk Choi	
Ring Theoretic Key Exchange for Homomorphic Encryption	759
Jack Aiston	
Response-Based Cryptographic Methods with Ternary Physical Unclonable Functions	781
Bertrand Cambou, Christopher Philabaum, Duane Booher, and Donald A. Telesca	

Implementation of Insider Threat Detection System Using Honeypot Based Sensors and Threat Analytics	801
Muhammad Mudassar Yamin, Basel Katt, Kashif Sattar, and Maaz Bin Ahmad	
Subject Identification from Low-Density EEG-Recordings of Resting-States: A Study of Feature Extraction and Classification	830
Luis Alfredo Moctezuma and Marta Molinas	
Early Detection of Mirai-Like IoT Bots in Large-Scale Networks through Sub-sampled Packet Traffic Analysis	847
Ayush Kumar and Teng Joon Lim	
Desktop Browser Extension Security and Privacy Issues	868
Steven Ursell and Thaier Hayajneh	
Exploring Cybersecurity Metrics for Strategic Units: A Generic Framework for Future Work	881
Mohammad Arafah, Saad Haj Bakry, Reham Al-Dayel, and Osama Faheem	
From Access Control Models to Access Control Metamodels: A Survey	892
Nadine Kashmar, Mehdi Adda, and Mirna Atieh	
A Potential Cascading Succession of Cyber Electromagnetic Achilles' Heels in the Power Grid	912
S. Chan	
On the Relation Between Security Models for HB-like Symmetric Key Authentication Protocols	935
Miaomiao Zhang	
Development of Students' Security and Privacy Habits Scale	951
Naurin Farooq Khan and Naveed Ikram	
Privacy and Security—Limits of Personal Information to Minimize Loss of Privacy	964
Hanif Ur Rahman, Ateeq Ur Rehman, Shah Nazir, Izaz Ur Rehman, and Nizam Uddin	
Towards Protection Against a USB Device Whose Firmware Has Been Compromised or Turned as ‘BadUSB’	975
Usman Shafique and Shorahbeel Bin Zahur	
A Multi-dimensional Adversary Analysis of RSA and ECC in Blockchain Encryption	988
Sonali Chandel, Wenxuan Cao, Zijing Sun, Jiayi Yang, Bailu Zhang, and Tian-Yi Ni	

Contents	xiii
Assessing the Performance of a Biometric Mobile Application for Workdays Registration	1004
Cristian Zambrano-Vega, Byron Oviedo, and Oscar Moncayo Carreño	
Closer Look at Mobile Hybrid Apps Configurations: Statistics and Implications	1016
Abeer AlJarrah and Mohamed Shehab	
Security and Privacy Issues for Business Intelligence in IoT	1038
Mohan Krishna Kagita	
Securing a Network: How Effective Using Firewalls and VPNs Are?	1050
Sun Jingyao, Sonali Chandel, Yu Yunnan, Zang Jingji, and Zhang Zhipeng	
A Software Engineering Methodology for Developing Secure Obfuscated Software	1069
Carlos Gonzalez and Ernesto Liñan	
Detecting Windows Based Exploit Chains by Means of Event Correlation and Process Monitoring	1079
Muhammad Mudassar Yamiun, Basel Katt, and Vasileios Gkioulos	
Analysing Security Threats for Cyber-Physical Systems	1095
Shafiq ur Rehman, Manuel De Ceglia, and Volker Gruhn	
Bayesian Signaling Game Based Efficient Security Model for MANETs	1106
Rashidah Funke Olanrewaju, Burhan ul Islam Khan, Farhat Anwar, Roohie Naaz Mir, Mashkuri Yaacob, and Tehseen Mehraj	
End-to-End Emotion Recognition From Speech With Deep Frame Embeddings And Neutral Speech Handling	1123
Grigoriy Sterling and Eva Kazimirova	
Algorithm and Application Development for the Agents Group Formation in a Multi-agent System Using SPADE System	1136
Alexey Goryashchenko	
On Compressed Sensing Based Iterative Channel Estimator for UWA OFDM Systems	1144
Sumit Chakravarty and Ankita Pramanik	
Development on Interactive Route Guidance Robot for the Mobility Handicapped in Railway Station	1153
Tae-Hyung Lee, Jong-Gyu Hwang, Kyeong-Hee Kim, and Tae-Ki Ahn	
Facilitate External Sorting for Large-Scale Storage on Shingled Magnetic Recording Drives	1159
Yu-Pei Liang, Min-Hong Shen, Yi-Han Lien, and Wei-Kuan Shih	

LTE Geolocation Based on Measurement Reports and Timing Advance	1165
Zaenab Shakir, Josko Zec, and Ivica Kostanic	
Hybrid Parallel Approach of Splitting-Up Conjugate Gradient Method for Distributed Memory Multicomputers	1176
Akiyoshi Wakatani	
Exceeding the Performance of Two-Tier Fat-Tree: Equality Network Topology	1187
Chane-Yuan Yang, Chi-Hsiu Liang, Hong-Lin Wu, Chun-Ho Cheng, Chao-Chin Li, Chun-Ming Chen, Po-Lin Huang, and Chi-Chuan Hwang	
Energy Aware LEACH Protocol for WSNs	1200
Nahid Ebrahimi Majd, Pooja Chhabria, and Anjali Tomar	
Evaluation of Parameters Affecting the Performance of Routing Protocols in Mobile Ad Hoc Networks (MANETs) with a Focus on Energy Efficiency	1210
Nahid Ebrahimi Majd, Nam Ho, Thu Nguyen, and Jacob Stolmeier	
Integrating User Opinion in Decision Support Systems	1220
Saveli Goldberg, Gabriel Katz, Ben Weisburd, Alexander Belyaev, and Anatoly Temkin	
SEAKER: A Tool for Fast Digital Forensic Triage	1227
Eric Gentry, Ryan McIntyre, Michael Soltys, and Frank Lyu	
Cyber-Physical Network Mapping Attack Topology	1244
Glenn Fiedelholtz	
Author Index	1251



Contextual Binding: A Deductive Apparatus in Artificial Neural Networks

Jim Q. Chen^(✉)

National Defense University, Washington, DC, USA

Abstract. The Artificial neural networks are key mechanisms in artificial intelligence, especially in machine learning. A close examination of this mechanism reveals some philosophical challenges in the current approach. With all the emphasis on the techniques used in the mining of the great volume of datasets collected and available, inductive reasoning is well employed to find out characteristics and identify patterns. However, deductive reasoning is not sufficiently utilized. In order to bring the current approach to the right track, this paper proposes a novel approach in which a deductive apparatus is made use of. This deductive apparatus is built on contextual binding, which sets a priority and provides guidance for the processing of various types of datasets collected and available, thus making the processing efficient and effective. The benefits and implementation of this new approach are also discussed.

Keywords: Artificial neural networks · Machine learning · Inductive method · Deductive method · Contextual binding

1 Introduction

1.1 A Subsection Sample

Artificial neural networks are key mechanisms in artificial intelligence, especially in machine learning. It is the mechanisms that provide deep learning. As mentioned by Schmidhuber [1], deep learning artificial neural networks “have won numerous contests in pattern recognition and machine learning”. “Shallow and deep learners are distinguished by the depth of their credit assignment paths, which are chains of possibly learnable, causal links between actions and effects”. As observed by Klauer [2], inductive reasoning is widely utilized. “Researchers in the field of artificial intelligence have constructed computer programs based on process models that aim to solve certain kinds of problems to test their theories of inductive reasoning”. “Even sophisticated mathematical models have been developed and tested that are able to predict how people process inductive problems, for instance, causal models or Bayesian models”.

There is no doubt that the inner-workings of an artificial neural network plays an important role in machine learning as varied patterns need to be identified and intertwined relations need to be established. Meanwhile, it has to be pointed out that the way in which input data are collected, verified, and processed also has a significant effect. Inductive reasoning does play an important role. However, pure inductive

method does not always lead to an accurate and prompt answer, as in most cases data collected are incomplete and in some cases they are even corrupted or damaged. It is hard to imagine that an inductive reasoning mechanism can turn garbage input into a trustworthy output.

This paper explores a creative approach in which deductive reasoning is extensively utilized. With such a deductive apparatus, intelligent hints can be dynamically generated based on varied contexts, thus avoiding unnecessary searches or operations and quickly identifying targets. This approach has significant impact upon digital forensic investigation, especially accurate and prompt attribution or target hunting in cyber operations. To a certain extent, it can serve as a force multiplier with the speed and capability that it provides.

The paper is organized as follows: Here in Sect. 1, an overview of the environment is provided. Next in Sect. 2, the current approach in artificial neural networks is examined and the issues in it are analyzed. Then in Sect. 3, an innovative approach that employs deductive reasoning is proposed. In Sect. 4, the benefits and the implementation of this new approach are discussed. In Sect. 5, a conclusion is drawn.

2 Issues

In explaining a deep learning artificial neural network, Dormehl [3] provides a diagram that consists of the input layer, the hidden layers, and the output layer. This diagram is displayed in Fig. 1.

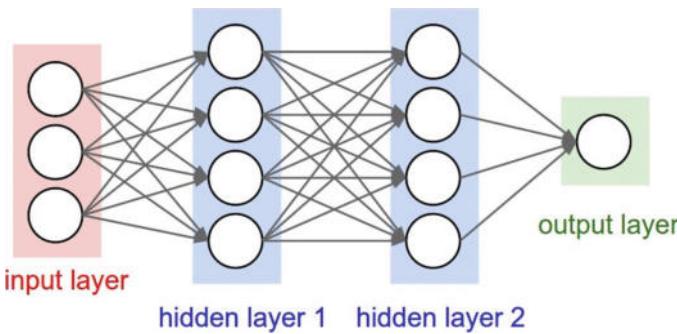


Fig. 1. Artificial neural network

In this diagram, deep learning occurs within the hidden layers, which may consist of more than two layers. An input dataset is examined by each subsequent node, which extracts “a different set of high-level features”. “By the time the fourth or fifth layer is reached, the deep learning net will have created complex feature detectors”, thus identifying an object. This mechanism works pretty well in identifying images. However, it may have some issues in conducting digital forensics. Should all the relevant pieces of evidence are captured in the input datasets, an accurate output can be generated with the help of the deep-learning layers, i.e. the hidden layers. Should some

piece of evidence are missing in the input datasets, an incomplete or even inaccurate output may be generated. This issue is directly related to the issues of inductive reasoning, as the conclusion is built on whatever inputs that it receives. This bottom-up approach is restrictive in nature as the datasets not captured in the input will certainly not be taken into consideration.

The Center for Academic Success at Butter College [4] describes inductive reasoning as a process that begins with specific and limited observations and ends with “a generalized conclusion that is likely, but not certain, in light of accumulated evidence”. “Much scientific research is carried out by the inductive method: gathering evidence, seeking patterns, and forming a hypothesis or theory to explain what is seen.” It is well pointed out that “conclusions reached by the inductive method are not logical necessities; no amount of inductive evidence guarantees the conclusion. This is because there is no way to know that all the possible evidence has been gathered, and that there exists no further bit of unobserved evidence that might invalidate my hypothesis.”

This methodological limitation may cause incomplete and inaccurate conclusion that may further lead to wrong decision-making, causing significant consequences. Further analysis clearly supports this argument.

2.1 Incompleteness of Input Datasets

The incompleteness exists both in depth and width.

Depth. In general, it is difficult to get all the datasets needed. In most cases, data collection is performed during a specific period of time. The data prior to that period of time might not be captured. The data that reveal the details might also not be captured. As a result, a decision has to be made based on the incomplete datasets. This may lead to a wrong direction for investigation. Let us look at the following hypothetical example. Assume the purpose of an investigation is to identify the individual who launched a cyber attack from the datasets collected. Assume that some pieces of data were not captured as they existed before the start of data collection. Further assume that the names captured are within the square bracket while the names not captured are outside the square bracket. It is possible that the following list is retained:

“Alan, Bill, [Charlie, Don, Ed, Frank, Gerald, Harry, Irene, Jack, Kevin, Leo, Mary, Nancy, Oliver, Peter]”

In this particular case, the system may fail to catch Bill if information about Bill failed to be captured or became corrupted. Without the whole datasets as inputs, the current approach is restricted in depth.

Width. In 1983, Gartner [5], an American developmental psychologist, came up with a theory of multiple intelligences. According to this theory, there are nine types of intelligence. They are: naturalist intelligence (nature smart), musical intelligence (sound smart), logical-mathematical intelligence (number/reasoning smart), existential intelligence (life smart), interpersonal intelligence (people smart), bodily-kinesthetic intelligence (body smart), linguistic intelligence (word smart), intra-personal intelligence (self smart), and spatial intelligence (picture smart). This theory clearly indicates that logical-mathematical intelligence is not the only intelligence and other types of intelligence have to be taken into account. However, the current approach in artificial

neural networks is only focused on logical-mathematical intelligence. Without examining other types of intelligence, the current approach is restricted in width. Likewise, Bradford [6] mentions the importance of having data from varied channels. He holds that “humans have five basic senses: touch, sight, hearing, smell and taste. The sensing organs associated with each sense send information to the brain to help us understand and perceive the world around us.” Obviously, having datasets from one dimension is not enough for analysis in some cases.

2.2 Inaccuracy of Output

It is hard to imagine that incomplete or damaged inputs may naturally generate accurate output. In most cases, it is garbage in and garbage out. Thus, the error rate goes up as the corresponding results are either inaccurate or totally wrong. As a consequence, the decision made via this mechanism can neither be trusted nor be put into action, rendering the output useless.

2.3 Issue Within the Hidden Layers

Within the hidden layers, the variables are unbound by any operators, thus making the process unguided and random to a certain extent. This lengthens the processing time and consumes a great amount of resources. In some cases, even doing this may not guarantee an accurate output. Let us use the summation operator for forward propagation as an example.

$$\text{netinput} = b + \sum_{i=1}^n (X_i * W_i) \quad (1)$$

As explained by Taylor [7], the “b” represents the input from a bias node. The “i”, the index of summation, begins with the first input node “1” and ends on the input node “n”. The “n” represents the total number of input nodes. The “ X_i ” represents a unique node. The “ W_i ” represents a unique weight situated on the nodes edge.

As it is not bound by any operator, the variable “b” can be substituted by any element in the process. If an irrelevant element gets injected into “b”, the result of the calculation becomes inaccurate, wasting the precious processing time. In the process of targeting during a cyber operation, a delay in accurate and prompt attribution may result in a failure of the cyber operation.

The analysis above clearly shows the inadequacy of inductive reasoning should it be used alone in artificial neural networks. Besides, it shows a severe consequence of having variables unbound in the process. To address these issues, an innovative approach is proposed. It is able to “help investigators get to relevant data more quickly, reduce the noise investigators must wade through, and help transform data into information and investigative knowledge”, as requested by Beebe [8]. It is also able to figure out a solution that is “near optimal (with high probability)”, as expected by Marti and Reinelt [9].

3 Proposal

The innovative proposal should address at least two issues. One is inductive reasoning. The other is unbound variables. The solution in dealing with inductive reasoning is the use of deductive reasoning in deep learning artificial neural networks. Fortunately, contextual binding provides such a function. The solution in dealing with unbound variables is the employment of contextual binding, which may serve as intelligent hints in the deep learning process.

The Center for Academic Success at Butler College [4] describes deductive reasoning as a process that “starts with the assertion of a general rule and proceeds from there to a guaranteed specific conclusion”. This means that if the original assertions are true, then the conclusion must also be true. How can deductive reasoning be conducted in deep learning artificial neural networks? The binding of a variable by an operator is a perfect candidate, as an operator always possesses general rules while a variable is always a specific application of the rule. If an operator is true, the variable that it binds must be true.

Chen [10] defines the notion of contextual binding. It is used for intelligent targeting in Chen [11]. The definitions are cited below:

Assume that X is a variable, CO is a contextual operator. To denote the relationship between the two entities, the subscripts “ i ” and “ j ” are used. If X is the same as OP or X is a member of OP , the subscript “ i ” is assigned to both entities, i.e. CO_i and X_i . If they are within the same setting, the following representation is used:

$$CO_i \{ \dots X_i \dots \} \quad (2)$$

In (2), X_i is contextually bound by CO_i .

If X is different from OP or X is not a member of OP , then the subscript “ i ” is assigned to OP and the subscript “ j ” is assigned to X , i.e. CO_i and X_j . If they are within the same setting, the representation below is used:

$$CO_i \{ \dots X_j \dots \} \quad (3)$$

In (3), X_j is not contextually bound by CO_i .

The Basic Contextual Binding Condition:

The entity X_i is directly related to CO_i iff (if and only if) CO_i provides a context that the interpretation of X_i solely depends on.

Based on the Basic Contextual Binding Condition, the Restrictive Contextual Binding Condition is derived in Chen [10].

The Restrictive Contextual Binding Condition:

Assume X is an entity, and CO is a contextual operator. In a specialized time, location, environment, and background, if X is directly related to CO with respect to all the attributes such as action-initiator (who), action (what), action-recipient (who/what_recipient), time (when), location (where), method (how), and purpose (why) in such a setting:

$$\begin{aligned}
 & CO_i[WHO1, WHAT2, WHAT_RECIPIENT3, WHEN4, WHERE5, \\
 & \quad HOW6, WHY7] \\
 & \{ \dots X_i[WHO1, WHAT2, WHAT_RECIPIENT3, WHEN4, \\
 & \quad WHERE5, HOW6, WHY7] \dots \}
 \end{aligned} \tag{4}$$

Here, X_i is contextually bound by CO_i in a restrictive way. If one contextual attribute in the variable is not directly related to the corresponding attribute in the contextual operator, the variable is not contextually bound by the contextual operator in the restrictive sense.

Applying the contextual binding to the summation operator for forward propagation in (1), the following formula can be derived:

$$netinput = CO_j(bj + \sum_{i=1}^n (Xi * Wi)) \tag{5}$$

Here in (5), a contextual operator “ CO_j ” is an entity consisting of a general property while the variable “ b_j ” is an entity containing the particular property that is part of the general property held in “ CO_j ”. Hence, if the entity “ CO_j ” is true, the variable “ b_j ” must be true. Thus, deductive reasoning is added into the deep learning artificial neural networks.

Interestingly enough, this solution naturally takes care of the issue of having unbound variables. In (5), the variable “ b_j ” is no longer an unbound variable. This means that it is not random and it cannot be substituted by any element in the process. If an irrelevant element gets injected into this position, it will not be processed, or it will be pushed out from that position as it does not satisfy the Basic Contextual Binding Condition.

Clearly, the integration of contextual binding into deep learning artificial neural networks can kill two birds with one stone, as it successfully addresses the two issues in the current approach. A contextual operator is able to provide intelligent hints or guidance in deep learning, speeding up the whole process by skipping irrelevant datasets and avoiding unnecessary calculations.

In addition, various types of intelligence other than logical-mathematical intelligence are added into the deep learning system. Each one is treated as a new dimension. For example, linguistic intelligence, interpersonal intelligence, spatial intelligence, and musical intelligence are all included. The attributes or features associated with a target in all these dimensions are loaded within a corresponding contextual operator, with the relations among them clearly defined. This extends the approach to a multi-dimensional or multi-level one, represented below in (6).

Here, the contextual operator “ CO ” consists of the indexes of “ j, k, l ”. The index “ j ” represents the characteristics in one dimension, say logical-mathematical dimension. The index “ k ” represents the characteristics in another dimension, say linguistic dimension. The index “ l ” represents the characteristics in still another dimension, say interpersonal dimension. Putting the unique characteristics of all these dimensions together, the target can be quickly and accurately identified.

$$\begin{aligned} \text{netinput} = COj, k, l\{ & [bj + \sum_{i=1}^n (Xi * Wi)] \\ & [bk + \sum_{i=1}^n (Xi * Wi)] \\ & [bl + \sum_{i=1}^n (Xi * Wi)] \} \end{aligned} \quad (6)$$

This representation is more powerful than the one in (5) as it covers more dimensions of the same target and it links these dimensions together with the unique features held within one contextual operator.

In the next section, the benefits of this new approach are discussed. So is its implementation.

4 Discussion

Let us take a look at the advantages of the new approach first.

4.1 Advantages of the New Approach

There are a number of advantages for the integration of contextual binding into deep learning artificial neural networks.

First, deductive reasoning is used. This complements inductive reasoning and takes care of the issues that cannot be solved by inductive reasoning alone. Consequently, new perspectives are enabled, thus enriching the deep learning system as varied methodologies are simultaneously adopted.

Second, a contextual operator can serve as an intelligent hint that guides a fast and accurate search. Not all the input datasets are needed. Only the relevant ones are summoned for processing. This approach is able to quickly and accurately identify a target while saving time and resources in the process.

Third, supervised learning is thus enabled in deep learning artificial neural networks. The availability of pre-defined default settings speed up the analysis process even if intelligent hints are not offered.

Fourth, the priority list within a contextual operator can be dynamically changed based on contexts. This dynamic parameter-setting capability makes it possible for the deep learning system to adapt to changes constantly, thus making the system itself more intelligent.

Fifth, the multi-level or the multi-dimensional approach provides perspectives from different angles. The unique characteristics from different dimensions are all related to a target. They are captured as unique features in a contextual operator. Knowing the relationship among these features, the contextual operator can quickly figure out a priority list for that specific search, thus speeding up the process in precisely identifying the target.

There are other advantages in addition to the ones discussed above. In a nutshell, this new approach can lead to fast and accurate attribution. It also can further improve the current approach in deep learning artificial neural networks.

4.2 Implementation

As shown in the Restrictive Contextual Binding Condition, a contextual operator consists of the following attributes: “who”, “why”, “how”, “when”, “where”, and “what”. In Chen [12], weight is assigned to each attribute, displayed below:

Weight in probability for each attribute:

$$\begin{aligned} \text{“who”}: 0.3 &\leq W_1 < 1 \\ \text{“why”}: 0.25 &\leq W_2 < 0.3 \\ \text{“how”}: 0.15 &\leq W_3 < 0.25 \\ \text{“when”}: 0.1 &\leq W_4 < 0.15 \\ \text{“where”}: 0.1 &\leq W_5 < 0.15 \\ \text{“what”}: 0.1 &\leq W_6 < 0.15 \end{aligned}$$

The total weight of probability equals 1.

If an attribute is known, it carries the value “1”. Otherwise, it has the value “0”.

The probability of successful attribution is expressed as follows:

$$P(X) = \sum_{i=1}^6 (X_i * W_i) \quad (7)$$

Given the weight of each attribute listed above, the formula in (7) can be rewritten as follows:

$$\begin{aligned} P(X) &= \sum_{i=1}^6 (X_i * W_i) \\ &= (X_1 * W_1) + (X_2 * W_2) + (X_3 * W_3) + (X_4 * W_4) \\ &\quad + (X_5 * W_5) + (X_6 * W_6) \\ &= (1 * 0.3) + (1 * 0.25) + (1 * 0.15) + (1 * 0.1) + (1 * 0.1) + (1 * 0.1) \\ &= 0.3 + 0.25 + 0.15 + 0.1 + 0.1 + 0.1 \\ &= 1 \end{aligned} \quad (8)$$

Once all the attributes are known, the target is successfully attributed to.

Meanwhile, when all the attributes are properly addressed in an expected way, the Restrictive Contextual Binding Condition listed in (4) above is satisfied, as the variables are properly bound by their corresponding contextual operators.

A contextual operator can help to derive a priority list for investigation. For example, if a target should be found but no single attribute is known, the default priority list is recommended based on the weight of each attribute shown above.

Within the list, the attribute “who” has the highest priority as it may possess the highest weight, i.e. up to 30% of the total weight. Accordingly, the following default priority list can be generated:

$$\text{“who”} > \text{“why”} > \text{“how”} > \text{“when”} \geq \text{“where”} \geq \text{“what”}$$

Nonetheless, the default priority list above can certainly be adjusted or revised should the environment be changed. For example, if the attributes “who”, “why”, “when”, and “what” are figured out and the attributes “how” and “where” remain unknown, the priority list is thus adjusted to focus on the attribute “how”, as it possesses more weight than the attribute “where”. The adjusted priority list thus looks like this:

$$\text{“how”} > \text{“where”}$$

If the purpose “why” and the method “how” need to be discovered while the rest of the attributes are known, then the attribute “why” has the priority in the adjusted priority list as it possesses more weight than the attribute “how”, as shown below:

$$\text{“why”} > \text{“how”}$$

Once the priority list is set, the corresponding contextual binding is adjusted accordingly, and revised investigation guidance is thus generated.

Assume what needs to be identified in the investigation is the person who launched an attack. Further assume the input datasets from different dimensions (i.e. log files, digital analysis results, corresponding images and videos) are available, the deep learning formula in (6) is thus activated. Here, at least two levels are established to reflect two different dimensions. The contextual operator is responsible for the establishment of priority lists for both levels. If the same individual is identified within both levels, the probability of a successful attribution is greatly increased.

As shown here, the implementation of this new approach can lead to successful attribution in digital forensic investigation in a fast and accurate way. Based on this design, corresponding algorithms can be developed and prototype of a novel system can be built.

5 Conclusion

The current approach in deep learning artificial neural networks lacks deductive apparatus and variable binding. The proposed approach addresses these challenges by utilizing contextual binding, which not only supports deductive reasoning but also provides intelligent hints and guidance in searches for investigation. This new approach can certainly lead to fast and accurate attribution while making deep learning more efficient and effective in support missions.

References

1. Schmidhuber, J.: Deep learning in neural networks: an overview. Technical Report IDSIA-03-14, the Swiss AI Lab IDSIA, University of Lugano & SUPSI, Switzerland (2014)
2. Klauer, K., Phye, G.: Inductive reasoning: a training approach. *Rev. Educ. Res.* **78**(1), 85–123 (2008). <https://doi.org/10.3102/0034654307313402>
3. Dormehl, L.: What is an artificial neural network? Here's everything you need to know. Emerging Tech, Digital Trends, May 11, 2018. <https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network/>. Last accessed 2 July 2018
4. The Center for Academic Success at Butte College: Deductive, inductive and abductive reasoning, <http://www.butte.edu/departments/cas/tipsheets/thinking/reasoning.html>. Last accessed 10 May 2018
5. Gardner, H.: *Frames of Mind: The Theory of Multiple Intelligences*, 3rd edn. Basic Books, London (2011)
6. Bradford, A.: The five (and more) senses. Livescience, October 23, 2017. <https://www.livescience.com/60752-human-senses.html>. Last accessed 28 May 2018
7. Taylor, M.: *Neural Networks Math: A Visual Introduction for Beginners*. Blue Windmill Media (2017)
8. Beebe, N.: Digital forensic research: the good, the bad and the unaddressed. In: Peterson, G., Shenoi, S. (eds.) *Advances in Digital Forensics*, vol. V, pp. 17–36. Springer, Berlin (2009)
9. Marti, R., Reinelt, G.: Heuristic methods. In: Marti, R., Reinelt, G. (eds.) *The Linear Ordering Problem: Exact and Heuristic Methods in Combinatorial Optimization* 175, pp. 17–40. Springer, Berlin (2011). https://doi.org/10.1007/978-3-642-16729-4_2
10. Chen, J.: On contextual binding and its application in cyber deception detection. In: *Proceedings of the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 215–218. IEEE, New York (2015)
11. Chen, J.: Contextual binding and intelligent targeting. In: *Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 701–704. IEEE, New York (2016)
12. Chen, J.: An intelligent path toward accurate attribution. In: *Proceedings of the 2018 IEEE/SAI Computing Conference*, pp. 1134–1139. IEEE, New York (2018)



Intelligent Signal Classifier for Brain Epileptic EEG Based on Decision Tree, Multilayer Perceptron and Over-Sampling Approach

Jimmy Ming-Tai Wu¹, Meng-Hsiun Tsai², Chia-Te Hsu²,
Hsien-Chung Huang^{2(✉)}, and Hsiang-Chun Chen²

¹ Shandong University of Science and Technology, 579 Qianwangang Road, Huangdao District, Qingdao 266590, Shandong Province People's Republic of China
wmt@wmt35.idv.tw

² National Chung Hsing University, 145 Xingda Rd., South Dist, Taichung City 402, Taiwan (R.O.C.)
{mht, arthurhsu}@nchu.edu.tw,
hchwang@dragon.nchu.edu.tw, cosirdcs@gmail.com

Abstract. Epilepsy is a chronic neurological disease induced by abnormal electrical discharges of brain which tends to irregular seizures. The seizures may cause the patients to lose consciousness and the patients couldn't control their muscles. Epilepsy even possibly endangers one's life. Electroencephalogram (EEG) is a common tool used in the clinical diagnosis and analytics of epilepsy. However, the visual examination of EEG is time-consuming and the diagnostic result is also easily influenced by the viewer's subjective judgement. Therefore, the purpose of this study is to construct an automatic classifier, which could be helpful to analyze, for the epileptic EEG signals. The EEG recordings of patients with intractable epilepsy, which are collected by Boston Children's Hospital, are used in this study. The features of EEG signals in time and frequency domains are collected from results of the Fast Fourier Transform. The Synthetic Minority Oversampling Technique (SMOTE) is used to solve the data imbalance problem. Four machine learning algorithms including C4.5, Classification and Regression Tree (CART), Chi-Square Automatic Interaction Detector (CHAID) and Multilayer Perceptron (MLP) are used to classify the data. As a result, the accuracy rate of the proposed classifier is 99.48%. It might be a clinical assistant tool for doctors to make a more reliable and objective diagnosis.

Keywords: Epilepsy · Electroencephalogram · Fast fourier transform · Oversampling technique · Machine learning algorithms

1 Introduction

1.1 What Is Epilepsy

Epilepsy is a chronic neurological disease induced by abnormal electrical discharges of brain. The location of the discharge will result in different clinical seizures. The patient might have short-term loss of consciousness and muscle twitching. Two main types of seizures are observed: generalized seizures and focal seizures [1–3].

Generalized seizures: The most typical one is the tonic-clonic seizure. It usually causes upward gaze, cyanotic lips, spasticity, stiff limbs, uncontrollable drooling or even incontinence.

Focal seizures: The symptoms are partial tic or numbness, losing ability of expression and involuntary behaviors such as chewing, blinking, swallowing, talking to himself, scratching or walking around.

Although epileptic seizures are mostly temporary, it still causes great inconvenience to patients because of its irregularity and frequency. Patients must always be accompanied and can't get a driver's license.

1.2 Diagnosis and Treatment

Nowadays, the most important diagnosis of epilepsy is based on comprehensive descriptions offered by witnesses during patients' epileptic seizures. It is necessary that the information detailed such as situation, time of duration and consciousness of patients. Examinations include physical checkup, blood test, neurological assessment, Computed Tomography scan (CT scan), Nuclear Magnetic Resonance Imaging (NMRI), Electroencephalography (EEG), etc. [3–6]. By using EEG, doctors could determine whether patients are suffered from epilepsy, classify the types of epilepsy and observe the performance of treatments.

Although epilepsy is an incurable disease, the seizures of about 70% patients could be controlled after they received the treatment and took Anti-Epileptic Drugs (AEDs) [7]. If the medicines are not effective, the patients are able to have a brain surgery for treating epilepsy after evaluation. Either surgical resection or surgical treatment has side effects, thus patients should be carefully examined by EEG before the surgery.

1.3 Electroencephalogram (EEG)

The cerebral cortex contains a large number of neurons, which open and close ion channels all the time, produce and transmit electrical signals. Neurons would generate a tiny electric field during propagating electrical signals and these actions of potentials can be recorded by sticking dozens of electrodes on patient's scalp. These potential signals are very weak, so it is necessary to amplify the variations of waveforms by signal amplifier, so-called brainwave or Electroencephalogram (EEG). EEG is a common tool of noninvasive diagnosis which could be used to record important physiological parameters of brains. With the advantages, including long-term monitoring and without involving radioactivity, its application in medicine is very extensive.

According to International Federation of Clinical Neurophysiology (IFCN), the frequencies of EEG can be divided into four sub-bands: Delta (δ , 0.5–4 Hz), Theta (θ , 4–8 Hz), Alpha (α , 8–13 Hz) and Beta (β , 13–30 Hz). Details of frequency sub-bands are shown in Table 1 [8].

Table 1. Comparison of EEG bands

Band	Waveform	Description
Delta		Frequency: 0.5–4 Hz Unconscious dimension Mainly occurs in deep sleep or unconsciousness
Theta		Frequency: 4–8 Hz Subconscious dimension Becomes the main wave when we enter status of deep relaxation or drowsiness to fall asleep
Alpha		Frequency: 8–13 Hz A bridge between conscious and subconscious dimensions A periodic wave generally appearing in conscious, quiet and resting situation
Beta		Frequency: 13–30 Hz Conscious dimension Usually occurs when we are conscious, awaken suddenly, nervous or excited

1.4 Imbalanced Data

Imbalanced data, which unevenly distributed the data of each category, indicate that the proportion of one category is much larger than another category in a dataset and the dataset is called an imbalanced dataset [9]. In this kind of dataset, minor categories are more important. For example, in the diagnosis of few diseases, there are usually more cases of normal patients and fewer cases of sick patients. That is, in the overall case data, the proportion of sick case data is much lower than normal case data. Such a dataset often tends to interpret cases as majority when classifying in machine learning and it causes bad classified results of few categories.

Currently, the methods have been proposed to solve the problems of imbalanced data mainly divided into two general directions: from algorithmic dimension and from data dimension. This research focused on the improvements of imbalanced data from data dimension: pre-processing datasets to balance the amount of data between each category. The general methods of data level are as follows:

Undersampling:

Random Undersampling: Data of categories which has higher proportion would be removed randomly until the proportion of each category tends to balance. This method randomly selects samples to delete and the important data of training process would be lost easily. As a result, the effect of machine learning is not ideal.

Easy Ensemble: The major categories of sample data in the dataset are equally divided into many sample subsets and the quantity of each subset is about the same as the minor categories. Each subset and the minor categories are combined into a new training sample, then each training sample is trained with a classifier. The result of each training sample is averaged as the final classified model [10].

Oversampling:

Random Oversampling: Randomly select data from minor categories of dataset to generate subsets, and directly copy data in the subset and add into sample dataset [9, 11]. This method is easy to overfitting because of directly copying the data into training samples.

Synthetic Minority Over-sampling Technique (SMOTE): In order to solve overfitting caused by a Random Oversampling of SMOTE, which increase data of minor categories, is extended from Random Oversampling [12]. Instead of randomly copies a data of the minor categories, this method randomly selects samples from the minor categories and randomly generates new samples between selected samples and adjacent samples of the same categories. As a result, new generated samples will be between samples of minor categories, but it wouldn't overlap original samples.

1.5 Research Process

Section 1 presents the background of research and the challenges to conquer. Section 2 provides the research materials and methods. Section 3 discussed the experiments and results of the research. Section 4 concludes the whole research scheme.

2 Materials and Methods

2.1 Source

The brainwave signals used in this study were epileptic patients' data at Boston Children's Hospital released by physionet. Patients' data in public databases only leave age and gender [13], other private information was deleted. Original data is shown in Fig. 1.

There are 23 cases from 5 male cases (3–22 years old) and 17 female cases (1–19 years old) in brainwave dataset (1st and 21st case was the same patient and 18 months between two cases). 9–42 continuous brainwave signal fragments in each case and each fragments' length is 1–2 or 4 h. There are 198 epileptic seizure records in 664 fragments. The signal acquisition of this dataset adopts international 10–20 standard electrode positions and 18 channels at 256 Hz.

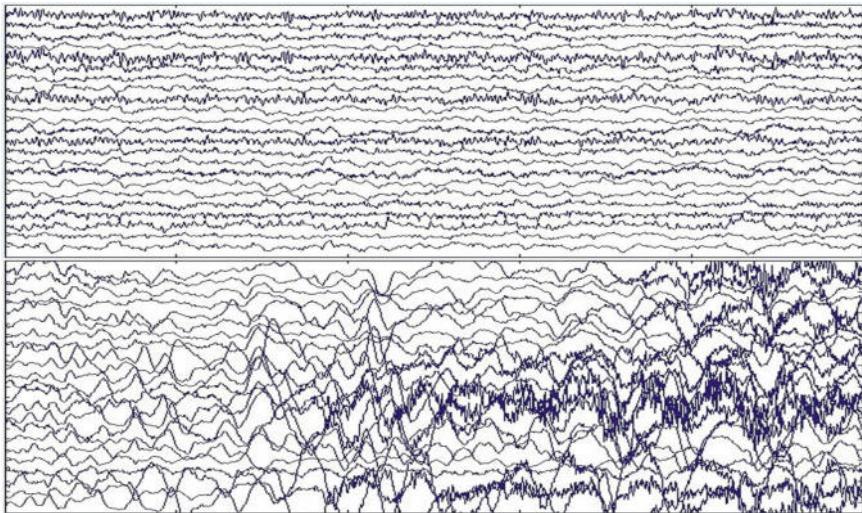


Fig. 1. Original data (up: without epileptic seizure, down: with epileptic seizure)

2.2 Fast Fourier Transform

Fast Fourier Transform (FFT) is a fast algorithm for computing discrete Fourier transforms and inverse transforms. Based on discrete Fourier Transform, FFT improved its disadvantage of expensive calculation [14]. In this study, FFT is used to convert the brainwave signals from time-domain to frequency-domain so that the amount of energy carried by each sub frequency can be analyzed.

2.3 Data Preprocessing

Discretization: Discretization, a technique to convert continuous data into discrete data by using cluster analysis to group numerical data [15]. Because continuous data values often order by relative relationship, the value range is cut into several intervals by using separation techniques to replace original data values and then marked, then the continuous data can be converted into discrete data.

Information Gain: The Information Gain can evaluate attributes by measuring corresponding attribute's gain value. For example, attribute A have n values $\{a_1, a_2, \dots\}$ in dataset S, S can be divided into n subsets by $\{a_1, a_2, \dots\}$. Gain (A), the gain value of using A in S as branching attribute, is equal to total Gain value of S subtract Gain value after branching. Defined as formula (1):

$$Gain(A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} Entropy(S_i) \quad i = 1, 2, \dots, n \quad (1)$$

The larger Gain value of the attribute is, the better to use the attribute as branching attribute.

SMOTE: The methods to improve data imbalance can be divided into under-sampling and oversampling. In case of the samples of this research has fewer data of minor categories, SMOTE is adopted to improve data imbalance.

2.4 Decision Tree

Three common decision trees are used in this study: C4.5, CART and CHAID. The comparison is shown in Table 2.

Table 2. Comparison of decision tree algorithms

	C4.5	CART	CHAID
Data type	Discrete, continues	Discrete, continues	Discrete
Branches based	Information gain	Gini coefficient	Chi-square test
Branches method	Multiple	Binary	Multiple
Pruning method	Base on the error	Cost complexity	N/A

2.5 Multiple Layer Perceptron

Multiple Layer Perceptron is a multi-layer feedforward neural network. Besides an input layer and an output layer, there are one or several middle layers, which are also called hidden layers [16], in the multi-layer. The structure of Multiple Layer Perceptron is shown as Fig. 2.

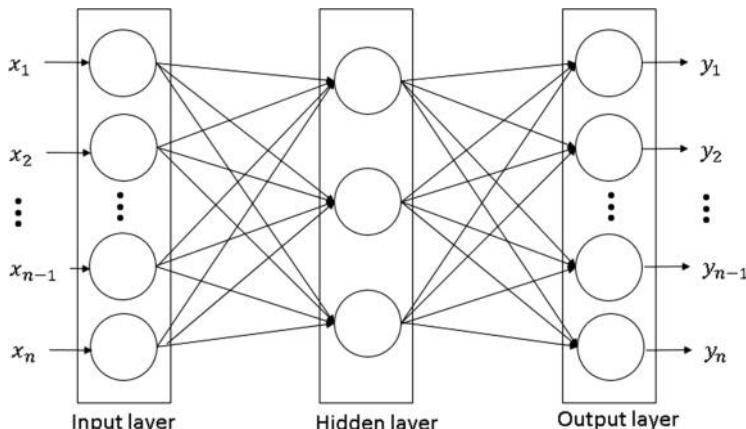


Fig. 2. Multiple layer perceptron (one hidden layer)

3 Experiments and Results

3.1 Research Process

The procedure of this study is shown in Fig. 3. There are three parts: signal processing, data preprocessing and classifier construction. Several software was applied in this study (MATLAB, WEKA and Statistica). The versions of software are MATLAB R2013a, WEKA3.6, Statistica 13 and the system environment is Window 7.

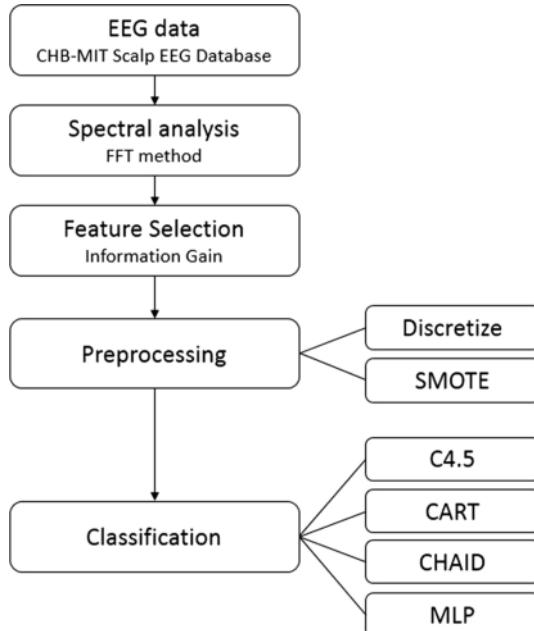


Fig. 3. Research process

3.2 Signal Processing and Feature Extraction

Signal Processing: The raw data used in this study was multi-channel EEG, which contained 23 channels of brainwave signals. Based on the research conducted by Ali Shoeb and John Guttag in 2009 [17], the rhythmic activities of epilepsy are the most notable and least noisy in channel T8-P8. Therefore, T8-P8 channel was extracted as experimental data in this study. Figure 4 shows an EEG with epileptic seizure in channel T8-P8. Moreover, linear marks indicate the beginning of a seizure.

After extraction of the T8-P8 channel, first step is to remove the long-time noise part due to environmental interference or equipment problems during the acquisition of the brainwave signal (as shown in Fig. 5).

Second, cutting each EEG as a section per 30 s, this step collects 19,046 un-seizure activities and 221 seizure activities.

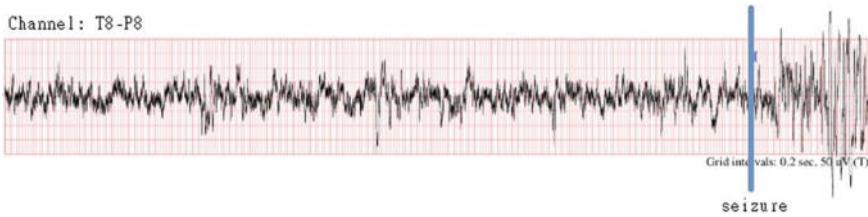


Fig. 4. EEG of channel T8-P8 in a minute

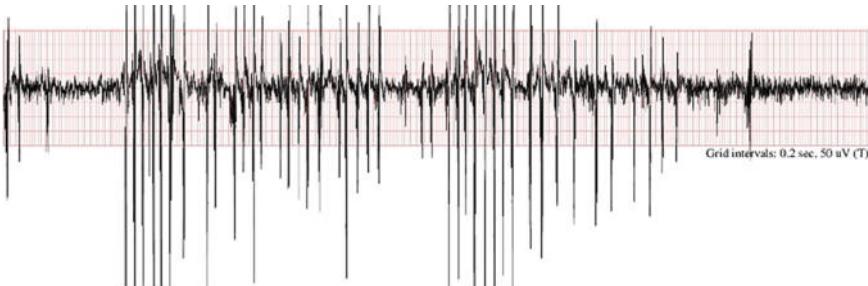


Fig. 5. EEG with noise

In the previous step, the long-time noise caused by the equipment problem has been removed. However, during the brainwave acquisition, there are some short-time noise will be recorded by the machine due to sudden eye movement or muscle tightening. Therefore, the Savitzky-Golay filter is used in this study to remove these short-time noise. Savitzky-Golay filter, which is a low-pass filter by using the least squares method to smooth signal processing, is widely used in signal smoothing and eliminating noise [18]. Figure 6 shows an original 30 s brainwave signal versus the signal waveform processed through the Savitzky-Golay filter.

Time-Domain Features: In time-domain, maximum amplitude and standard deviation (SD) of the brainwave signal segment are extracted in this study as features of brainwave to classify epilepsy. The standard deviation of amplitude can indicate whether the change of amplitude during this period is significant.

Frequency-Domain Features: To extract frequency-domain features, this study uses FFT to convert the brainwave signal form time-domain to frequency-domain. Figures 7 and 8 show that before and after brainwave signals using FFT. Figure 7 are brainwave signals with un-seizure activities and Fig. 8 are brainwave signals with seizure activities.

According to the study by Anusha et al. [19], the energy carried by each sub-band in frequency-domain is used as features. In this study, based on International Federation of Societies for Electroencephalography and Clinical Neurophysiology, the spectrum was divided into four sub-bands: Delta wave (δ , 0.5 Hz–4 Hz), Theta wave (θ , 4 Hz–8 Hz), Alpha wave (α , 8–13 Hz) and Beta wave (β , 13–30 Hz), then take the maximum energy and average energy in each sub-band as features for epilepsy classification in frequency-domain.

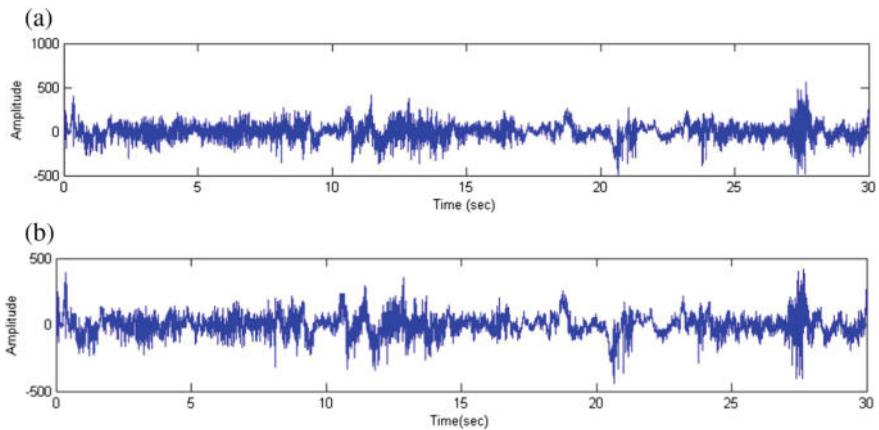


Fig. 6. The signal filtering charts (**a** before filtering, **b** after filtering)

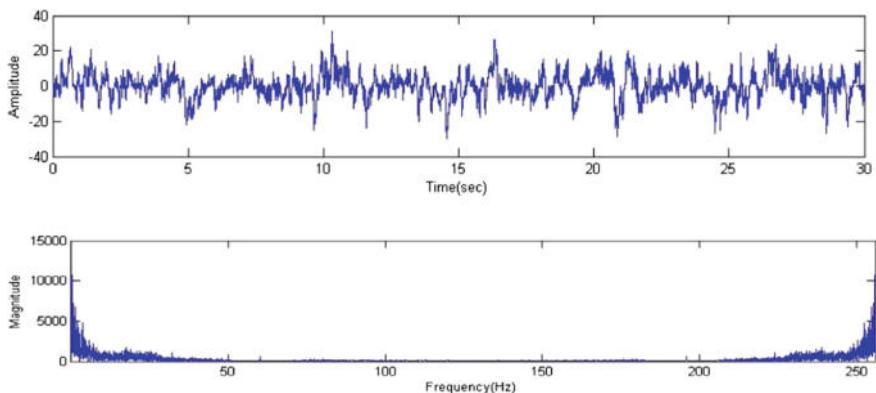


Fig. 7. Signal charts using FFT with un-seizure activities (before and after)

3.3 Result of Data Preprocessing

Discretization: The features of each EEG, which obtained through the previous feature extraction steps, are not only more complex, but also need to be classified by using decision tree classification methods. Hence, features of each data should be discretized. In this study, the discretization method uses equal-width discretization method to discretize ten continuous features.

Feature Extraction: Before constructed the predicting model, features were extracted by Information Gain. Then, two features of maxD and meanD were decided to delete by the results of Information Gain, which have big difference in Gain value.

SMOTE (Synthetic Minority Over-Sampling Technique): After signal segmentation, this study has 19,046 un-seizure activities and 221 seizure activities. As the

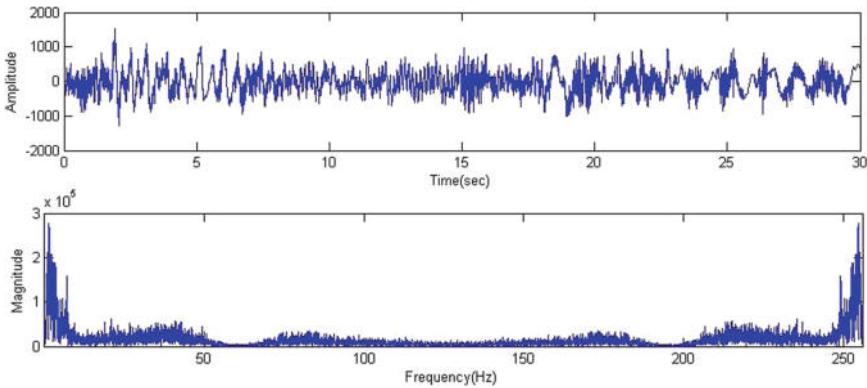


Fig. 8. Signal charts using FFT with seizure activities (before and after)

ratio of positive and negative cases varies widely, it is an imbalanced dataset. In order to avoid the misclassification problems, this study employ the algorithm SMOTE to synthesize number of seizure samples and set the parameter of nearest neighbors as 5 [20]. That is, in the seizure samples, select sample X and randomly selected one of samples from five seizure samples that is closest to X to generate a new synthesized sample.

Since number of synthetic samples produced by SMOTE in WEKA software is double the original number of samples, SMOTE is allowed roughly to balance between two labels (seizure and un-seizure). After using SMOTE to process six times, we finally obtain 19,046 un-seizure samples and 14,414 seizure samples.

3.4 Construction of Classification

Classification Model with feature of single sub-band: During the process of establishing decision tree model, we found out that the average energy of Theta band in C4.5, CART and CHAID is in the node of first classification. The second node is the maximum energy of Theta band in C4.5 and CART, indicates that the Theta band might be an important attribute in classification of epilepsy by EEG signal. In order to verify the assumptions, two features, which are extracted from time-domain, are divided into one group. Other groups of features are based on each frequency band. Details of groups are shown in Table 3. After grouping, the original data table is split

Table 3. Property grouping table

Group	Feature
TD (time domain)	max, std
Alpha	maxA, meanA
Beta	maxB, meanB
Theta	maxO, meanO

into four new tables and each table contains a set of characteristics and “type” attribute. Similar to previous data preprocessing, after discretizing and SMOTE each data table, the classifiers are constructed respectively with MLP, C4.5, CART and CHAID. The accuracy of classification constructed by single sub-band feature is shown in Table 4.

Table 4. Comparison of classification accuracy established by single sub-band feature

Feature group	Classification methods			
	C4.5 (%)	CART (%)	CHAID (%)	MLP (%)
TD	91.45	91.48	89	91.39
Alpha	87.54	87.55	85.48	87.12
Beta	81.67	81.69	78.63	81.68
Theta	95.46	95.41	95.47	95.58
All	99.39	99.41	95.44	99.48

In the experiment of constructing a classifier based on single frequency band, it can be seen from the classification accuracy table that the accuracy of classification based on Theta band group is not only significantly better than that of the other groups, but also the accuracy of classification is higher than 95% proved that the hypothesis of proposed Theta band as an important attribute in the classification of epilepsy by EEG signals.

Construction of classifier with random sampling: The methods to improve imbalanced data could be divided into undersampling and oversampling. Because the quantity of epileptic samples is small, SMOTE, which is a method of oversampling, is chosen to synthesize artificial samples to construct classifier. To verify the performance of SMOTE, the result of classifier construction with Random Undersampling as pre-processing method is shown in this section. In this experiment, control samples, which are extracted by random sampling, are the same, twice, three times, four times or five times as amount of epileptic seizure samples. In order to avoid the extreme samples from affecting the experimental results, each quantity is extracted ten times and classified individually.

The results showed that in the classification, which has same quantity of control samples as epileptic seizure samples, the accuracy was between 89–97% and the average accuracy is 94.43%. When control samples have double the amount of epileptic seizure samples, the accuracy was between 91–97% and the average accuracy is 95.46%. When control samples have three times the amount of epileptic seizure samples, the accuracy was between 94–97% and the average accuracy is 96.05%. When control samples have four or five times the amount of epileptic seizure samples, the accuracy was between 95–97% and the average accuracy is 96.81%. Overall, the classification accuracy increases with the amount of samples. In the construction of classifier, if number of samples are too small, which will easily lead to a few extreme samples, affect the result of whole classification. Therefore, SMOTE is chosen to solve the problem of imbalanced data in this research. Artificial samples of epileptic seizure samples, which are synthesized by SMOTE, could avoid probable problems of extreme samples by other methods.

3.5 Evaluation of Classification Model

Evaluation Indicators: In order to confirm whether the proposed classification model has a good classified ability of epileptic seizure. Several common classified indicators, including accuracy, specificity and area under the curve (AUC) of ROC, are used to evaluate the proposed classification model in this research.

Model Evaluation: Four classification models proposed in this study are respectively evaluated by accuracy, specificity and AUC. The results are shown in Table 5.

Table 5. Evaluations of classification models

	Accuracy (%)	Specificity (%)	AUC
C4.5	99.39	99.73	0.998
CART	99.41	99.75	0.998
CHAID	95.44	95.35	0.991
MLP	99.48	99.71	0.999

In this research, the accuracy of C4.5, CART, MLP is higher than 99% and CHAID is lower than 95.44%. To determine whether the classification of epileptic seizure EEG classifies accurately, specificity is an important indicator in this research. Although the accuracy and AUC of C4.5 and CART are slightly lower than MLP, the specificity is higher. Three classifiers, C4.5, CART and MLP are perform well in all respects, which prove that artificial neural network and decision trees are good methods for analyzing EEG signals.

4 Conclusion

FFT is used to process EEG signals in this research. Ten features are extracted respectively in time-domain and frequency-domain, then this research deducts variables by information gain. The method of oversampling, which is seldom used by people in previous related researches, is used after filtering out two features (Delta sub-band) with lowest gain values. The classified accuracy of C4.5 is 99.39%, CART is 99.41% and MLP is 99.48%. All of them have good standards and are superior to researches which used the same dataset in the past.

In the process of decision tree algorithms' construction, which have been found that features extracted from the Theta sub-band (4–8 Hz) in brainwave are very important attributes in the classification of epileptic seizure EEG signals. Consequently, the classifiers are constructed by two algorithms and only use features extracted from Theta sub-band. The accuracy of both CART and MLP is all over 99%. Although the estimation with multiple features is usually used to effectively classify the epileptic seizure EEG signals in clinical diagnosis, this discovery could be a reference of epileptic seizure EEG research in future.

Acknowledgements. The authors would like to thank the reviewers for their valuable suggestions and comments that are helpful to improve the content and quality for this paper. This paper is supported by the National Science Council of Taiwan, ROC, under the contract of MOST 106-3114-E-005-008-, MOST 106-2119-M-005-006- and the National Chung Hsing University-Chung Shan Medical University cooperative research project, under the contract of NCHU-CSMU 10707.

References

1. Awad, I.A., Rosenfeld, J., Ahl, J., Hahn, J.F., Lüders, H.: Intractable epilepsy and structural lesions of the brain: mapping, resection strategies, and seizure outcome. *Epilepsia* **32**, 179–186 (1991)
2. Fisher, R.S., Boas, W.V.E., Blume, W., Elger, C., Genton, P., Lee, P., et al.: Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia* **46**, 470–472 (2005)
3. Smeets, V.M., van Lierop, B.A., Vanhoutvin, J.P., Aldenkamp, A.P., Nijhuis, F.J.: Epilepsy and employment: literature review. *Epilepsy Behav.* **10**, 354–362 (2007)
4. Gotman, J., Gloor, P.: Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG. *Electroencephalogr. Clin. Neurophysiol.* **41**, 513–529 (1976)
5. Yoo, J., Yan, L., El-Damak, D., Altaf, M.A.B., Shoeb, A.H., Chandrakasan, A.P.: An 8-channel scalable EEG acquisition SoC with patient-specific seizure classification and recording processor. *IEEE J. Solid-State Circuits* **48**, 214–228 (2013)
6. Shorvon, S.D.: *Handbook of Epilepsy Treatment*. Wiley, New York (2010)
7. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386 (1958)
8. Ray, W.J., Cole, H.W.: EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science* **228**, 750–752 (1985)
9. Chawla, N.V.: Data mining for imbalanced datasets: an overview. In: *Data Mining and Knowledge Discovery Handbook*, pp. 853–867. Springer, Berlin (2005)
10. Liu, X.-Y., Wu, J., Zhou, Z.-H.: Exploratory under-sampling for class-imbalance learning. 965–969 (2006)
11. Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 148–156 (1994)
12. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 321–357 (2002)
13. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, L., Ivanov, P.C., Mark, R.G., et al.: Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation* **101**, 215–220 (2000)
14. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* **19**, 297–301 (1965)
15. Han, J., Kamber, M., Pei, J.: *Data Mining: Concepts and Techniques*. Elsevier, Amsterdam (2011)
16. Gardner, M.W., Dorling, S.: Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* **32**, 2627–2636 (1998)
17. Shoeb, A.H.: *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. Massachusetts Institute of Technology (2009)

18. Azami, H., Mohammadi, K., Bozorgtabar, B.: An improved signal segmentation using moving average and Savitzky-Golay filter (2012)
19. Anusha, K., Mathews, M.T., Puthankattil, S.D.: Classification of normal and epileptic EEG signal using time & frequency domain features through artificial neural network. In: 2012 International Conference on Advances in Computing and Communications (ICACC), pp. 98–101 (2012)
20. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing, pp. 878–887 (2005)



Korean-Optimized Word Representations for Out-of-Vocabulary Problems Caused by Misspelling Using Sub-character Information

Seonhghyun Kim^(✉), Jai-Eun Kim, Seokhyun Hawang,
Berlocher Ivan, and Seung-Won Yang

AI Labs, Saltlux Inc., Seoul, Republic of Korea
`{seonghyunkim, jekim, shhwang, ivan, swyang}@saltlux. com`

Abstract. In this paper, we propose Korean-optimized word representations that can better address the out-of-vocabulary (OOV) problem caused by misspelling. This problem is an important issue in many applications based on natural language processing. However, previous models do not fully consider the representations of misspelled OOV words. To overcome this problem, we propose sub-character information obtained from Korean Jamo units and also adopt additional sub-character information to better withstand the misspelling. Finally, experimental results show that our model is about 2.3 times more accurate than the conventional model in case of the misspelled word while still maintaining the semantic relationship of the words.

Keywords: Word embedding · Word representation · Machine learning · Korean · Out-of-vocabulary · Misspelling · Sub character · Natural language processing

1 Introduction

Continuous word representations play a major role in many natural language processing (NLP) applications based on neural networks approaches such as named entity recognition (NER) or machine reading comprehension (MRC). Previous studies that measure the semantic relatedness between words using word embedding, such as Word2Vec [1, 2], have been successfully implemented in these applications [3, 4].

However, previous word representations have some limitations. First, they allow computing word vectors for only words that appear in the training data. Second, the linguistic characteristics of languages are not considered for the word representation. In most languages, the semantics of a word are determined by combining subword information, such as morphemes, syntax, and root.

To overcome these problems, recent studies [5, 6] have considered the subword information represented as a bag of character n-grams. Thus, word representations can be more effectively learned by embedding subword information. Moreover, these models calculate the vector of the first word to be seen, often called the out-of-vocabulary (OOV), by using character n-grams. However, directly utilizing subword

information from character n-grams would be vulnerable to misspelled words in some languages based on a featural writing system, such as Korean [7]. This is because misspelling, even of a single character, is a critical factor to change the semantics of subword in a featural writing system [8].

In this paper, we focus on the Korean language. Korean is unique in that each character within a word is composed of two or three sub-character units called Jamo [9, 10]. In this work, we propose optimized word representations for Korean that utilize the proposed sub-character units. We verify that the proposed word representations are better to address the OOV problem than the existing public model for misspelled words while still maintaining a good semantic relationship between words.

2 Related Works

Many studies on word representations have used a morpheme as the subword information because it is the smallest unit of meaning in linguistics [1, 2, 11–13]. References [1, 13] proposed a continuous bag of words (CBOW) model that predicts the current word from the surrounding context words and a skip-gram model that uses the current word to predict the surrounding context words. Reference [2] proposed a GloVe vector model that utilizes the probability of the co-occurrence between words. However, these methods cannot solve the OOV problem. The Facebook AI Research group proposed a word representation method [6, 14, 15] called FastText, which utilizes subword information. This method solves the OOV problem by vectorizing with subwords. According to their results, OOV words also can be vectorized with good semantic relations because subword (character n-grams) vectors contain semantic characteristics. The cosine similarity between -adolesc- from preadolescent (OOV word) and young shows a high semantic relation. However, in our experiments, when OOV occurs by misspelling, word similarity is significantly lower than expected because a subword character is changed.

3 Methods

In this section, we describe an efficient word representation method for the Korean language directly inspired by previous research [6]. Their model can construct OOV word vectors using the vector summation of syllable-based n-gram subword units. In contrast to many other languages, Korean words are not just a concatenation of characters by syllable units [9]. Our main contribution is to propose optimized word representations for Korean by considering its featural characteristics.

3.1 Jamo-Static Model

To illustrate the Korean linguistic features, we show the diverse linguistic levels for the sentence, 오렌지는 맛있다 (Orange is delicious) in Fig. 1. This sentence is

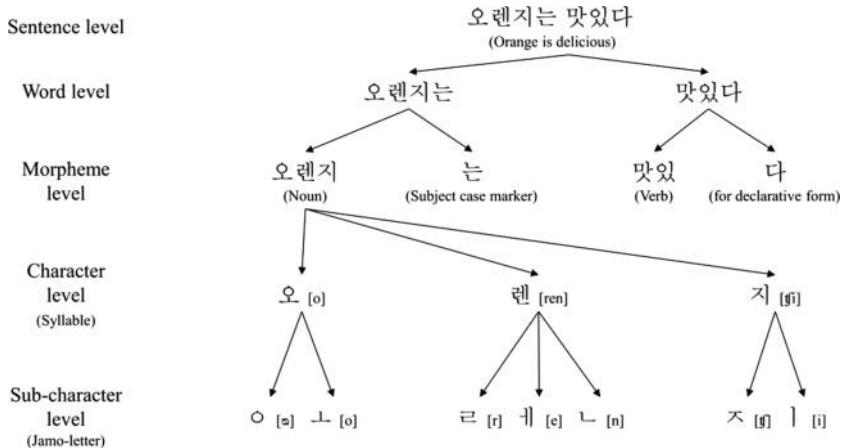


Fig. 1. Hierarchical structure of Korean.

constructed using two words 오렌지는 (Orange) and 맛있다 (is delicious). Words can be decomposed into the morpheme level and the character level.

In the case of 오렌지 (Orange), the word consists of three characters: 오, 렌, and 지. Each character consists of two or three sub-characters (Jamo), which are an initial consonant and a vowel or sometimes a consonant placed under a vowel. The sub-character level is the basic unit of Korean, and it is an important unit for word representation.

To learn word representation based on these Korean characteristics, we tokenize a word into a morpheme unit using special symbols, < and >, at the beginning and end of the words in the same manner as a previous subword model [6]. The morpheme units are decomposed to the character level. Each character is disassembled into sub-characters. We also add a special boundary symbol, ↘, between the characters as a mark to separate them. Taking the word, 오렌지 (Orange), as an example, it can be represented by sub-characters as follow:

<○ ㅗ ↘ ㅓ ㅓ ㄴ ↘ ㅈ ㅣ>

We first conducted an experiment based on this sub-character as a Jamo-static model.

3.2 Jamo-Advanced Model for Misspelling

In Korean, most misspelling problems occur at the sub-character level, especially a vowel, due to similar pronunciations or typos. For example, 렌 [ren], ㅓ [e] is a vowel that can be written using a similar pronunciation, such as 랜 [ræn], 랜 [ryæn], or 련 [ryen].

This type of misspelling causes OOV problems even though the corrected word is part of the vocabulary. Although the meaning of a word and the subword information

are completely distorted by the misspelling, this problem is very common in Korean. Therefore, a novel method is required to correct the semantic distortion and to preserve the subword information of misspelled OOV words.

In our method, we design the additional sub-character n -gram vectors that contain a set of sub-characters without a vowel for each character. As with the word 오렌지 (Orange), each vowel is excluded as ᄂ 렌지, 오ᄅᆞᄂ 지, and 오렌즈, and the additional sub-characters are:

Without ᄂ: <○ ᄁ ᄂ ᄃ ᄂ ᄃ ᄂ ᄃ>

Without ᄃ: <○ ᄂ ᄁ ᄃ ᄂ ᄃ ᄂ ᄃ>

Without] : <○ ᄂ ᄁ ᄃ ᄂ ᄃ]>

In order to compare how this method affects word representation in OOV, we conducted an experiment with Jamo-static model that learns the word vector based on the sub-character unit. Next, we adapted the additional sub-character vectors as a Jamo-advanced model.

4 Experimental Setup

4.1 A Model Variations

Baseline. The pre-trained published word vector model based on character units [6].

Jamo-static. A model that learns the word vector based on the sub-character units.

Baseline. The Jamo-static model noted above but with our additional sub-character n -gram vectors.

4.2 Datasets

We trained our model on the same parameters and training data (Wikipedia dumps of September 2016) used in previous research [6]. In order to evaluate the word representations for misspelled words, we designed a new comprehensive test set containing 250 paired OOV words consisting of misspelled OOV words and corrected words.

4.3 Implementation Details

To tokenize the Wikipedia dumps into morpheme units, we used open API software for morpheme analysis in Korean.¹ Our model embeds words that appear at least five times in the training set into a 300 dimensional space. We use a 1-5 context window size and a 10-4 rejection threshold. The step size is set to 0.025 (for more details, see [6]). The n -gram range is set to 6-12 and includes 2-3 syllables with special boundary symbols.

¹ <http://www.adams.ai/apiPage?tms=POS>.

5 Results

5.1 Correlation with Human Judgment

We first evaluated the semantic relationship between words by computing Pearson's correlation coefficient between human judgement and the cosine similarity of the vector representations. We used the translated WordSim353 dataset [16].

Table 1. The results of translated WordSim353.

Dataset	Model	Correlation
WordSim353	Baseline	0.726
WordSim353	Jamo-static	0.745
WordSim353	Jamo-advanced	0.744

As shown in Table 1, the semantic relation of our Jamo-static and Jamo-advanced models is as strong as the baseline model. We also noticed that our proposed model is more highly correlated with human judgement than the baseline model.

5.2 Similarity Between Misspelled Words and Corrected Words

Cosine Similarity. Next, we compared the cosine similarity between the misspelled OOV words and the corrected words. If the misspelled word is semantically related to the corrected word, it will show a high cosine similarity. Statistical significance was tested using Student's t-test.

Consequently, both the Jamo-static and Jamo-advanced models showed higher similarity scores than the baseline model (Fig. 2). Moreover, the difference between the Jamo-advanced model and the Jamo-static model was statistically significant ($P < 0.001$). These results indicate that subword information is better reflected in the word representation based on sub-character units than word units.

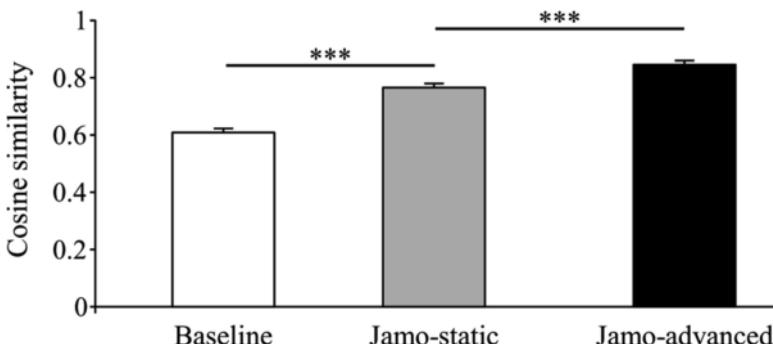


Fig. 2. Cosine similarity between a misspelled word and a corrected word.

Most Similar Words. Finally, to evaluate how close the misspelled word is to the corrected word in the word vector space, we observed the 10 words that are most similar to the misspelled words. The accuracy of the proposed method was determined by checking whether the corrected word was included in the words that are most similar based on the range of words.

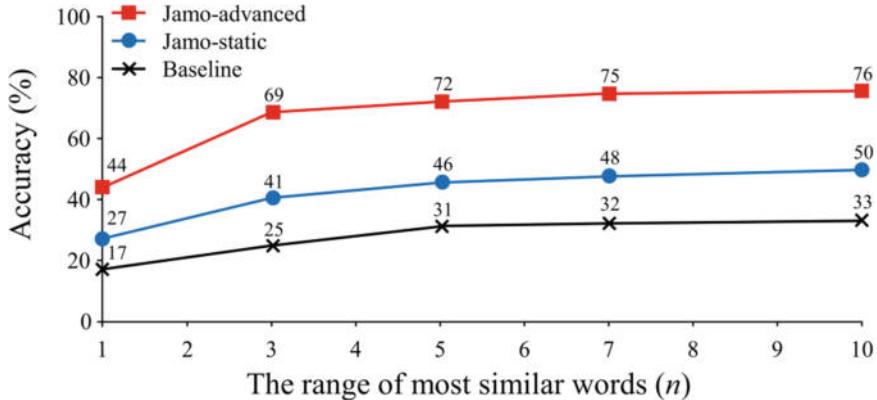


Fig. 3. The ratio that the corrected word is included in the top n most similar words from the misspelled OOV words.

As shown in Fig. 3, the baseline model was found to be less than 50% accurate over the entire range. The accuracy was greater for the Jamo-static model than the baseline model. When we observed the 10 words that are most similar, the corrected word was included in more than half of the test words. The accuracy was dramatically increased for the Jamo-advanced model. In contrast to the most similar word in an OOV word exactly matched its corrected word with 17% accuracy for the baseline model; Jamo-advanced model shows 44% accuracy. When the number of words in the most similar range was increased to 10, the accuracy increased to 76%, which is approximately 2.3 times higher accuracy than the baseline model with 33% accuracy.

Table 2 shows the top 3 most similar words of the baseline model and Jamo-advanced model as an example with misspelled OOV words (bold text). In the vocabulary word, the baseline model and the Jamo-advanced model show similar words in top 3, while in the OOV words, the baseline model shows quite unrelated most similar words compare to Jamo-advanced model. Moreover, the corrected words are contained in the most similar words of top3. These results indicate that the misspelled OOV word was successfully represented the semantic relationship of the words as the original corrected word by using our Jamo-advanced model.

Table 2. Top 3 neighboring words based on cosine similarity for some example words that contain misspelled OOV words (bold text) in baseline model and Jamo-advanced model.

Input word	Model	Most similar words in range of top 3		
페이스북 (Facebook)	Baseline	트위터 (Twitter)	공식페이스북 (Official Facebook)	인스타그램 (Instagram)
	Jamo-advanced	트위터 (Twitter)	SNS (Social Media)	인스타그램 (Instagram)
월트디즈니 (Walt Disney)	Baseline	디즈니 (Disney)	디즈니툰 (DisneyToon)	디즈니주니어 (Disney Junior)
	Jamo-advanced	디즈니툰 (DisneyToon)	디즈니채널 (Disney Channel)	디즈니사 (Walt Disney Company)
타자 (Hitter)	Baseline	톱타자 (Lead-off man)	좌타자 (Left-handed hitter)	다음타자 (Next hitter)
	Jamo-advanced	우타자 (Right-handed hitter)	좌타자 (Left-handed hitter)	장타자 (Power hitter)
페널티 (Penalty) OOV word	Baseline	리날디 (Rinaldi)	페레티 (Ferretti)	마세티 (Machete)
	Jamo-advanced	페널티골 (Penalty goal)	페널티 (Penalty)	드록바 (Drogba)
나포탈렌 (Naphthalene) OOV word	Baseline	야렌 (Yaren)	콜루바라 (Kolubara)	몽클로아아라바카 (Moncloa-Aravaca)
	Jamo-advanced	나프탈렌 (Naphthalene)	테레프탈산 (Terephthalic acid)	아디pic산 (Adipic acid)
스테이크 (Steak) OOV word	Baseline	스탠포프 (Stanhope)	스탠纳드 (Stannard)	화이트스네이크 (White Snake)
	Jamo-advanced	롱테이크 (Long take)	비프스테이크 (Beefsteak)	스테이크 (Steak)

6 Conclusions and Further Works

We introduce word representations specialized in Korean using morpheme analysis and extra sub-characters (Jamo). The similarity test shows higher correlation than the baseline model; thus, our model is competitive for word representations with semantic relations. For misspelled OOV words, our model shows a significantly increased cosine similarity between misspelled words and corrected words. Moreover, among the 10 most similar misspelled OOV words, the corrected words are included almost 80% of the time. Therefore, our model successfully solved the OOV problem caused by misspellings in Korean. Furthermore, it is very promising that our model can be utilized as a spelling correction method. We will try to apply our model for spelling correction and utilize it as an input for NLP applications, such as NER or MRC. By applying our model, we expect that the performance of NLP applications will be increased.

Acknowledgements. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2013-0-00109, WiseKB: Big data based self-evolving knowledge base and reasoning platform).

References

1. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space, arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
2. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
3. Sienčnik, S.K.: Adapting word2vec to named entity recognition. In: Proceedings of the 20th Nordic Conference of Computational Linguistics, Nodalida 2015, May 11–13, 2015, Vilnius, Lithuania, pp. 239–243. Linköping University Electronic Press (2015)
4. Hu, M., Peng, Y., Qiu, X.: Reinforced mnemonic reader for machine comprehension. CoRR, abs/1705.02798 (2017)
5. Wieting, J., Bansal, M., Gimpel, K., Livescu, K.: Charagram: embedding words and sentences via character n-grams, arXiv preprint [arXiv:1607.02789](https://arxiv.org/abs/1607.02789) (2016)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information, arXiv preprint [arXiv:1607.04606](https://arxiv.org/abs/1607.04606) (2016)
7. Sampson, G.: Writing Systems. London (1985)
8. Choi, H., Kwon, H., Yoon, A.: Improving recall for context-sensitive spelling correction rules using conditional probability model with dynamic window sizes. J. KIISE **42**(5), 629–636 (2015)
9. Kang, S.-S., Kim, Y.T.: Syllable-based model for the Korean morphology. In: Proceedings of the 15th Conference on Computational Linguistics, vo. 1, pp. 221–226. Association for Computational Linguistics (1994)
10. Stratos, K.: A Sub-character architecture for Korean language processing, arXiv preprint [arXiv:1707.06341](https://arxiv.org/abs/1707.06341) (2017)
11. Luong, T., Socher, R., Manning, C.: Better word representations with recursive neural networks for morphology. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning, pp. 104–113 (2013)
12. Botha, J., Blunsom, P.: Compositional morphology for word representations and language modelling. In: International Conference on Machine Learning, pp. 1899–1907 (2014)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
14. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification, arXiv preprint [arXiv:1607.01759](https://arxiv.org/abs/1607.01759) (2016)
15. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext. zip: Compressing text classification models, arXiv preprint [arXiv:1612.03651](https://arxiv.org/abs/1612.03651) (2016)
16. Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., Ruppin, E.: Placing search in context: the concept revisited. In: Proceedings of the 10th International Conference on World Wide Web, pp. 406–414. ACM, New York (2001)



A Regressive Convolution Neural Network and Support Vector Regression Model for Electricity Consumption Forecasting

Youshan Zhang^{1(✉)} and Qi Li²

¹ Computer Science and Engineering, Lehigh University,
Bethlehem, PA 18015, USA

yoz217@lehigh.edu

² Department of Automation, BOHAI University,
Jinzhou 121013, Liaoning, China
liqi199507@gmail.com

Abstract. Electricity consumption forecasting has important implications for the mineral companies on guiding quarterly work, normal power system operation, and the management. However, electricity consumption prediction for the mineral company is difficult since electricity consumption can be affected by various factors. The problem is non-trivial due to three major challenges for traditional methods: insufficient training data, high computational cost and low prediction accuracy. To tackle these challenges, we firstly propose a Regressive Convolution Neural Network (RCNN) model, but RCNN still suffers from high computation overhead. Then we utilize RCNN to extract features from data and Regressive Support Vector Machine (SVR) trained with features to predict the electricity consumption. The experimental results show that RCNN-SVR model achieves higher accuracy than using the traditional RCNN or SVM alone. The MSE, MAPE, and CV-RMSE of RCNN-SVR model are 0.8564, 1.975, and 0.0687% respectively, which illustrates the low predicting error rate of the proposed model.

Keywords: Electricity consumption forecasting · Regression convolution neural network · Support vector machine

1 Introduction

The electricity consumption of large enterprises has been a major factor of the cost control and the operational efficiency. Specifically, mineral companies consume large quantities of electricity in the coal production process daily. The electricity consumption forecasting has important implications for the mineral companies on guiding quarterly work, the normal power system operation and power management. Besides, the prediction accuracy of electricity consumption

directly determines the power construction, network planning and the planning of electricity marketing strategies [1–4]. Therefore, predicting the electricity consumption accurately is demanded and crucial to mineral companies.

Since the complicated dynamic of the electrical power system, it is difficult to establish an explicit model. Many traditional methods are applied to predict the electricity consumption, such as Gray prediction, regression analysis, time series, artificial neural network (ANN), support vector machine (SVM) [3, 5–9]. However, these methods have their respective disadvantages. For example, traditional ANN train data are mostly based on the gradient, and it may fail into local minimum easily [4]. One common limitation is that these methods are strongly depended on the number of training data, which discover the relationship between predictive value and model. Also, some statistical analysis models such as Kalman filters, and Autoregressive Integrated Moving Average (ARIMA) [10–12] were also applied in electricity consumption prediction. However, they still have constraints of insufficient data size. In [13], Hu presented a neural-network-based gray prediction (NNGM(1,1)) method, which can overcome the limitation of the traditional gray prediction method. It can easily determine the developing coefficient and control variables in the gray prediction model. Therefore, NNGM(1,1) can improve load forecasting accuracy. Similarly, in [2], Song et al. modified the gray prediction method and proposed a rolling gray prediction (NOGM(1,1)) model. Ding et al. [2] overcame the deficiencies of fixed structure and poor adaptability in the original gray prediction model. The empirical results showed the NOGM(1,1) model has higher prediction accuracy than original gray prediction model. However, the prediction accuracies of these methods are still not satisfying.

The major challenge is that electricity consumption prediction of the mineral company is different from the traditional electricity load prediction since mineral company electricity consumption is affected by various factors (e.g., ore grade, processing quantity of the crude ore, Ball milling fill rate). Conventional methods only consider the electricity values and ignore the influential factors. Therefore, it is necessary to build a new model that not only considers electricity values and influential factors but predicts the monthly electricity consumption of mineral company. In this paper, we will solve three issues by our proposed electricity consumption prediction model: (1) reduce the computational cost; (2) train the model with limited data; and (3) improve the prediction accuracy. Convolution Neural Network (CNN) [14–16] has become a popular method for solving image classification, segmentation, and regression problem recently. However, there is no such a Regressive CNN (namely RCNN, ending with a regression layer) architecture for predicting electricity consumption of mineral company.

In this study, we present a new electricity consumption forecasting model based on regressive convolution neural network and support vector regression (RCNN-SVR). Compared with traditional methods, the RCNN model is capable of extracting more representative features of history electricity consumption data, while SVR model can reduce the computation overhead. The forecasting accuracy of the proposed model is higher than several baseline models such as BP

neural network and SVM [3]. There are two major contributions of this paper: (1) build the RCNN-SVR architecture to predict the electricity consumption of electricity; (2) compare prediction performances of our model with several baseline forecasting methods. We describe the RCNN and the SVR model, and introduce the model architecture in Sect. 2. Experiments are conducted to verify our model and the comparisons with previous methods are available in Sect. 3. Based on results in Sect. 3, we discuss the experimental results, make a conclusion and explore future work in Sect. 5.

2 Methodologies

In this section, we first introduce the regressive convolution neural network (RCNN), and support vector regression (SVR) model, separately. Then, we present our RCNN-SVR architecture for predicting electricity consumption.

2.1 Data Prepossessing

The electricity consumption data was collected from a mineral company in Liaoning province, China. It contained the monthly electricity consumption from 2012 to 2017 (only two months data are provided in 2017) with total 62 months. We split the data into training data and testing data. Testing data are not used during the training process. Training data contain 8×50 influential factors (IFs) 8 is eight IFs of each month, and 50 is the number of month. 50×1 true electricity consumption values (EVs). Testing data contain 8×12 IFs, 12×1 true EVs. For the input for RCNN, and RCNN-SVR model, we reshape the influential factors into a 4-D array, for example, influential factors change into $8 \times 1 \times 1 \times 50$ for training and testing dataset, 8, 1 and 1 represents for length, height, and depth.

2.2 RCNN Architecture

We first propose a regressive convolution neural network model (RCNN, shows in Fig. 1), which is similar to DeepEnergy in [17]. But our RCNN model has fewer layers because of limited data, the input is influential factors (IFs), and the last layer is regression layer which represents the electricity consumption values (EVs). In this network, it contains two main steps: feature extraction, and prediction. It only has eight layers. The feature extraction is performed by two convolution layers (Conv1, Conv2), and two max-pooling layer, (Maxpool1, Maxpool2), one rectified linear units (ReLU) layer, and one normalization (Norm) layer. The prediction step consists of a fully-connected layer and a regression layer. The input layer is comprised of 8×1 influential factors (one month), Conv1 and Conv2 have the filter size (F) of 1×1 , and filter number (N) 25 with padding size (P) 0; Maxpool1 and Maxpool2 have the stride size (S) of 2×2 . Therefore, after the max-pooling layer, the dimension of feature map is divided by 2. The ReLU layer reduces the number of epochs to achieve the training error rate greater than traditional tanh units. The normalization layer increases

generalization and reduces the error rate. Also, ReLU and normalization layer does not change the size of the feature map. The pooling layers summarize the outputs of adjacent pooling units.

One of the most obvious merits of RCNN is more features can be extracted from different layers. With more features, we can easily build the relationship between model and the predicted value. For example, if the input size (I) is 8×1 , we assume that feature map size is 8×1 . In Conv layer, the feature map size can be calculated as: $((I - F + 2P)/S + 1) \times N$. And feature map size is equal to $I/S \times N$ in max-pooling layers. In the Conv1 layer, the feature map size is: $((8 - 1)/1 + 1) \times 20 = 8 \times 25$; the feature map size in Maxpool1 is $(8/2) \times 25 = 4 \times 25$. Again, the feature size is: $((4 - 1)/1 + 1) \times 25 = 4 \times 25$ in the Maxpool2 layer, and feature size becomes: $(4/2) \times 25 = 2 \times 25$ in the Maxpool2 layer. The total number of features is increased (50 in maxpooling2 layer vs. 8 in input layer), and this is one reason why RCNN can generate a better-predicted result than other neural networks which only use input data as feature map.

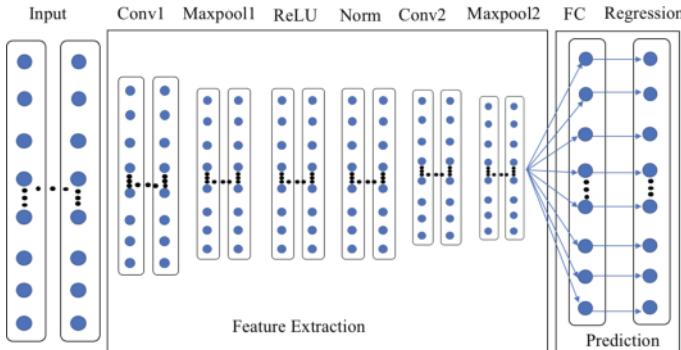


Fig. 1. The RCNN structure, it contains only eight layers (to prevent the overfitting of limited data), the input layer is the: influential factors. And the regression layer generates electricity consumption values. In the training stage, the RCNN will extract features of the influential factors, and check if the MSE is convergent. By using the trained RCNN classifier, we can predict the electricity consumption values of test data. (Conv: convolution layer, NA: normalization layer, FC: fully-connection layer)

Electricity Prediction Using RCNN As shown in Fig. 1, with more features extracted in the Maxpool2 layer, we will connect it into FC layer and flat all features into one dimension. In the training stage, the input size is: 8×50 . The size of the fully-connected layer is 50×1 , and it has the same size as the regression layer, and this why points in FC layer are only connected to one point in regression layer. During the training process, if the desired Mean Square Error (MSE) is not reached in the current epoch, the training will continue until the maximal number of epochs or desired MSE is reached. On the contrary, if the

maximal number of epochs is reached, then the training process will stop regardless the MSE value. Final performances are evaluated to demonstrate feasibility and practicability of the proposed method. During the test stage, we input the test data set 8×12 , and by using the training RCNN model, we can predict the electricity consumption of each month.

2.3 SVR

The original linear support vector machine (SVM) is proposed for binary classification problem. Given data and its labels: (x_n, y_n) , $n = 1, \dots, N$, and $y_n \in \{-1, +1\}$. It aims to optimize following equation:

$$\begin{aligned} \min_{w,b} \frac{1}{2} \|w\|^2 + \lambda \sum_n \xi_n^2 & \quad \text{s.t.} \\ y_n(w^T x_n + b) \geq 1 - \xi_n & \quad (\forall n), \quad \xi_n \geq 0 \quad (\forall n), \end{aligned} \quad (1)$$

where λ controls the width of margin (smaller margin with smaller λ); ξ_n is a non-negative slack variable and penalizes data points which against the margin; b is the bias.

Linear SVM can also be used as a regression method (called SVR), there are few minor differences comparing with SVM for classification problem. First of all, the output of SVR is a continuous number, but not the classes in the classification problem. Besides, there is a margin of tolerance ε in the SVR. However, the main idea is always the same: minimize the error and maximize the margin. Figure 2 describes the one-dimensional SVR, it aims to optimize following constrained function:

$$\begin{aligned} \min_{w,b} \frac{1}{2} \|w\|^2 + \lambda \sum_n (\xi_n + \xi_n^*) & \quad \text{s.t.} \quad y_n - (w^T x_n + b) \leq \varepsilon + \xi_n \\ y_n - (w^T x_n + b) \leq \varepsilon + \xi_n^* & \quad (\forall n), \quad \xi_n, \xi_n^* \geq 0 \quad (\forall n), \end{aligned} \quad (2)$$

where ξ_n^* is another non-negative slack variable [18–20].

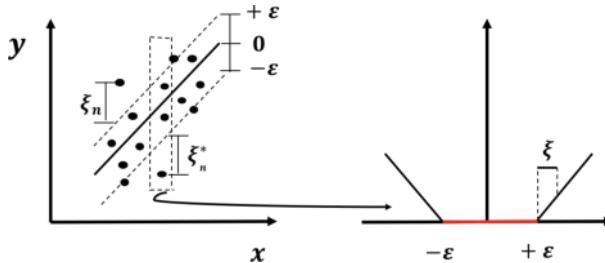


Fig. 2. The one-dimensional linear SVR with ε intensive band. Only the points which is out of $+\varepsilon$ and $-\varepsilon$ bound contribute to the cost. The dashed rectangle can be alternative visualized in the left image. Figure is modified from [18]

Electricity Prediction Using SVR To apply SVR method in predicting the electricity consumption of mineral company, we use SVR classifier to train the eight factors and predict the electricity consumption using the trained classifier. The SVR structure is shown in Fig. 3. In training stage, we train the SVR classifier using IFs, and we compare the predicting electricity value with true EVs and check whether the model is convergent; if not, the training stage will execute again. During test stage, we use IFs from test data set and predict electricity value.

2.4 RCNN-SVR

Inspired by the RCNN and SVR, we combine the deep neural network with SVR and design an RCNN-SVR model. Specifically, we train SVR classifier using features, which extracted from RCNN, then predict the electricity consumption using trained SVR classifier. Different from above RCNN architecture, we add more layers in the RCNN part to get more useful features. The RCNN-SVR architecture is shown in Fig. 4. Different from single RCNN and SVR model, RCNN-SVR combines the advantages of these two methods. RCNN-SVR can extract more features and use the less computational time to train the model. In our RCNN-SVR model, it also contains two steps: the feature extraction step is from RCNN model, and predicting step is from SVR model. Also, to extract the features, we fine-tuned the network. Different from the number of layers in RCNN, we add another Conv3 and Maxpool3 layer. To reduce error and prevent the overfilling, we use the drop out strategies, which adds a dropoutlayer after the Maxpool3 layer. For three Conv layers, the fitter size is 1×1 , and the filter number are: 20, 25, and 50, respectively. For three Maxpool layers, the stride size is 2×2 . Besides, we removed the last two layers (FC and regression layer), since we could not extract significant features from these two layers. The feature size of last dropout layer is the same as the feature map size in the Maxpool2 layer of RCNN. But the feature map is different; there is more information in feature map of RCNN-SVR model. As shown in Figs. 5 and 6, the feature map of RCNN-SVR model (both training and testing data) has more features than 8 layers RCNN model in Sect. 2.2. With more features extracted in RCNN model, it will provide enough information for SVR model to train the features. Further, we can build a better relationship between features and actually electricity consumption values.

Electricity Prediction Using RCNN-SVR To apply the RCNN-SVR model in predicting the electricity consumption of mineral company, we use RCNN to extract features of eight IFs and predict the electricity consumption using the trained SVR classifier. The RCNN-SVR structure is shown in Fig. 4. As shown in Algorithm 1, in training stage, we train the SVR classifier using features from RCNN model, and we compare the predict electricity value with true EVs and check whether the model is convergent. If not, the training stage will execute again. During test stage, we use IFs from test data set and predict electricity consumption values.

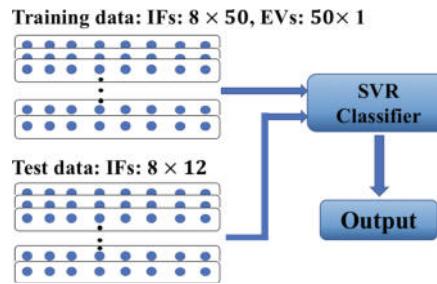


Fig. 3. The SVR structure, IFs: influential factors (eight factors with fifty months), EVs: electricity values. In the training stage, the SVR classifier trains the IFs, and check if the model is convergent, if not, then SVR model will train again. We can predict the electricity values using the trained SVR classifier

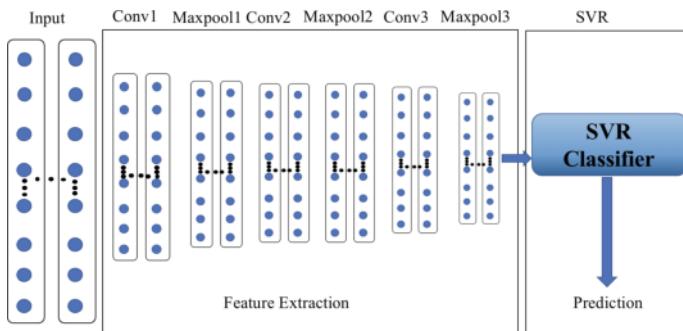


Fig. 4. The RCNN-SVR structure, which combines partial model from RCNN and SVR. Different from RCNN and SVR, RCNN-SVR extract features of data with more layers, and it trains SVR with more extracted features. By using RCNN-SVR model, we can extract more useful information and use less time to train the SVR classifier, and predict the electricity consumption values

Algorithm 1: Electricity prediction using RCNN-SVR model

Training Stage:

Input data: influential factors: 8×50 , and electricity values: 1×50
 Extract the features from RCNN, and train SVR classifier

Testing Stage:

Input data: influential factors: 8×12
 Predict electricity values of testing data using SVR classifier from testing stage
 Calculate the accuracy of predicted results

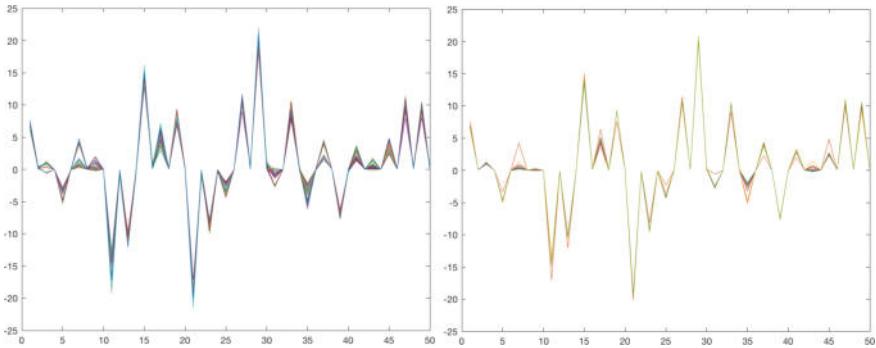


Fig. 5. The feature maps of training data (left one) and testing data (right one) from RCNN model. The x-axis is the number of features and y-axis is the range of features. These features are extracted from the maxpooling2 layer in RCNN model

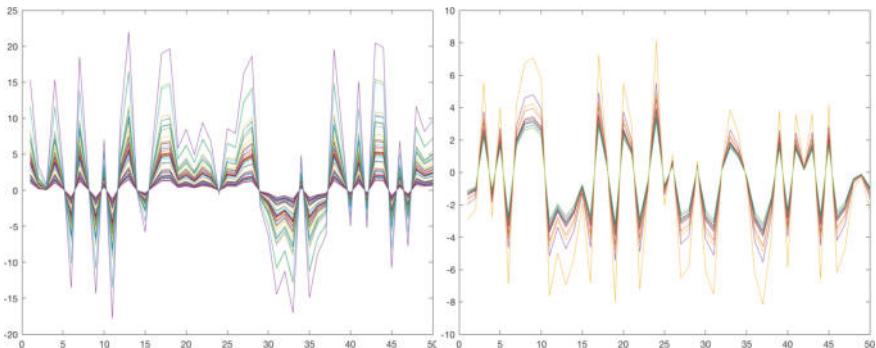


Fig. 6. The feature maps of training data (top one) and testing data (bottom one) from RCNN-SVM model. The x-axis is the number of features and y-axis is the range of features. We train the features of training data, and predict the electricity consumption of test data by using the extracted test features. Features are extracted from the maxpooling3 layer in RCNN-SVR model. There are 50 lines in the top image which are corresponding to the number of training data, and there are 12 lines in bottom image, which are corresponding to the number of testing data. Each line has different shapes which illustrate the difference of data. We could find that RCNN-SVM model have more features than RCNN model. And this can be a reason that predicting results of RCNN-SVR model are better than RCNN model

3 Results

In the experiment, we use data which are provided by a mineral company. Besides, the training data are the electricity consumption values of past 50 months, and the test data are 12 months electricity consumption values. The data were processed in Sect. 2.1. Figures 7 and 8 is the comparison predicting result of RCNN-SVR model with RCNN, SVR, MPSO-BP, and DeepEnergy. In

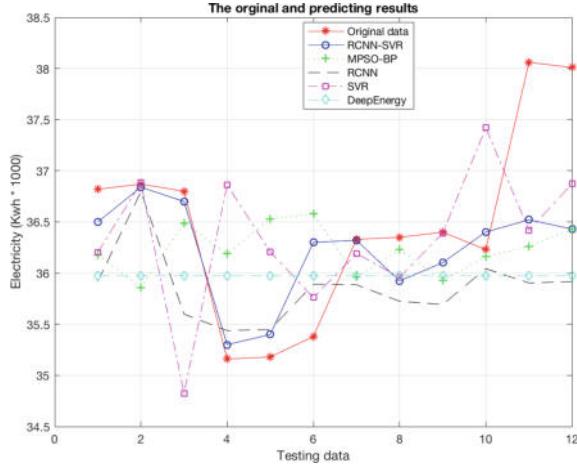


Fig. 7. The comparison results of predicting electricity values using RCNN-SVR, RCNN, SVR, MPSO-BP and DeepEnergy

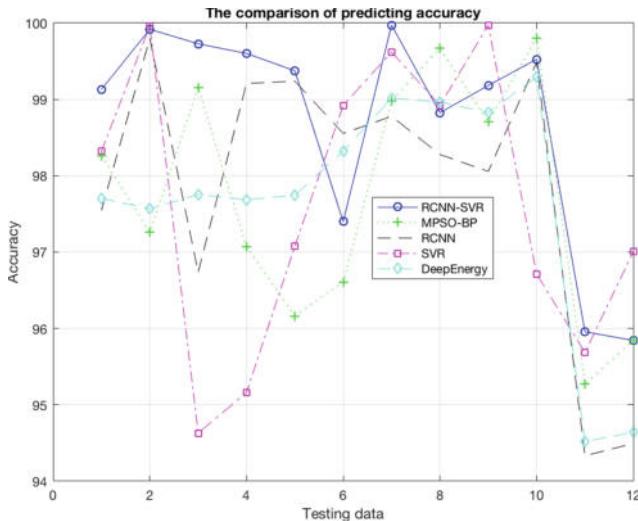


Fig. 8. The comparison results of predicting accuracy using RCNN-SVR, RCNN, SVR, MPSO-BP and DeepEnergy. SVR model is more likely to oscillate in the image which means the predicting results are not stable. And RCNN-SVR is more closed to the true data. Although DeepEnergy network have better accuracy in Fig. 8, but its actually predicted values not change with different testing data and this is caused by too many layers in the DeepEnergy network with limited training data (in Fig. 7)

Fig. 7, the vertical axes represent the electricity consumption (kWh), and the horizontal axes denote different test months. According to the results in Fig. 8, RCNN-SVR model has the highest accuracy among all models.

3.1 Evaluation of Model Accuracy

To evaluate the performance of predicting results, we employ three evaluation functions: Mean Standard Error (MSE), Mean Absolution Percentage Error (MAPE) and Cumulative Variation of Root Mean Square Error (CV-RMSE) [17]. And these evaluation functions are defined in Eq. (3), where y_i is the true electricity value, \hat{y}_i is the predicting value, N represents the data size.

$$\text{MSE} = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}, \text{MAPE} = \frac{1}{N} \sum_{i=1}^N \left| \frac{(y_i - \hat{y}_i)}{y_i} \right|,$$

$$\text{CV-RMSE} = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N \left(\frac{(y_i - \hat{y}_i)}{y_i} \right)^2}}{\frac{1}{N} \sum_{i=1}^N y_i} \quad (3)$$

The comparison results of four methods are shown in Table 1. As shown in Table 1, the MAPE and CV-RMSE of the RCNN-SVR model are the smallest, and the goodness of error is the best among all models, namely, MSE, average MAPE and CV-RMSE are 0.8564, 1.975 and 0.0687%, respectively. The MAPE of SVR model is the largest among all of the models; an average error is about 2.3341%. On the other hand, the CV-RMSE of SVR model is the largest among all models; an average error is about 0.0809%. According to the MSE, average MAPE and CV-RMSE values, the electricity consumption forecasting accuracy of tested models in descending order is as follows: RCNN-SVR, RCNN, MPSO-BP, DeepEnergy, and SVR. However, SVR uses less time than other models (1.82 s), comparing with the rest three methods, our model RCNN-SVR uses relatively less time than RCNN, SVR and DeepEnergy methods.

Table 1. Predicting results of the electricity consumption by different methods

	RCNN-SVR	RCNN	SVR	MPSO-BP [4]	DeepEnergy [17]
MSE	0.8564	1.0690	1.1639	0.9236	1.0720
MAPE (%)	1.975	2.1239	2.3341	2.2665	2.330
CV-RMSE (%)	0.0687	0.0755	0.0809	0.0745	0.0760
Time (s)	4.35	221.74	1.82	27.21	3758

From Table 1, we can find that our RCNN-SVR model has the smallest MSE, MAPE, and CV-RMSE, which means our model has the highest accuracy than other methods. Therefore, the RCNN-SVR model is the most suitable method for electricity predicting. We recommend using the RCNN-SVR model to predict the electricity consumption of mineral company.

3.2 Forecasting of Electricity Consumption of Each Month in 2018

Using the trained RCNN-SVR model, we predict the electricity consumption values of each month in 2018, as shown in Table 2. The electricity consumption

will increase in November and December, and this may cause by heavy pressure on the operation, maintenance, and supply heating of power system.

Table 2. Forecasting electricity consumption (kWh) of each month in 2018

Months	1	2	3	4	5	6
Evs (kwh/t)	38.78	38.56	37.01	36.83	35.76	35.94
Months	7	8	9	10	11	12
Evs (kwh/t)	37.33	36.84	36.76	36.06	37.78	38.28

4 Discussion

The traditional method, such as SVR, BP neural network has been applied in electricity consumption prediction. In this paper, these methods also provided a reasonable result (as shown in Table 1). Regarding SVR, the results are worst among these methods. One reason is that there are no enough features can be trained due to the limited data. According to the Table 1, the RCNN has a relative long computational time, and this is caused by the features extraction and training step. Our RCNN-SVR model has the lowest MSE, MAPE, and CV-RMSE comparing with other methods. Furthermore, the selection of extracting features from which layer in RCNN-SVR model is important, as shown in Fig. 9,

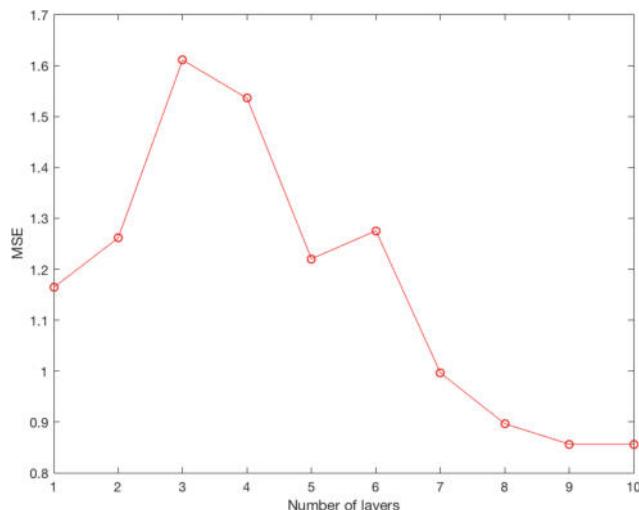


Fig. 9. The relationship between MSE and number of layer selected in RCNN-SVR model

MSE is overall reduced with the selected later layers. And this implies that the most useful features are shown in the last layers in the RCNN-SVR model. Therefore, we may get better results if we use the features from the last layer.

5 Conclusions

In this paper, we propose a regressive convolution neural network and support vector regression (RCNN-SVR) model for electricity consumption forecasting. The proposed model is validated by experiment with the electricity consumption data from the past five years. In the experiment, the data from a mineral company were used, and historical electricity demands are considered. According to the experimental results, the RCNN-SVR model can precisely predict electricity consumption in the next following months. Also, the proposed model is compared with four models that were used in electricity consumption forecasting. The comparison results showed that performance of our RCNN-SVR model is the best among all tested algorithms, which has the lowest values of MSE, MAPE, and CV-RMSE. According to all of the obtained results, the proposed method can reduce computation time. The proposed RCNN-SVR method successfully solves three issues which are mentioned above: (1) reduce the computational cost; (2) train the model with limited data; and (3) improve the prediction accuracy. Therefore, the RCNN-SVR model can be used to predict the electricity consumption of mineral company.

However, our paper has the limitation of data size. For future work, we will first test our model use more data, then we will expand the different neural networks, such as DenseNet, Adversarial neural network to extract the features of data. What's more, the novel model in this paper can be used in predicting electricity values in other fields, such as wind power generation system electricity prediction, and agricultural electricity consumption area.

References

1. Kavousi-Fard, A., Samet, H., Marzbani, F.: A new hybrid modified firefly algorithm and support vector regression model for accurate short term load forecasting. *Expert Syst. Appl.* **41**(13), 6047–6056 (2014)
2. Ding, S., Hipel, K.W., Dang, Y.: Forecasting China's electricity consumption using a new grey prediction model. *Energy* **149**, 314–328 (2018)
3. Kaytez, F., Taplamacioglu, M.C., Cam, E., Hardalac, F.: Forecasting electricity consumption: a comparison of regression analysis, neural networks and least squares support vector machines. *Int. J. Electr. Power Energy Syst.* **67**, 431–438 (2015)
4. Zhang, Y., Guo, L., Li, Q., Li, J.: Electricity consumption forecasting method based on MPSO-BP neural network model. In: Proceedings of the 2016 4th International Conference on Electrical & Electronics Engineering and Computer Science (ICEEECS 2016), vol. 50, pp. 674–678 (2016)
5. Akay, D., Atak, M.: Grey prediction with rolling mechanism for electricity demand forecasting of Turkey. *Energy* **32**(9), 1670–1675 (2007)

6. Bianco, V., Manca, O., Nardini, S.: Electricity consumption forecasting in Italy using linear regression models. *Energy* **34**(9), 1413–1421 (2009)
7. Abdel-Aal, R.E., Al-Garni, A.Z.: Forecasting monthly electric energy consumption in Eastern Saudi Arabia using univariate time-series analysis. *Energy* **22**(11), 1059–1069 (1997)
8. Ekonomou, L.: Greek long-term energy consumption prediction using artificial neural networks. *Energy* **35**(2), 512–517 (2010)
9. Wang, S., Yu, L., Tang, L., Wang, S.: A novel seasonal decomposition based least squares support vector regression ensemble learning approach for hydropower consumption forecasting in China. *Energy* **36**(11), 6542–6554 (2011)
10. Yuan, C., Liu, S., Fang, Z.: Comparison of China's primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM (1, 1) model. *Energy* **100**, 384–390 (2016)
11. Soubdhan, T., Ndong, J., Ould-Baba, H., Do, M.-T.: A robust forecasting framework based on the Kalman filtering approach with a twofold parameter tuning procedure: application to solar and photovoltaic prediction. *Solar Energy* **131**, 246–259 (2016)
12. Al-Hamadi, H.M., Soliman, S.A.: Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model. *Electr. Power Syst. Res.* **68**(1), 47–59 (2004)
13. Hu, Y.-C.: Electricity consumption prediction using a neural-network-based grey forecasting approach. *J. Oper. Res. Soc.* **68**(10), 1259–1264 (2017)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440 (2015)
16. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. [arXiv:1404.2188](https://arxiv.org/abs/1404.2188) (2014)
17. Kuo, P.-H., Huang, C.-J.: A high precision artificial neural networks model for short-term energy load forecasting. *Energies* **11**(1), 213 (2018)
18. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. *Stat. Comput.* **14**(3), 199–222 (2004)
19. Basak, D., Pal, S., Patranabis, D.C.: Support vector regression. *Neural Inf. Process.-Lett. Rev.* **11**(10), 203–224 (2007)
20. Tang, Y.: Deep learning using linear support vector machines. [arXiv:1306.0239](https://arxiv.org/abs/1306.0239) (2013)



Facial Expression Recognition and Analysis of Interclass False Positives Using CNN

Junaid Baber^{1(✉)}, Maheen Bakhtyar¹, Kafil Uddin Ahmed¹,
Waheed Noor¹, Varsha Devi², and Abdul Sammad³

¹ Department of CS and IT, University of Balochistan, Quetta, Pakistan
junaidbabber@ieee.org, maheen.bakhtyar@gmail.com, kafil.fast09@gmail.com,
waheed.noor@um.uob.edu.pk

² University of Grenoble Alpes, Grenoble, France
varsha.devi@etu.univ-grenoble-alpes.fr

³ Department of Computer Science, Habib University, Karachi, Pakistan
abdul.samad@sse.habib.edu.pk

Abstract. In this paper, the performance of Facial Expression Recognition (FER) using Deep Convolutional Neural Network (DCNN) model is evaluated. The expressions include *Angry*, *Disgust*, *Fear*, *Happy*, *Sad*, *Surprise*, and *Neutral*. In addition to performance evaluation, the analysis on Interclass false positives are also discussed which helps to analyze the underlying challenges to improve the model. All classifiers give low performance on Fer2013 datasets. DCNN gives 54.46% accuracy on test and 89.52% on training set, whereas, in case of different kernels of Support Vector Machines (SVM), the highest accuracy is 45% using cubic kernel on training set. Experiments show that certain facial expressions have more false positives and few of them are very dominant. In case of *Disgust* expression, it has *Angry* as dominant false positive. Based on false positives analysis, binary classifiers can be trained to improve the accuracy of the expressions with dominant false positives. Experiments on Fer2013 dataset confirms that the accuracy of expression *Disgust* is improved by piping the binary classifiers, *Disgust* vs. *Angry*, to existing DCNN of multi-class.

Keywords: DCNN · Facial expression recognition · Interclass analysis · Smart classroom

1 Introduction

Automatic recognition of human emotions and mental state is one of the important research problem in Human Computer Interaction (HCI). It is multidisciplinary research including a wide assortment of related fields, such as computer vision, audio/visual analysis, cognitive psychology and learning theories etc. Different types of signal are the input sources for better emotion recognition, which includes audio/video signals, texts, and bio-signals.

Visual cues and features can provide a lot of useful and reliable information for FER systems. FER based on visual contents are more robust and cheap compared to other source of signals. CCTV cameras are almost mounted in all universities, colleges, and organizations which provides monitoring and security, same CCTV signals can be utilized to analyze the facial expressions of the audience where CCTVs are mounted, i.e., classrooms, workshops, or any public and organized gathering of people.

FER in uncontrollable environment, such as evaluating the quality of lecture by students facial expressions in classroom, is still very challenging and need attention for better accuracy.

Many databases has been created for this research problem [1] and these datasets are collected in lab-controlled environments where the artificial expressions were generated by the subjects.

Only visual cues and information are focused in this paper. The input signal can be divided into three major steps, (1) Face Detection; (2) Feature Engineering, and (3) Expression Classification. Face detection is already very matured in literature therefore the main focus of research is on second and third steps, feature engineering and classifier, respectively. Number of features have been applied on facial expression recognition such as LBP-TOP [2], PHOG [3] and LPQ [4], but all handcrafted features give limited performance on small images [5]. The one of the main reason for poor performance is quality of the image, if the quality of image is better then the accuracy is better. As stated above, CCTV signals are mostly used to gather the faces, in this scenario, the faces are small and distorted. In case of Fer2013 dataset, dataset explain in experimental section, the input image is only 48×48 pixel of single channel (gray-scale image). In case of HOG based feature using SVM the best results on Fer2013 dataset is 45%. In case of high quality images, the results are far better. For example, JAFFE¹ dataset accuracy is $\approx 90\%$, the image size in JAFFE dataset is 256×256 pixels per image.

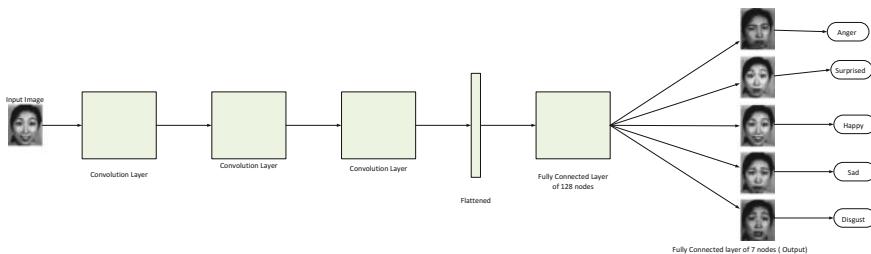


Fig. 1. Abstract network architecture of proposed DCNN. The configuration of filters and max pooling are shown in Table 1

Deep learning based FER have higher accuracy even on small images [5]. In this paper we have proposed DCNN model for FER on moderate hidden levels

¹ <http://www.kasrl.org/jaffe.html>.

which is easy to train on commodity hardware. The accuracy of proposed DCNN is evaluated on challenging dataset, Fer2013, false positives for each expression are also reported. Based on false positive analysis, the performances of FER can be increased, as discussed in experimental section.

Rest of our paper is organized as follows: Sect. 2 discusses the related work, Sect. 3 reports the proposed methodology of the DCNN along with explanation of dataset and evaluation matrices, Sect. 4 presents the results and discussion whereas conclusion is given in Sect. 5.

2 Related Work

Number of methods and approaches have been proposed and used for better expressions recognition on AFEW (Acted Facial Expressions in the Wild) dataset [6, 7]. Many authors used multiple kernel learning [8], multiple feature fusion [9], and score level fusion [10, 11]. In [12], zhong et al. analyzed and found that only few active patches of faces are useful for facial expression recognition and these patches can be find out by two stage multi-task sparse learning framework. These patches include common patches and specific patches. Common patches are for recognition of all kind of expressions while specific patches are used for the recognition of only one kind of expression. Common patches can be search out by using the multi-task learning while specific patches can be find out by using facial recognition. This sequential search raised the problem of searching of overlapped patches. To solve this problem, Liu et al. in [13] proposed a new technique by fusing two major techniques which are sparse vector machine and multi-task learning into one framework. Instead of using two different patches in two different phases for recognition of expressions, Liu et al. used a technique in which specific expression feature selection vector and common features selection vector are employed together. For more better facial expression classification and recognition, more discriminative features are used by [14], rather than using the hand-crafted features. Furthermore, recent work has been done by [15], in which the authors proposed a technique of fusing the two trained networks. These network were trained with facial landmarks and images. The two networks are deep temporal geometry network and deep temporal appearance network. To effectively join these two networks, a joint fine tuning strategy has been proposed. In [16], authors found that inception network architecture is good for classification of facial expressions. Authors performed multiple cross data experiments and showed the generality of learned model. Moreover Liu et al. in [17] applied 3D CNN to learn low level features from the videos, in order to utilize the temporal information from video based expressions recognition. After then, GMM (Gaussian Mixture Model) was trained on these features, and further covariance matrix for each component composes the expressionlet. Getting inspired from the information that facial expressions can be further decomposed into Active Units (AUs), [18] proposed a new model in which multiple filters are learned for detection of different facial part, known as deformable facial part model. To further cope with the pose and identity variations, a quadratic deformation cost is used.

Table 1. Number of convolution filters and their sizes used in each layer along with pooling size

Layer #	# Convolutional filter	Conv. size	Pooling size
Layer 1	32	3×3	3×3
Layer 2	64	3×3	3×3
Layer 3	128	3×3	3×3

Table 2. Ratio of each facial expression in test set and training set

Facial expression	Training set	Test set
Angry	3994 (13.91%)	492 (13.70%)
Disgust	436 (1.52%)	55 (1.53%)
Fear	4097 (14.27%)	528 (14.71%)
Happy	7215 (25.13%)	879 (24.48%)
Sad	4830 (16.83%)	594 (16.55%)
Surprise	3171 (11.05%)	416 (11.59%)
Neutral	4965 (17.29%)	626 (17.44%)

3 Proposed Deep Learning Model

Figure 1 shows the overview of proposed DCNN for FER which only contains 5-hidden layers. The network comprises of 3-Convolutional layers followed by max pooling layers, and there are two fully connected layers at the end. Convolution is widely used mathematical operation in image and signal processing, it is used for filtering, finding the patterns in the image (signal), detection of edges, and many other operations. In deep learning, Convolutional layers compute the output from given image and passes to next layers. The output of Convolutional layer is treated as feature maps which are obtained by sliding the $n_1 \times n_2$ window filter on the image in vertical and horizontal directional, mostly $n_1 = n_2$. The unit step of sliding is called the stride, \mathcal{S} , which is widely kept $\mathcal{S} \in \{1, 2\}$, pixels. Typically, more than 1 filters are used in each Convolutional layers, number of filters used for each layer in proposed DCNN are shown in Table 1. Before applying the pooling, feature map is passed to activation function.

There are number of activation functions, whereas, Leaky Rectified Linear Unit (Leaky ReLU) which is variation of simple ReLU, is used for the experiments.

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (1)$$

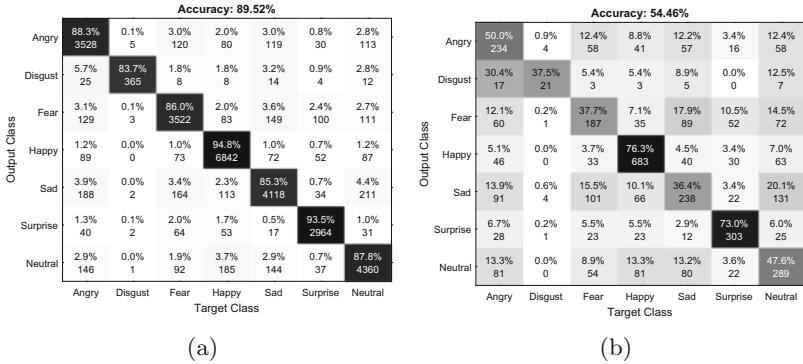
ReLU is widely used in deep learning as activation function. The main limitation of ReLU is that it behaves neutral on negative values. Sometimes, the negative values received by convolution have huge importance to the model. To

overcome the above mentioned problem, Leaky ReLU is used

$$\text{Leaky ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \frac{x}{\alpha} & \text{if } x < 0 \end{cases} \quad (2)$$

where $\alpha = 0.1$ in proposed DCNN.

Pooling is used for sub-sampling after convolution which reduces the dimension of feature maps that minimizes the over-fitting. During the experiments, stochastic pooling and max pooling are also evaluated, the difference between their accuracies were marginal, therefore, only max pooling results are reported in the paper.



(a)

(b)

Fig. 2. Accuracy of proposed DCNN on Fer2013 dataset. **a** shows the classification on training set, and **b** shows the classification on test test

After convolution layers, the feature map is flatten to single vector and passed to fully connected layers. The output of fully connected layers is computed by matrix multiplication, weighted matrix, followed by the offset bias.

The last layer of the network is set for classification which is softmax classification. Usually the last dense layer is scoring layer with real score for each class. The vector from the scoring layer is transform into a vector of values ranging from (0, 1) using softmax function. The softmax function for our 7 classes is given below:

$$p_i = \frac{e^{a_i}}{\sum_{k=1}^{N=7} e^{a_k}} \quad (3)$$

The above function outputs the probability distribution for scoring vector vector which makes more sense for classification task in probabilistic interpretation. Since now the output is the probability a cross entropy loss is used. It calculates the distance between the model believed probability distribution and the original distribution. It is defined as

$$H(y, p) = - \sum_i y_i \log(p_i) \quad (4)$$

4 Experiments and Results

Fer2013 dataset is used for experiments [19]. This is one of the challenging dataset since 2013. It has 7 facial expressions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Dataset is divided into two sets, training set and test set. The test set is further divided into private and public testing. Training set contains total 28,709 instances against all expressions, whereas, the public test set contains 3590 instances. The ratio of each facial expression in test and training set are described in Table 2. It can be seen that some expression tags have very few training instances such as *Disgust* which is only the 1.5% of total instances which makes dataset imbalance. Since deep learning models are data greedy models therefore the recognition heavily depend on the number of training instances per class. It can be seen in Fig. 2 that accuracy is low for those expression whose training instances are marginally less. Deep learning models which give substantial improvement have mostly the big data for training.

Figure 2 shows the accuracy of proposed DCNN on training and test sets. The accuracy on training set is comparatively higher than the test set. It might be the case of over-fitting, but the fact cannot be ignored that dataset is also imbalance and small. During experiments, we trained the DCNN on different sizes of data, ranging from 40 to 100% of the training set. On each iteration the accuracy increases, from 33 to 54.46% on test set whereas the accuracy on training set is always between (69–89%), which confirms that if the dataset size is increased then the accuracy may also be increased. It can be seen in Table 2 and Fig. 2 that *Surprise* has maximum number of instances during training and it also gives highest accuracy compared to other facial expressions.

Figure 3 shows the ratio of true positives and false positives for each facial expression where false positives are in increasing orders. The expression *Disgust* has one of the minimum accuracy, 37.5%, on test set having expression *Angry* as false positive with accuracy 30.4%. Whereas, the expression *Angry* is also the one of the dominant false positive for all other expressions, on average 13.6% times classified as false positives against other expressions. An other expression, *Neutral*, also dominantly decreases the performance of the DCNN, on average 12% times other expressions are classified as *Neutral*. Based on current false positive analysis, we trained two binary classifiers for *Angry* and *Disgust* and piped next to proposed DCNN, whenever the proposed DCNN outputs *Angry* we pass same image to our binary classifier to check either it is *Disgust* or *Angry*. The binary classifier is trained on same Fer2013 dataset. Just piping one binary classifier increases the accuracy of *Disgust* expression up to 63.6% without hurting the accuracies of other expressions.

Similar idea to improve the accuracies of particular class based on false positive analysis is proposed in Sindhi OCR [20]. The correlation between the shapes of different handwritten characters and template classes are computed, based on the correlation scores with template classes, the output class from template classes is chosen with maximum correlation of given handwritten character. The template classes are numerals from 0 to 9 generated from computer. The accuracy of correlation based classification is very low but false positive analysis on

the confusion matrix for 10 classes gives the a lot of information about the shapes of characters which are very similar and leads to miss-classification.

The same hypothesis is applied on the expressions of *Angry* and *Disgust* and the performance of *Disgust* is improved.

		Target Class							Target Class							
		Angry	Fear	Neutral	Sad	Happy	Surprise	Disgust	Sad		Neutral	Fear	Angry	Happy	Surprise	Disgust
Target Class	Angry	50.0%	12.4%	12.4%	12.2%	8.8%	3.4%	0.9%	36.4%	20.1%	15.5%	13.9%	10.1%	3.4%	0.6%	
		234	58	58	57	41	16	4	238	131	101	91	66	22	4	
Target Class	Disgust	Angry	Fear	Neutral	Sad	Happy	Surprise	Disgust	Sad		Neutral	Fear	Angry	Happy	Surprise	Disgust
		37.5%	30.4%	12.5%	8.9%	5.4%	5.3%	0.0%	73.0%	6.7%	6.0%	5.5%	5.5%	2.9%	0.2%	
Target Class	Disgust	21	17	7	5	3	3	0	303	28	25	23	23	12	1	
		Disgust	Angry	Neutral	Sad	Fear	Happy	Surprise	Surprise	Angry	Neutral	Fear	Happy	Sad	Disgust	
Target Class	Fear	37.7%	17.9%	14.5%	12.1%	10.5%	7.1%	0.2%	47.6%	13.3%	13.3%	13.2%	8.9%	3.6%	0.0%	
		187	89	72	60	52	35	1	289	81	80	54	22	0	0	
Target Class	Fear	Fear	Sad	Neutral	Angry	Surprise	Happy	Disgust	Neutral	Angry	Happy	Sad	Fear	Surprise	Disgust	
		76.3%	7.0%	5.1%	4.5%	3.7%	3.4%	0.0%	Neutral	Angry	Happy	Sad	Fear	Surprise	Disgust	
Target Class	Happy	683	63	46	40	33	30	0	Happy	Neutral	Angry	Sad	Fear	Surprise	Disgust	
		Happy	Neutral	Angry	Sad	Fear	Surprise	Disgust	Target Class							

Fig. 3. False positives analysis. For each facial expression, the column 1 is the true positive of given class, and rest all other are false positives. Some expressions are mostly misleading to the classifier such as *Angry*. Most of the expressions are classified as *Angry*

5 Conclusion

In this paper we have proposed simple and small DCNN. We have reported the results of facial expression recognition on Fer2013 dataset, which is quite challenging dataset in literature. The proposed DCNN achieves 89.52% accuracy on training set which comprises of 28,709 instances for seven different facial expressions, and 54.46% accuracy on test set which comprises of 3590 instances. We also reported the false positives for each class to analyze the reasons of miss-classification. Some expressions, such as *Neutral*, have significantly increased the false positives for other classes such as *Sad* which was classified neutral 20% times. The expression *Disgust* accuracy is very low due to the expression *Angry* which was reported false positive 30.4% times for *Disgust*. On the basis of above said analysis, binary classifier on *Disgust* and *Angry* is trained on same dataset and DCNN. For given any expression, if the classifier outputs *Angry* then it passed to the binary classifier to further confirmation. This improves the accuracy of *Disgust* up to 63.6% without hurting the overall accuracies of *Angry* and other expressions. In future we are interested to propose extended framework which should improve the performance of facial expression based on the analysis of false positives.

Acknowledgements. This research is supported by University of Balochistan under the project UBRF with grant number UOB/ORIC/17/UBRF-17/022, and Higher Education Commission of Pakistan (HEC).

References

1. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101. IEEE, New York (2010)
2. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 915–928 (2007)
3. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) *Image and Signal Processing*, pp. 236–243 (2008)
4. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, pp. 401–408 (2007)
5. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435–442 (2015)
6. Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion recognition in the wild challenge 2013. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 509–516. ACM, New York (2013)
7. Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion recognition in the wild challenge 2014: baseline, data and protocol. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 461–466. ACM, New York (2014)
8. Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., Bartlett, M.: Multiple kernel learning for emotion recognition in the wild. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 517–524. ACM, New York (2013)
9. Liu, M., Wang, R., Huang, Z., Shan, S., Chen, X.: Partial least squares regression on Grassmannian manifold for emotion recognition. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, pp. 525–530. ACM, New York (2013)
10. Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., Chen, X.: Combining multiple kernel methods on Riemannian manifold for emotion recognition in the wild. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 494–501. ACM, New York (2014)
11. Sun, B., Li, L., Zuo, T., Chen, Y., Zhou, G., Wu, X.: Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 481–486. ACM, New York (2014)
12. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2562–2569. IEEE, New York (2012)
13. Liu, P., Zhou, J.T., Tsang, I.W.-H., Meng, Z., Han, S., Tong, Y.: Feature disentangling machine-a novel approach of feature selection and disentangling in facial expression analysis. In: European Conference on Computer Vision, pp. 151–166. Springer, Berlin (2014)

14. Liu, M., Li, S., Shan, S., Chen, X.: Au-aware deep networks for facial expression recognition. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–6. IEEE, New York (2013)
15. Jung, H., Lee, S., Park, S., Lee, I., Ahn, C., Kim, J.: Deep temporal appearance-geometry network for facial expression recognition. arXiv preprint [arXiv:1503.01532](https://arxiv.org/abs/1503.01532)
16. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE, New York (2016)
17. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1749–1756. IEEE, New York (2014)
18. Liu, M., Li, S., Shan, S., Wang, R., Chen, X.: Deeply learning deformable facial action parts model for dynamic expression analysis. In: Asian Conference on Computer Vision, pp. 143–157. Springer, Berlin (2014)
19. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., Bengio, Y.: Challenges in representation learning: a report on three machine learning contests. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) Neural Information Processing, pp. 117–124 (2013)
20. Sanjrani, A.A., Baber, J., Bakhtyar, M., Noor, W., Khalid, M.: Handwritten optical character recognition system for Sindhi numerals. In: 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube), pp. 262–267 (2016)



CASCADENET: An LSTM Based Deep Learning Model for Automated ICD-10 Coding

Sheikh Shams Azam¹(✉), Manoj Raju¹,
Venkatesh Pagidimarri¹, and Vamsi Chandra Kasivajjala²

¹ Foundation Inc., Marina Del Rey, San Francisco, CA 90292, USA
s.shams.official@gmail.com,
rmanuuu@gmail.com,
venki.460@gmail.com

² Healthcare Information and Management Systems Society, Bengaluru, India
vamsichandra@gmail.com

Abstract. In this paper, a cascading hierarchical architecture using LSTM is proposed for automatic mapping of ICD-10 codes from clinical documents. The fact that it becomes increasingly difficult to train a robust classifier as the number of classes (over 93k ICD-10 codes) grows, coupled with other challenges such as the variance in length, structure and context of the text data, and the lack of training data, puts this task among some of the hardest tasks of Machine Learning (ML) and Natural Language Processing (NLP). This work evaluates the performance of various methods on this task, which include basic techniques such as TF-IDF, inverted indexing using concept aggregation based on exhaustive Unified Medical Language System (UMLS) knowledge sources, as well as advanced methods such as SVM trained on a bag-of-words model, CNN and LSTM trained on distributed word embeddings. The effect of breaking down the problem into a hierarchy is also explored. Data used is an aggregate of ICD-10 long descriptions along with anonymised annotated training data provided by few of the private hospitals from India. A study of the above-mentioned techniques leads to the observation that hierarchical LSTM network outperforms other methods in terms of accuracy as well as micro and macro-averaged precision and recall scores on the held out data (or test data).

Keywords: Deep learning · Natural language processing · Medical coding

1 Introduction

Medical Coding is the process of classifying the clinical documents and health records according to standard coding conventions. Some of the popular coding conventions widely used are International Classification of Diseases (ICD)

[1], Current Procedural Terminology (CPT) [2], Logical Observation Identifiers Names and Codes (LOINC) [3] etc. ICD-10, the 10th revision of ICD codes, is a coding standard released by World Health Organization (WHO) for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or disease. There are a total of over 93k codes in this system as of the time of this work.

Annotation of unstructured clinical documents using ICD codes also marks the foundation for various subsequent analysis such as identification of statistical health trends globally, patient similarity analysis that can help in the prediction of disease onset, development of clinical decision support system (CDSS) etc. It is also traditionally used as the first step in insurance claim filing process.

Since the annotation of ICD codes from clinical notes requires significant amount of time and expert knowledge in the field of medicine, the task is prone to errors both in terms of incorrectly attributed codes as well as missed diagnoses due to the high number of documents that are to be processed manually by human coders.

In order to tackle the posed issues of accuracy and time consumption this study aims at evaluating various techniques for automating the process of ICD coding. The work leads to the conclusion that a cascaded hierarchical architecture of Long Short-Term Memory (LSTM) [4] where inference from each level cascades to the subsequent levels perform particularly well on this task. The state-of-the-art performance in this classification task is achieved by this proposed architecture.

The rest of the paper is organized as follows. Section 2 discusses the related work in the field. Section 3 explains the background concepts such as organization of ICD-10 data and UMLS etc. Section 4 reviews the optimization objectives for the classifiers. Section 5 explains the architecture of the system, preprocessing and training strategies to achieve the best results. Section 6 reports the experimental results and discussions based on evaluation metrics. Finally, Sect. 7 concludes the work and discusses the limitations and future scope of this work.

2 Related Work

The number of research works in the field of automatic classification of clinical text is quite large. But a major portion of these publications is devoted to either classification of a carefully selected subset of codes or classification as per ICD-9 conventions.

Farkas and Szarvas [5] presented automated construction of ICD-9-CM codes using hand-crafted rule based system. They use preprocessing techniques such as lemmatization, punctuation-removal, removal of negated diagnoses (using manually collected indicator words such as can, may etc.) to normalize the data. They experimented with multi-label classification using binary relevance [6] as well as label powerset [7] methods. The best model trained got a 88.93% F measure on the test set. The work also points out to the issue of data sparseness during formulation of problem as multi-label classification. This issue will only compound

as the number of classes increase, which is the case with jump to ICD-10 from ICD-9. Hence, direct application of multi-label classification was ruled out.

Pariera et al. [8] propose a semi-automated ICD-10 coding system using mapping between MeSH [9] and ICD-10 extracted from UMLS metathesaurus [10], and usage of drug prescriptions by exploiting mapping between prescription drugs and relevant ICD-10 codes. The aim of the system is to assist the human coders by narrowing down on code sets to choose from using the contextual information. While this system achieves a recall as high as 68% under some settings, it is dependent on the quality of contextual information recorded.

Lita et al. [11] presented data mining based approaches using Support Vector Machines (SVM [12]) and Bayesian Ridge Regression [13] for ICD-9 classification, but these cannot be extended to ICD-10 because of the high number of classes. These methods do not take into consideration the semantic aspect of the words during text classification. There are various efforts which achieve a varying degree of success based on the specific use case. But these works were not a general effort to be applied to the entire set of ICD codes.

One of the notable works presented by Subotin et al. [11] considers clinical text and builds a 2-level hierarchical classifier to predict ICD-10 PCS code using regularized logistic regression. The work uses a bag of tokens available in their training data as the feature set for ICD-10 PCS code prediction. They build a code concept mapping model and concept code mapping model to rank the codes.

Since the organization of the codes in ICD-10 differs substantially (in terms of hierarchical structure, number of codes and granularity of the descriptions) when compared to ICD-9, it is only natural that the work done on later cannot be effectively extended to former.

Our work differs from the ones presented because it not only takes into account the semantic aspect of words in a text but also breaks down the classification to basic levels to achieve maximum accuracy and good decision boundaries among classes. Also, ICD codes are annotated using the description of diagnoses only and does not depend on supplementary information such as prescriptions etc. Our original contribution includes the formulation of task as a hierarchy, limiting the number of classes at each level by transforming the classification task to character level, the architecture design of the system where inference from each hierarchy cascades as a feature to the succeeding stages, use of the distributed word embeddings (Word2Vec [14]) to take into account the semantics of the word, and considering the words as a sequence so that loss of context due to change in word order is kept at a minimum. The final architecture can be loosely seen as a combination of multi-class classification and multi-label classifier chains [15].

3 Background

3.1 ICD-10 Coding System

ICD-10 is the latest in-use version of ICD. ICD-10 is split into two systems namely, ICD-10 Clinical Modification (ICD-10-CM) and ICD-10 Procedure Coding System (ICD-10-PCS). ICD-10-CM is the diagnostic coding used by health-care providers while ICD-10-PCS is used for inpatient procedure reporting by hospitals. The ICD-10 coding standard is a seven character coding convention and follows a curated hierarchical structure unlike the previous version of ICD, i.e. ICD-9.

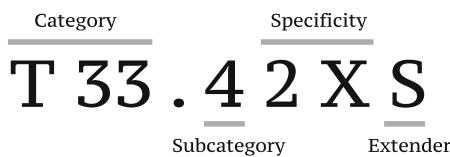


Fig. 1. ICD-10 code structure

Figure 1 gives a summary of the function of each of the seven characters in the ICD-10 code. A minimum of first 3 characters, together called category must be present to form a valid ICD-10 code. Among the three, the first character is an alphabet (except U) that points to the type of diagnosis eg. injury, poisoning, infection etc. The second two characters are numeric and together they point out the specific ailment or diagnosis of the given type. Occasionally, a dot separator is used after the 3rd character but it is not mandatory. The character at position four through seven may or may not follow the category. These are filled on the basis of the precision of information present in the description of diagnosis such as severity, etiology (the cause, the set of causes or the manner of causation of disease or condition), laterality etc. The last character is an extender or extension, which is used to specify the type of encounter, namely, initial encounter (patient is receiving active treatment), subsequent encounter (encounter after an active phase of treatment), or sequela (complication or condition as a direct result of an injury). “X” is used as a placeholder if a position is to be skipped for specifying the next characters.

Various revisions of the ICD-10 coding system are released periodically which appends, modifies and removes the previous errors and discrepancies in the codes, spellings etc. According to the statistics presented by [16], ICD-10 is cited in more than 20,000 scientific articles and is adopted across more than 100 countries. With the growing norm of adoption of ICD coding, there is very high demand for automatic coding systems, but there are not many such solutions available for use by hospitals and other consumers of ICD information.

3.2 Clinical Documents Versus Diagnosis Phrases

Clinical Documents can be defined as a digital or analog record detailing medical treatment, medical trial or clinical tests. These are generally a narrative of a physician's evaluation and follow up notes of a given patient that can be presented in either tabular or free text formats. The clinical documents may be in form of discharge summaries, surgical reports, progress notes etc.

It is important to note the difference between the clinical documents and the diagnosis phrases. A clinical document is a detailed report of the patient history of findings, medications and procedures. But diagnosis phrases are chunks or phrases extracted from these clinical documents that contain only the relevant information about the patient diagnosis. For example, a clinical document may have the following line in the patient diagnosis, "Patient is a 48-year old male and presented himself to the Emergency Department with chest pain. AFib and massive cardiac arrest was observed.", but only the phrases, "chest pain", "AFib" (short form of Atrial fibrillation) and "cardiac arrest" are the relevant diagnosis phrases.

The predictive models such as LSTM, Convolutional Neural Networks (CNN) [17], term-frequency inverse document frequency (TF-IDF) etc. mentioned in this work are trained on the diagnosis phrases and not on the entire unprocessed clinical document for the purpose of predicting the ICD codes, i.e. training here only refers to formulation of task that leads to successful ICD coding of diagnosis phrases. Apart from this we also explore and present the preprocessing pipeline to retrieve these phrases of interest from real-world discharge summaries. These supporting systems are briefly discussed in Sects. 3.3 and 3.4.

3.3 Q-Map

Q-Map [18] is a simple yet powerful system that can sift through large datasets to retrieve structured information aggressively and efficiently. It is backed by an effective mining algorithm based on curated knowledge sources, that is both fast and configurable.

It works in two phases, namely training (indexing) and testing. The pre-processing options are available through configuration that can switch on the modules as and when required. The heart of the system lies in the Aho-Corasick [19] algorithm that builds a finite-state machine (very similar to tree structure) with indexed failure state for each node during the training phase.

By using the configuration options in semantic types, one can filter out only the diagnosis terms. Because it is using UMLS as the knowledge source, filtered out terms and phrases are elementary as that is the intrinsic property of UMLS concepts. The elementary nature of concepts retrieved ensures that each output concept has one and only one ICD-10 code associated with it. Table 1 gives a sample output from Q-Map system.

Table 1. Q-Map output

Document	Output
Ligament tear observed in radiograph	Diagnosis: Ligament tear
Patient is a 53 year old female and has been recommended for unilateral total knee replacement	Procedures: Radiograph
Patient was given Heparin as a prophylactic treatment to prevent DVT	Total knee replacement
	Prophylactic treatment
	Medicines: Heparin

3.4 Negation Detection

Farkas et al. [5] lay strong emphasis on detection of negation among diagnoses as it plays an important role in determining the final performance of such coding systems. For this reason, all the diagnosis phrases retrieved from free clinical narrative texts via Q-Map during the real world application, are also passed to an algorithm called NegEx [20] (which is a rule based negation detection algorithm that can resolve complex sentence structures such as negations of negations etc.) to weed out the negated phrases.

4 Optimization Objectives

The study begins by defining the ICD classification task and the corresponding optimization objectives. Given an input sequence of arbitrary length, M , the goal is to produce an ICD code of appropriate length, N , which can vary between 3 to 7 based on the precision of the description present in the word sequence. Let the input sequence consist of M words x_1, x_2, \dots, x_M coming from a fixed vocabulary ν of size $|\nu| = V$. Each word is represented using an indicator vector x_i for $i \in \{1, 2, \dots, M\}$, where the indicator vector can be either one-hot encoded over a bag of words model, or a distributed representation of the word based on the Word2Vec [21] training. Furthermore, the notation $x_{[i\dots j]}$ is used to indicate the sub-sequence from index i to index j .

4.1 One-Step Classification

In a one-step classification method, there is a fixed set of possible output classes, $Y = \{y_1, y_2, \dots, y_l\}$, where $l = 93,830$ (number of classes). A one-step classifier takes x as input and outputs $y \in Y$ to maximize the probability, $p(y|x; \theta)$ where θ is the parameter of the classifier learnt through training.

The classifier tries to find the optimal θ such that,

$$\arg \max_{\theta} s(x, y) \quad (1)$$

under a scoring function $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$. If a window size W is used and conditional log probability is chosen as the scoring function, then $s(x, y)$ from (1) can be approximated using the W^{th} order markov [22] assumption as follows,

$$s(x, y) = p(y|x; \theta) \approx p(y|x_{[1\dots W]}; \theta) \quad (2)$$

4.2 Hierarchical Classification

In hierarchical classification, the single-step classification is split into several levels. The breakdown of classes into levels is explained in Table 2 and Fig. 2. Looking at the hierarchies, a natural question arises as to why is the second and third digit not broken down into individual chunks, and the reason lies in the fact that under the implementation of ICD-10 coding they are treated as a single unit instead of giving individual functions to each. Following this very intuition, levels have been broken down based on how the ICD-10 codes are designed, which is explained below in Sect. 5.1.

Table 2. Classification hierarchies

Hierarchy	Classes	Characters
Level 1 (L1)	25	First
Level 2 (L2)	106	Second and Third
Level 3 (L3)	23	Fourth
Level 4 (L4)	14	Fifth
Level 5 (L5)	19	Sixth
Level 6 (L6)	24	Seventh

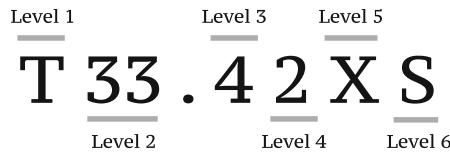


Fig. 2. Hierarchical breakdown of ICD-10 code

Following the new structure, the classifier at hierarchy i , takes as input x along with the outputs from preceding classifiers (if any) $y^{(i-1)}, \dots, y^{(1)}$ and outputs a $y^{(i)} \in Y^{(i)}$, where $Y^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \dots, y_{l_i}^{(i)}\}$ and l_i is the number of classes at hierarchy i . The notation $y^{(i)}$ is used to denote variable at level i . Similarly, notation $y^{(i\dots j)}$ is used to denote corresponding variables through levels i to j .

In hierarchical classification, a classifier at level i tries to find the optimal θ such that,

$$\arg \max_{\theta} s(x, y^{(1 \dots i)}) \quad (3)$$

For a window size, W and conditional log probability as the scoring function, $s(x, y^{(1 \dots i)})$ is defined as,

$$\begin{aligned} s(x, y^{(1 \dots i)}) &= p(y^{(i)} | y^{(1 \dots (i-1))}, x; \theta) \\ &\approx p(y^{(i)} | y^{(1 \dots (i-1))}, x_{[1 \dots W]}; \theta) \end{aligned} \quad (4)$$

5 Architecture and Procedure

5.1 Training Data

The training data is obtained from the official release of ICD-10 Data, 2017 version released by Centers for Disease Control and Prevention [23] and the annotated data (containing diagnosis phrase from clinical documents and manually annotated ICD codes) provided by some of the private hospitals from India.

The data received from the hospital is in compliance with the Information Technology Act, 2000 and Information Technology (Amendment) Act, 2008 under the Indian Penal Code. The data is received by the hospital with the signed patient consent and is anonymized to mask all the sensitive personal data and information (SPDI) at the source itself.

The official data released by CDC consists of ICD-10 codes along with their prescribed long and short descriptions. The data has 93,830 unique codes, where each code is accompanied by corresponding long and short descriptions. The long descriptions are the actual descriptions of the ICD codes, while the short description use the abbreviations which might not be universally accepted. For example, consider the short description “Intraop and postproc comp and disord of eye and adnexa, NEC” for the corresponding long description “Intraoperative and postprocedural complications and disorders of eye and adnexa, not elsewhere classified”. The short description presents the abbreviation “NEC” for “Not Elsewhere Classified”, but there are several other full forms in use such as “Necrotizing Enterocolitis”. Similarly, the short descriptions have a lot of acronyms and shortened word forms (such as encr. for encounter, sql. for sequela) which increase the number of Out of Vocabulary (OOV) words when using the pre-trained word embeddings (explained in Sect. 5.2).

The data contributed by the private hospitals come in form of phrases extracted from discharge summaries and clinical notes that consist of only the relevant information about a diagnosis along with ICD-10 codes manually annotated by doctors from various specialities. The codes are re-annotated by the consulting clinician and inter-rater reliability (IRR) is calculated using kappa statistic. The scores are as high as 0.94 for the first character which gradually falls to 0.72 as we reach the seventh character.

Since the institutions contributing data are multi-speciality hospitals with around 600 beds, the data has codes from all major classes of ICD-10. A total of

134,733 data rows are contributed which have a total of 69,777 distinct descriptions that point to 14,692 distinct ICD-10 codes.

The metrics regarding the distribution of codes and lengths of descriptions in the data given by the hospitals with respect to the ones present in ICD-10 data released by CDC is presented in Tables 3 and 4 respectively.

Table 3. Distribution of distinct ICD-10 codes

Granularity	CDC data	Hospital data
7-Character codes	93,830	14,692
3-Character codes	1910	1620
Level 1 (L1)	25	25
Level 2 (L2)	106	105
Level 3 (L3)	23	23
Level 4 (L4)	14	14
Level 5 (L5)	19	16
Level 6 (L6)	23	16

Table 4. Code description length metrics

Metrics	CDC data	Hospital data
Mean	10.05	5.06
Median	9	4
Mode	7	4
Max	32	30
Min	1	1

Together, the number of data records available is around 163k with a total vocabulary size of 13,634, i.e. $|\nu| = 13,634$. A standard stratification train-test split of 70:30 is performed on the data to ensure that none of the codes is missed out during the training phase. Due to the lack of data, stratification does not ensure the same set of classes in the train and test splits. The metrics about the class coverage are presented later in the paper while discussing the results.

5.2 Word Embeddings

For the purpose of training, input sequences are fed as distributed vector representations learned from Word2Vec instead of one-hot encoded vectors over a bag-of-words model.

Some of the advantages of this technique are the removal of data sparsity as Word2Vec is a dense model, dimensionality reduction as vector dimension in Word2Vec is several orders of dimension smaller than the bag-of-words model, and the inclusion of a sense of semantic distance which is an intrinsic property of Word2Vec representations.

The pre-trained open-source Word2Vec model released by Biomedical natural language processing lab (BioNLP) [24] was used for training. The vector representations were trained on text from PubMed [25], PMC [26] and English Wikipedia [27]. The pre-trained model from BioNLP has a vocabulary size of 5,443,656 with each word having 200-dimensional vector representation. This model is apt for our use case as the documents used for training the model include medical documents as well as general English documents, which makes the vector space representations inclusive of the medical context of words like seizure, procedure etc.

During our training, we encountered a total of 760 OOV cases. Most of these were irregular words such as lemli, xyy etc. which might be the result of spelling mistakes or usage of acronyms in the data.

Removal of data possessing OOV cases from training data did not have a positive effect on the evaluation metrics, pointing to the fact that the advantages of contextual information attributed by the word embeddings far outweighed the disadvantages of OOV words.

5.3 Preprocessing

Basic text preprocessing [28] on the training data is performed in the form of handling punctuations, extra spaces in the text. No other preprocessing such as stop-word removal, stemming or lemmatization is performed. The LSTM algorithm is inherently capable of understanding the context and importance of a word in a sentence and the amount of significance it plays in the classification process.

5.4 Experiments

The work begins by employing various different techniques such as TF-IDF, Inverted Indexing, SVM, CNN and LSTM for the single-step classification task and observing their performance which is summarized in Table 5. The top performers in this task are re-applied on the hierarchical classification to achieve a performance boost. The accuracy on the held-out data (test dataset mentioned in Sect. 5.1) is taken as the final metric of performance for evaluation of models.

For the TF-IDF method, standard smoothed IDF function was used. The accuracies observed was as low as 29.3%, a strong reflection of the fact that word importance and document lengths used for representing the concepts is very random in the medical domain, e.g. “pyrexia” and “fever” are semantically very similar but such resemblance is not captured in TF-IDF method. Similarly,

Table 5. One-step classification performance

Model used	Accuracy
TF-IDF	0.2932
Inverted index	0.3727
SVM	0.1407
CNN	0.5454
LSTM	0.5798

“diabetic foot” and “diabetes mellitus with foot ulcer” represent the same diagnosis, but the difference in sentence length attributes low TF-IDF scores to such examples.

To overcome the drawback of semantic loss and sequence lengths explained above in TF-IDF method, an attempt was made to bag synonyms of a medical concept under a single bucket using Q-Map (an optimized version of Metamap [29]) which is explained in Sect. 3.3. After concept retrieval, an inverted index is built, wherein each bucket points to the set of ICD-10 codes it occurs in. The accuracy achieved using this method was 37.27%, which is a considerable jump from the previous method. A drawback observed with this method was the excessive dependency on the word order and the organisation of sentences in the concept retrieval step. For example, while “cancer of brain”, “brain cancer” etc. is a part of UMLS database and is readily identified in text, a string like “cancerous tissue mass in temporal lobe” is not identified as a single concept because even though UMLS is a very comprehensive aggregation of medical concepts it is not exhaustive.

Following the above two approaches the popular machine learning technique, SVM model was also evaluated for the single step classification process using bag-of-words model. This model gave an accuracy of 14.07%. It is posited that the major reason for such low accuracy can be attributed to the high cardinality of the feature set i.e. the vocabulary size of 13634. The similar low performance was observed by applying the SVM on features extracted using Q-Map.

Observing the drawbacks of the previous methods, the work pivoted to deep learning techniques, namely, CNN and LSTM because of the advantages they offer in overcoming the issues of word order, sentence structure, word importance, semantics of words etc. For setting up these models for single-step classification, the input layer is weighed using the distributed word embedding from Word2Vec (explained in Sect. 5.2) and the model parameters are learnt, providing a softmax score which indicates the likelihood of belonging to a class. Both CNN and LSTM outperformed the previous techniques. CNN converged to an accuracy of 54.54%, while the accuracy of LSTM was observed to be 57.98%. The authors posit that the LSTM performs better than CNN because it takes into consideration the word order in a document which plays a significant role e.g. word order and semantics are crucial in differentiating descriptions like “Hypertensive chronic

kidney disease” and “Chronic kidney disease”, where the former is coded as “I12” while later as “N18”.

Among the methods evaluated on single-step classification, LSTM and CNN give the most promising results. Hence, the hierarchical classification is implemented using this method and the experimental results maintained the order or performance that was observed in one-step classification, i.e. LSTM outperformed CNN. Following this, the hierarchical LSTM was further optimized by hyper-parameter tuning and changing network architecture as explained below in Sect. 5.5. The results are explained in detail in the sections that follow.

There are few adjustments made in the coding scheme for implementing the hierarchical classification. One major hindrance in the process is the uncertainty in the length of expected ICD code (ranges between 3 and 7). This is tackled by padding the codes with X’s and getting a uniform length of codes i.e. 7. This padding is harmless because X’s are originally utilized as a placeholder as explained in Sect. 5.1.

In the hierarchical classification, there are 6 classifiers set up, one for each layer listed in Table 2. The inference from each classifier is fed to the next level as represented mathematically in (4). So, the system consists of stacked LSTM classifiers. The architecture can be better understood from Fig. 3.

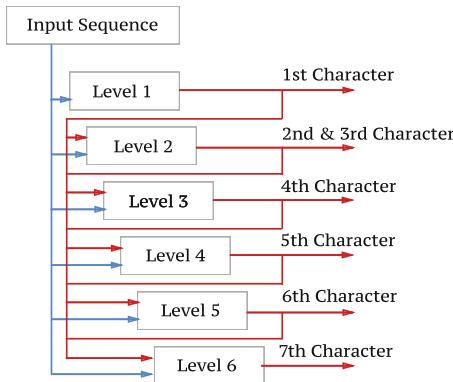


Fig. 3. Cascaded hierarchically stacked architecture

The hierarchical LSTM model outperformed the corresponding model under single-step classification. Evaluation metrics of the model are listed in Table 6. It includes the accuracy, macro/micro precision, recall and F1-scores along with class coverage (number of labels) for both training and test datasets for individual classifiers.

Table 7 lists the similar evaluation metrics for the cumulative coding at a given level. Cumulative coding means the aggregation of predicted codes at any level given the codes predicted at previous levels. Since the minimum number of characters in a valid ICD-10 codes is three, these metrics are presented from level 2 onwards only.

Table 6. Evaluation metrics results of individual classifiers in the hierarchical setup

Hierarchy	Accuracy	Micro			Macro			Classes
		Precision	Recall	F1	Precision	Recall	F1	
L1 Test	0.9586	0.9586	0.9586	0.9586	0.9334	0.9268	0.9301	25
L1 Train	0.9823	0.9823	0.9823	0.9823	0.9724	0.9691	0.9707	25
L2 Test	0.8656	0.8656	0.8656	0.8656	0.8701	0.8645	0.8673	106
L2 Train	0.9205	0.9205	0.9205	0.9205	0.9300	0.9255	0.9277	106
L3 Test	0.8067	0.8067	0.8067	0.8067	0.8199	0.7624	0.7901	22
L3 Train	0.8652	0.8652	0.8652	0.8652	0.9301	0.8418	0.8838	23
L4 Test	0.8935	0.8935	0.8935	0.8935	0.9067	0.8754	0.8908	13
L4 Train	0.9410	0.9410	0.9410	0.9410	0.9498	0.9488	0.9493	14
L5 Test	0.9627	0.9627	0.9627	0.9627	0.9123	0.9018	0.9070	19
L5 Train	0.9851	0.9851	0.9851	0.9851	0.9529	0.9299	0.9413	19
L6 Test	0.9928	0.9928	0.9928	0.9928	0.9923	0.9825	0.9874	24
L6 Train	0.9976	0.9976	0.9976	0.9976	0.9984	0.9943	0.9963	24

Table 7. Cumulative evaluation metrics results of the hierarchical setup

Hierarchy	Accuracy	Micro			Macro			Classes
		Precision	Recall	F1	Precision	Recall	F1	
L2 Test	0.8995	0.8995	0.8995	0.8995	0.8610	0.8145	0.8371	1634
L2 Train	0.8532	0.8532	0.8532	0.8532	0.8636	0.8421	0.8527	1879
L3 Test	0.7833	0.7833	0.7833	0.7833	0.7319	0.7366	0.7342	5888
L3 Train	0.7885	0.7885	0.7885	0.7885	0.7290	0.7369	0.7329	8935
L4 Test	0.7741	0.7741	0.7741	0.7741	0.7391	0.7502	0.7446	11560
L4 Train	0.7919	0.7919	0.7919	0.7919	0.7890	0.7919	0.7905	21912
L5 Test	0.7743	0.7743	0.7743	0.7743	0.8014	0.8133	0.8073	17683
L5 Train	0.7959	0.7959	0.7959	0.7959	0.8796	0.8813	0.8805	36418
L6 Test	0.7205	0.7205	0.7205	0.7205	0.8776	0.9939	0.8857	23429
L6 Train	0.7149	0.7149	0.7149	0.7149	0.8079	0.8092	0.8086	71361

5.5 LSTM Architecture and Training

The experiments started out by using vanilla LSTM but later few design modifications were introduced which further improved the performance of the network.

As a first step, the entire word sequence along with the inputs from previous layers was handled as one long sequence fed to the LSTM layer.

As a next step, a different architecture wherein word sequence and the inferences from previous layers are handled as separate entities, and only the word sequence is fed into the LSTM layers while the other inputs are fed into a dense

layer and then concatenated before normalizing and adding dropout [30]. It is observed that dropout and batch-normalization [31] help in generalizing the classifier by a considerable margin as the ratio of training accuracy to the test accuracy is much closer to unity in models after normalization and dropout. Figure 4 gives a brief generalized summary of the individual level LSTM classifier architecture. Exact architectural details along with model parameters for individual level classifiers can be seen in Figs. 5, 6, 7, 8, 9 and 10 .

Also, the input word sequence lengths were varied keeping the rest of the parameters static in order to study the effect of window size, W on the accuracy. Table 8 lists the final window size parameters that were found most optimal by observing the increase in accuracy for an increase in window size. The results are in line with the fact that the later characters in ICD-10 coding are to do with the precision of the diagnosis which is often captured in longer sentences.

The text preprocessing (mentioned in Sect. 5.3) for LSTM is kept at a very minimal intrusion because it is capable of drawing more informed context from a well-formed sentence.

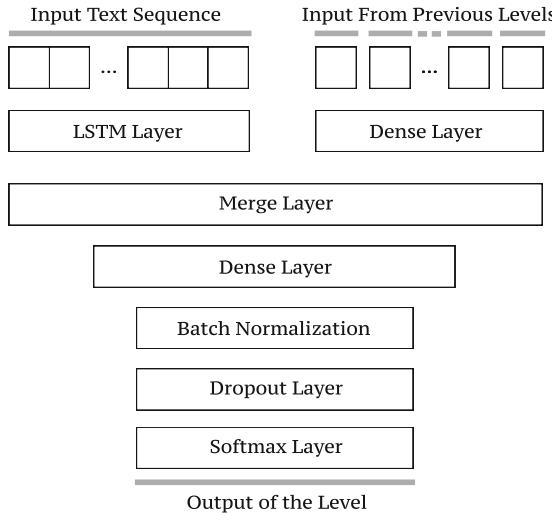


Fig. 4. Hybrid network structure at each level using LSTM



Fig. 5. Level 1 classifier

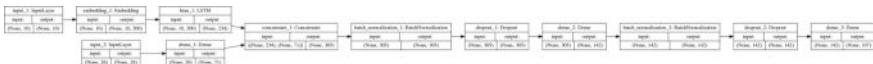
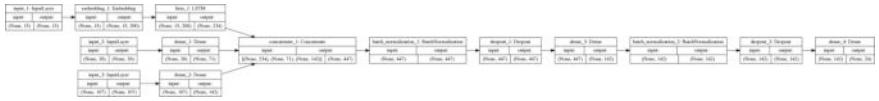
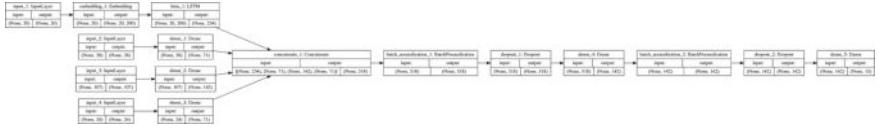
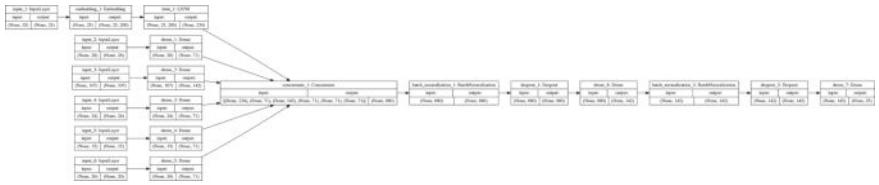


Fig. 6. Level 2 classifier

**Fig. 7.** Level 3 classifier**Fig. 8.** Level 4 classifier**Fig. 9.** Level 5 classifier**Fig. 10.** Level 6 classifier**Table 8.** Window size for classifier levels

Hierarchy	Window size
Level 1 (L1)	10
Level 2 (L2)	10
Level 3 (L3)	15
Level 4 (L4)	20
Level 5 (L5)	20
Level 6 (L6)	25

Once the model is defined, one can estimate the model parameters to minimize multi-log loss, the cost function which is given by,

$$\begin{aligned}
 L_{log}(Y, P) &= -\log Pr(Y|P) \\
 &= -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} y_{i,k} \log p_{i,k}
 \end{aligned} \tag{5}$$

where the true labels are a set of samples encoded as a 1-of-K binary indicator matrix, \mathbf{Y} , i.e. $y_{i,k} = 1$ if sample i has label k taken from a set of K labels, and \mathbf{P} is a matrix of probability estimates, with $p_{i,k} = Pr(t_{i,k} = 1)$. The networks are trained as long as there is an improvement in test accuracy, using a learning rate that decreases by order of 10 as the training epochs proceed.

6 Results and Discussion

Our results from LSTM are presented in Tables 6 and 7. Table 9 lists out a few examples of the automatic coding by the system. It was first noted that the conventional NLP methods such as TF-IDF and Inverted Indexing show relatively poor performance, indicating that neither term frequency and its importance in corpus or bagging techniques alone are sufficiently discriminative for decision making.

Table 9. Model output

Input	Code predicted	ICD-10 description
Patient has fever	R509XXX	Fever, unspecified
Brain cancer	C719XXX	Malignant neoplasm of brain, unspecified
Right eye has retinal vein occlusion and is stable	H348112	Central retinal vein occlusion, right eye, stable
Diabetic foot	E11621X	Type 2 diabetes mellitus with foot ulcer

Both deep learning techniques, i.e. LSTM and CNN are better suited to draw the inferences from the training data available, but the advantages of LSTM over CNN in the natural language domain are evident from performance metrics on single step classification. It is also observed that breaking down the task into hierarchical classification is advantageous and better convergence for the networks are achieved when compared to single step classification among the massive number of classes that the data presents.

In particular, LSTM model that cascades the inference from one level to another helps achieve much better classification metrics. While it gives an accuracy of 72.05% on the seven-digit coding scheme, it is worth noting that the accuracy for predicting the equally important three-digit code (cumulative accuracy at level 2) is considerably high at 89.95%. The lower values in the macro

metrics in Table 7 can be attributed to the high-class imbalance which can be observed in the class coverage of the same table. The high accuracy of 99.28% for level 6 in Table 6 might be due to the fact that the extenders of ICD code which are represented by the 7th character in the code are generally associated with a combination of words that do not present a lot of variations over the texts irrespective of the source as explained in Sect. 5.1.

Table 10. Invalid codes predicted

Hierarchy	Data instances	Invalid codes ratio
L2 Test	32722	0.0019
L2 Train	130885	0.0012
L3 Test	32722	0.0172
L3 Train	130885	0.0156
L4 Test	32722	0.0202
L4 Train	130885	0.0299
L5 Test	32722	0.0280
L5 Train	130885	0.0322
L6 Test	32722	0.0287
L6 Train	130885	0.0379

It can be noted that while the actual number of classes in ICD-10 is limited to 93,830, the sample space for classes in the hierarchical model is very high ($25 * 106 * 23 * 14 * 19 * 24 = 389,104,800$), if all the possible combinations are considered for classifiers at each level. The number of invalid codes predicted is studied at each level presented in Table 7 and the findings are summarized in Table 10.

It can be seen that the total number of invalid codes predicted are very low with the maximum of 3.79% at seven-digit code prediction level. The same is as low as 0.19% for the three-digit coding. The low percentage of invalid code predictions suggests that the levels in LSTM network in the hierarchical set up are statistically strengthened to prevent prediction of invalid subclasses for a given superclass by means of passing information from one classifier level to the following levels. Invalid codes are truncated at the preceding level based on a lookup table to avoid such predictions in a real-world scenario and production deployments.

It must be noted that the system is only trained for predicting correct code given that the input data is a valid diagnosis string. The instance of an input string from any other domain, invalid diagnosis strings or concatenation of strings belonging to varied classes of ICD-10 code would lead to haphazard results. This is because inherently, the neural network is trained to output a single class for a given input sequence. But when sequences containing concepts

belonging to more than one class are passed in a concatenated form, the network will not work with maximum efficiency.

We prevent the occurrence of such cases by devising a set of steps which include the best practices to achieve the most efficient coding when applied to clinical free text. A system for splitting diagnosis documents into suitable valid diagnosis phrases that can be annotated by individual codes plays an important role in maintaining the performance of the system in predicting correct codes when applied to a real-world clinical or discharge note. These obstacles are tackled using the proprietary clinical concept extraction system, Q-Map (explained in Sect. 3.3), and detection of negation of retrieved concepts using the NegEx algorithm (explained in Sect. 3.4).

7 Conclusion

The authors designed an efficient LSTM based hierarchical network for automatic classification of ICD-10 coding using a limited amount of dataset by exploiting cascaded hierarchical classification. As a next step, this architecture will be extended to various other hierarchical coding schemes, such as CPT and LOINC and other architecture options such as LSTM with attention mechanism, LSTM-CNN parallel model will be explored. Also, there is a scope to improve the accuracy by deriving attention based summaries from the discharge notes. Both the future works pose additional challenges in terms of availability of training data.

The utility of this model is already proving to be very useful. Following the development of this model, systems have been deployed on-premise for various hospitals that helps in suggestive ICD annotation of clinical documents. It is also helpful in creating registries of patients for doctor references. Similarly, using the data from MIMIC III [32], systems are developed for next disease prediction and the onset of disease prediction using methods such as collaborative filtering [33]. Apart from this, it is helping in other scenarios such as insurance claims review and document search based on diseases and diagnostic related groups (DRGs).

References

1. World Health Organization: International statistical classification of diseases and related health problems, vol. 1. World Health Organization (2004)
2. Beebe, M., et al.: Current Procedural Terminology: CPT. American Medical Association (2007)
3. McDonald, C., et al.: Logical observation identifiers names and codes (LOINC) users' guide. Regenstrief Institute, Indianapolis (2004)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
5. Farkas, R., Szarvas, G.: Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinform.* **9**3 BioMed Central (2008)
6. Boutell, M.R., et al.: Learning multi-label scene classification. *Pattern Recogn.* **37**(9), 1757–1771 (2004)

7. Tsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.* **23**(7), 1079–1089 (2011)
8. Pereira, S., et al.: Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding. *MIE* (2006)
9. Lipscomb, C.E.: Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.* **88**(3), 265 (2000)
10. Schuyler, P.L., et al.: The UMLS Metathesaurus: representing different views of biomedical concepts. *Bull. Med. Libr. Assoc.* **81**(2), 217 (1993)
11. Lita, L.V., et al.: Large scale diagnostic code classification for medical patient records. In: Proceedings of the Third International Joint Conference on Natural Language Processing, vol. II (2008)
12. Hearst, M.A., et al.: Support vector machines. *IEEE Intell. Syst. Appl.* **13**(4), 18–28 (1998)
13. Hoerl, Arthur E., Kennard, Robert W.: Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970)
14. Mikolov, T., et al.: Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
15. Read, J., et al.: Classifier chains for multi-label classification. *Mach. Learn.* **85**(3), 333 (2011)
16. International Classification Of Diseases, 10th Revision (ICD-10). World Health Organization. N.p., 2018. Web. 9 July 2018
17. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint [arXiv:1408.5882](https://arxiv.org/abs/1408.5882) (2014)
18. Azam, S.S., et al.: Q-Map: clinical concept mining with phrase sense disambiguation. arXiv preprint [arXiv:1804.11149](https://arxiv.org/abs/1804.11149) (2018)
19. Aho, A.V., Corasick, M.J.: Efficient string matching: an aid to bibliographic search. *Commun. ACM* **18**(6), 333–340 (1975)
20. Chapman, W.W., et al.: A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.* **34**(5), 301–310 (2001)
21. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)
22. Gagniuc, P.A.: Markov Chains: From Theory to Implementation and Experimentation. Wiley, New York (2017)
23. National Center for Health Statistics. Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 11 June 2018 www.cdc.gov/nchs/icd/icd10cm.htm
24. Biomedical Natural Language Processing. Bio.nlplab.org, bio.nlplab.org/
25. Home - PubMed - NCBI, , U.S. National Library of Medicine www.ncbi.nlm.nih.gov/pubmed
26. Home - PMC - NCBI, , U.S. National Library of Medicine www.ncbi.nlm.nih.gov/pmc/
27. Main Page. Wikipedia, Wikimedia Foundation, 8 July 2018 www.en.wikipedia.org/wiki/Main_Page
28. shams-sam. Shams-Sam/Logic-Lab. GitHub, www.github.com/shams-sam/logic-lab/blob/master/TextPreprocessing/text.preprocessing.py. Accessed 9 July 2018
29. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium. American Medical Informatics Association (2001)
30. Srivastava, N., et al.: Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**(1), 1929–1958 (2014)

31. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint [arXiv:1502.03167](https://arxiv.org/abs/1502.03167) (2015)
32. Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**, 160035 (2016)
33. Schafer, J.B., et al.: Collaborative filtering recommender systems, pp. 291–324. The adaptive web. Springer, Berlin (2007)



Automated Gland Segmentation Leading to Cancer Detection for Colorectal Biopsy Images

Syed Fawad Hussain Naqvi, Salahuddin Ayubi, Ammara Nasim^(✉),
and Zeeshan Zafar

Department of Electrical Engineering, Bahria University, Islamabad, Pakistan
fawadnaqvi46@gmail.com, {salahuddinayubil,
zeeshanzafar19@yahoo.com, ammara.nasim@bui.edu.pk

Abstract. Glandular formation and morphology along with the architectural appearance of glands exhibit significant importance in the detection and prognosis of inflammatory bowel disease and colorectal cancer. The extracted glandular information from segmentation of histopathology images facilitate the pathologists to grade the aggressiveness of tumor. Manual segmentation and classification of glands is often time consuming due to large datasets from a single patient. We are presenting an algorithm that can automate the segmentation as well as classification of H and E (hematoxylin and eosin) stained colorectal cancer histopathology images. In comparison to research being conducted on cancers like prostate and breast, the literature for colorectal cancer segmentation is scarce. Inter as well as intra-gland variability and cellular heterogeneity has made this a strenuous problem. The proposed approach includes intensity-based information, morphological operations along with the Deep Convolutional Neural network (CNN) to evaluate the malignancy of tumor. This method is presented to outpace the traditional algorithms. We used transfer learning technique to train AlexNet for classification. The dataset is taken from MCCAIC GlaS challenge which contains total 165 images in which 80 are benign and 85 are malignant. Our algorithm is successful in classification of malignancy with an accuracy of 90.40, Sensitivity 89% and Specificity of 91%.

Keywords: Colorectal cancer · Malignant · Benign · Glands · Segmentation · Convolutional neural networks (CNN)

1 Introduction

Colorectal cancer grows in colon or rectum which are part of large intestine. It is deliberated to be one of the most fatal disease worldwide. In the beginning of most cases, it starts with polyps and with the passage of time it spreads to other parts of body which makes it cancerous. Automated analysis has got much importance since the last few years, as it is very helpful in cancer grading. The sample of tissues are taken and then they are examined microscopically. Slides are prepared from it and H&E (Hematoxylin and Eosin) stained. These slides are digitized for the use in CAD (Computer Aided Diagnosis) using whole slide scanners. The variation between the

images obtained is very large and it even increases on going toward highly malignant ones. These differences are based on contour, intensity or special dimensions. Hence, we need a solution which will be suitable for all the types of images.

Some of the techniques used earlier include, segmentation by globally optimized clustering in [1]. Size of graph of the image in [2] is reduced along with the complexity of affinity matrix. This simplicity aided in normalized cuts algorithm in [3, 4]. Learning glands done by weekly supervised method [5], the representation based on sparse dictionary [6], Model fitting and pixel level clustering are the mostly used approaches.

The segmentation and classification in [7] is done by labeling glands using features and PPMM (Probabilistic Pairwise Markov Model). Textural feature alone [8] or the combination of cytological and the textural features [9] are used for the detection of cancer regions by the classification of image pixels and patches. Wu et al. [10] Presented a method based on region growing in which seed is initialized in the large empty space and is expanded till it reaches the boundary of epithelial nuclei. This is helpful for benign glands but failed to produce fruitful results in case of malignant. For the clustering of nucleus, stroma and lumen [11] used textural features. After that they removed the region containing stroma and lumen, while separated the region of nucleus.

Deep Learning (DL) has been widely used for detection of objects for the past few years. DL is composed of multiple convolutional layers and improves iteratively over learned representations of underlying data for the attainment of class detachability. Tissue classification [12], mitosis detection [13–16] immunohistochemical staining and many more have proved the potential of DL for being unifying approach in variety of tasks in DP (Digital Pathology). Adequate data is used for training to obviate the need of manual classification by generalizing the system to unperceived situations.

Neural architectures were first proposed in [17], for the recognition of handwritten characters achieving the accuracy of 99.2% on MNIST dataset [18]. This was considered as the origin of modern CNN (Convolutional Neural Networks). Researchers got their inspiration from the mammal's way of visually perceiving the world, by utilizing layered architecture of neurons in encephalon.

We altered some of the layers of pre-trained CNN for image classification which can get better efficiency in case of small dataset.

2 Methodology

The histology images obtained from dataset is passed through the three main steps: The images (in RGB) are preprocessed to reduce unwanted information in the images as far as possible. Next step is application of intensity-based segmentation followed by morphological operations. Convolutional Neural Networks (CNN) are utilized for categorization of images into benign and malignant. CNN is trained on annotated images available in the data sets, Using, the trained model segmented images are tested for malignancy of tissues. Figure 1 shows the block diagram of proposed algorithm.

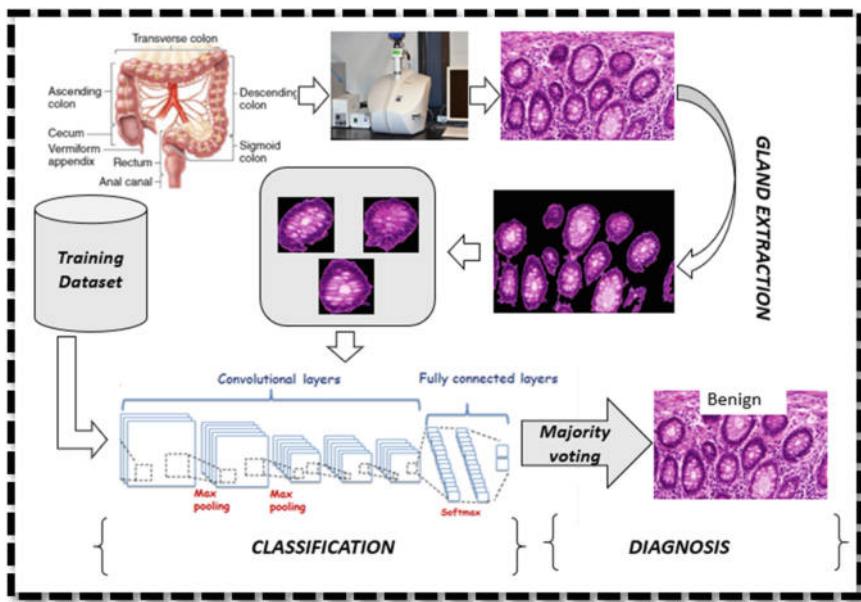


Fig. 1. Proposed approach

2.1 Image Acquisition

Images used for this research are taken from the dataset of **MCCAI GLaS 2015** challenge on gland segmentation of histology images [6]. The images are obtained from digitization of hematoxylin and eosin stained slides of Whole Slide Scanner with pixel resolution of $522 \times 775 \times 3$.

2.2 Preprocessing

Initially we separated Red, Green and Blue channels of the original RGB image. For the purpose of contrast enhancement, each RGB channel is passed through the process of histogram equalization (Fig. 2).

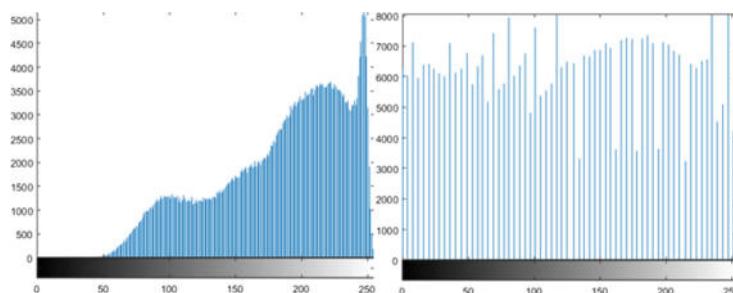


Fig. 2. Histogram equalization

The discrete approximation of transformation function for histogram equalization is given below.

$$s_k = T(r_k) = (L - 1) \sum_{j=0}^k p_r(r_j) \quad (1)$$

$$s_k = \frac{(L - 1)}{MN} \sum_{j=0}^k n_j \quad k = 0, 1, 2, \dots, L - 1 \quad (2)$$

where

N Total number of pixels

n_j Pixels with intensity value r_j

L Possible intensity levels.

2.3 Image Segmentation

We applied Otsu algorithm [19] to individual Red, Green and Blue channels (Fig. 3a–c) for image binarization furthermore intersection of all the three binarized images is taken to scrutinize the information (Fig. 3d).

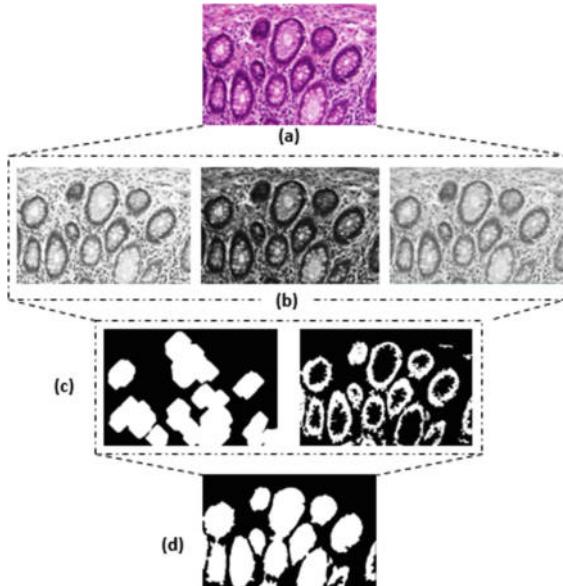


Fig. 3. Input image is labeled as (a) and the Red, Green and Blue images as (b). The resultant of Otsu [21] and morphological processing is labeled as (c), while final segmentation is labelled as (d)

$$I_{original} = (I_{r(gray)} + I_{g(gray)} + I_{b(gray)}) \quad (3)$$

$$I_{binary} = (I_{r(binary)} \cap I_{g(binary)} \cap I_{b(binary)}) \quad (4)$$

The images obtained after binarization, contain many small nuclei (I_{nuclei}) which are undesirable in our case, these components must be removed as much as possible to get the glands detached.

After the intersection of all the three Red, Green and blue components the resultant image contains lot of noise. We used morphological techniques to get the glands separated from other unwanted components. These techniques utilize the spatial and contour-predicated characteristics of objects present in the binary image. At first, we determined the objects connectivity and also computed the area occupied by each of these components. The extracted values labeled the objects based on their spatial dimensions. The pixels which corresponds to certain glands have proportionately greater connectivity than those of the nuclei, are stored discretely in variable ($I_{gland+nuclei}$). Then we extracted the nuclei components (I_{nuclei}) leaving the glandular objects. As depicted in Eq. 5, we have taken Exclusive Disjunction logical operation between both the ($I_{gland+nuclei}$) and (I_{nuclei}) to abstract nuclei left unremoved in ($I_{gland+nuclei}$) Fig. 3c.

$$I_{glands} = I_{glands+nuclei} \oplus I_{nuclei} \quad (5)$$

This process seems to be efficient for the benign tissues with the darker glandular boundaries but some of the benign tissues contain very dizzy and light glands, due to which their contrast and preprocessing was not fruitful enough to segment the glands. So, we familiarized another technique on the basis of morphological properties of those glands. Most of their glands include virtually white lumen. We have utilized region growing algorithm starting from the brightest pixels towards the boundary quite similar to method used in [10], the only difference arises in generation of seed. This step is usually not required for malignant glands where the boundaries of objects are more gathered towards lower histogram bins. The algorithm is accurate for distinguishing glandular region from the image by 97%.

2.4 Classification

Computer Vision and DP has been revolutionized by the involvement of DL. Building a new model from scratch and training it is far lengthy processes then using an existing network. It can give even higher value of accuracy if the model which is taken is related to the model of interest, but after training of CNN on dataset, the size turns out to be much bigger in this case as it is pre-trained on very large dataset. This method is known as Transfer Learning (TL).

The AlexNet model (available in MATLAB) which is used in our methodology is consisted of total 25 different layers of which 5 are convolutional layers, 3 are fully connected layers. Pooling and activation layers are in-between these layers.

This pre-trained model is modified slightly to make it suitable for classification of histopathology images of tissues. As it is previously trained to classify 1000 different types of images, so we reformed one of the fully connected layers of AlexNet, which is consisted of 1000 neurons for those classes and altered it to the required number of neurons. As we have solely two different types of classes (i.e. Benign and Malignant) so we substituted it with the network consisted of 2 different classes. We have also replaced the output layer which has learned the classification of AlexNet for 1000 different classes with an empty layer that can learn the classification of our dataset. Table 1 illustrates the updated layer information. 20% randomly selected images from both classes were used for training purposes. AlexNet is trained to ground truth images for which we need a bigger dataset. To increase the dataset, we extracted all glands and nuclei elements identified in native images of both the classes and stored them separately. 20% images from both classes were used for training purposes. Those separated objects are then varied by alteration of angle, contrast and illumination of pixels to increase the training data. The data includes more than thousand images for each class. The input images size is restricted by AlexNet so the data is converted to $227 \times 227 \times 3$, on which it is to be trained.

Table 1. AlexNet layers description [21]

S. No.	Name	Features
1	Data	$227 \times 227 \times 3$ images with normalization
2	Conv1	96 $11 \times 11 \times 3$ convolutions with stride [4 4] and padding [0 0]
3	Relu1	ReLU
4	Norm1	Cross channel normalization with 5 channels per element
5	Pool1	3×3 max pooling with stride [2 2] and padding [0 0]
6	Conv2	256 $5 \times 5 \times 48$ convolutions with stride [1 1] and padding [2 2]
7	Relu2	ReLU
8	Normi2	Cross channel normalization with 5 channels per element
9	Pool2	3×3 with stride [2 2] and padding [0 0]
10	Conv3	384 $3 \times 3 \times 256$ convolutions with stride [1 1] and padding [1 1]
11	Relu3	ReLU
12	Conv4	384 $3 \times 3 \times 192$ convolutions with stride [1 1] and padding [1 1]
13	Relu4	ReLU
14	Conv5	256 $3 \times 3 \times 192$ convolutions with stride [1 1] and padding [1 1]
15	Relu5	ReLU
16	Pool5	3×3 max pooling with stride [2 2] and padding [0 0]
17	Fo6	4096 fully connected layer
18	Relu6	ReLU
19	Drop6	50% dropout
20	Fc7	4096 fully connected layer
21	Relu7	ReLU
22	Drop7	50% dropout
23	Fc8	2 fully connected layers
24	Prob	Softmax
25	output	Benign or Malignant

3 Result and Discussion

The testing dataset we used, contain native 165 images out of which 80 are benign and 85 malignant. Each image contains 5 to 20 glands. The data required for training of CNN is further generated by separating glands using the available annotations provided by MCCAIS GlaS [6] and through intensity, orientation and morphological variations of the native glands. Almost 10,000 glands are fed to CNN for training purposes.

We have used the segmented gland images resized to $227 \times 227 \times 3$ for further classification. Each of the gland is assigned a class by CNN. An image is referred as benign or malignant on the basis of majority voting criterion.

Figure 4 demonstrate segmentation results of benign (a) as well as malignant glands (b). These results illustrate a computationally efficient algorithm in comparison to conventional algorithms utilizing cumbersome techniques e.g. SLIC and DBSCAN [20]. Proposed system segments both benign and malignant structures quite accurately. The classification of malignancy of image is accurate by 90.4% with sensitivity of 89% and specificity of 91%.

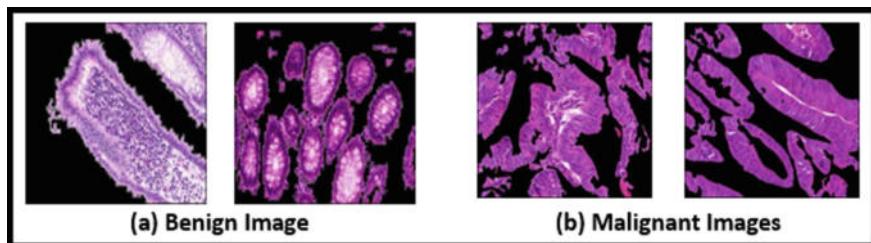


Fig. 4. Classification results **a** Image Classified as benign **b** Image Classified as malignant

4 Conclusion

A fully automated malignancy detection system which successfully classify tumor tissues into malignant and benign at an accuracy of 90.4%, Sensitivity of 89% and Specificity of 91% is proposed. The algorithm is based on intensity-based segmentation and CNN which is trained on more than 10,000 ground truth images. This system can further be extended towards TNM (Tumor, Nodes and Metastasis) grading of colorectal cancer.

Acknowledgements. We are thankful to MCCAIS GlaS 2015 contest for providing us with the relevant data [6].

References

- Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**(8), 888–905 (2000)
- Tao, W., Jin, H., Zhang, Y.: Color image segmentation based on mean shift and normalized cuts. *IEEE Trans. Syst. Man. Cybern. B. Cybern.* **37**(5), 1382–1389 (2007)

3. Xu, J., Madabhushi, A., Janowczyk, A., Chandran, S.: A weighted mean shift, normalized cuts initialized color gradient based geodesic active contour model: applications to histopathology image segmentation. In: Proceedings of SPIE, vol. 7623, April 2016, pp. 76230Y–76230Y–12 (2010)
4. Xu, Y., Zhang, J., Eric, I., Chang, C., Lai, M., Tu, Z.: Context-constrained multiple instance learning for histopathology image segmentation. In: International Conference on Medical Image Computing and Computer Intervention, MICCAI, vol. 15, no. Pt 3, pp. 623–30 (Jan 2012)
5. Gao, Y., Liu, W., Arjun, S., Zhu, L., Ratner, V., Kurc, T., Saltz, J., Tannenbaum, A.: Multi-scale learning based segmentation of glands in digital colorectal pathology images. In: Proceedings of SPIE, vol. 9791. pp. 97910M–97910M–6 (2016)
6. Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.-A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Ben Cheikh, B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D.R.J., Rajpoot, N.M.: Gland segmentation in colon histology images: the GlaS challenge contest, pp. 1–24 (2016)
7. Monaco, J., Tomaszewski, J., Feldman, M., Hagemann, I., Moradi, M., et al.: High-throughput detection of prostate cancer in histological sections using probabilistic pairwise Markov models. *Med. Image Anal.* **14**, 617–629 (2010)
8. Doyle, S., Feldman, M., Tomaszewski, J., Madabhushi, A.: A boosted Bayesian multi-resolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE Trans. Biomed. Eng.* **59**, 1205–1218 (2012) (F) Article title. *Journal* **2**(5), 99–110 (2016)
9. Nguyen, K., Jain, A., Sabata, B.: Prostate cancer detection: fusion of cytological and textural features. *J. Pathol. Inform.* **2**, 2–3 (2011)
10. Wu, H.-S., Xu, R., Harpaz, N., Burstein, D., Gil, J.: Segmentation of intestinal gland images with iterative region growing. *J. Microsc.* **220**(3), 190–204 (2005)
11. Farjam, R., et al.: An image analysis approach for automatic malignancy determination of prostate pathological images. *Cytom. Part B: Clin. Cytom.* **72**(4), 227–240 (2007)
12. Cruz-Roa, A., Basavanhally, A., González, F., Gilmore, H., Feldman, M., Ganesan, S., et al.: Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. *SPIE Med. Imaging* **9041**, 904103–904103-15 (2014)
13. Veta, M., van Diest, P.J., Willemse, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., et al.: Assessment of algorithms for mitosis detection in breast cancer histopathology images. *Med. Image Anal.* **20**, 237–248 (2015)
14. Roux, L., Racoceanu, D., Loménie, N., Kulikova, M., Irshad, H., Klossa, J., et al.: Mitosis detection in breast cancer histological images An ICPR 2012 contest. *J Pathol. Inform.* **4**, 8 (2013)
15. Ciresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. *Med. Image Comput. Comput. Assist. Interv.* **16**(Pt 2), 411–418 (2013)
16. Chen, T., Chef'd'hotel, C.: Deep learning based automatic immune cell detection for immunohistochemistry images. In: Wu, G., Zhang, D., Zhou, L., (eds.) *Machine Learning in Medical Imaging. (Lecture Notes in Computer Science)*, vol. 8689, pp. 17–24. Springer International Publishing, Berlin (2014)
17. LeCun, Y., et al.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86** (11), 2278–2324 (1998)
18. Hubel, D.H., Wiesel, T.N.: Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**(1), 215–243 (1968)
19. Otsu, N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)

20. Nasim, A., Hassan, T., Akram, M.U., Hassan, B., et al.: Automated identification of colorectal glands morphology from benign images. In: International Conference on IP, Computer Vision, and Pattern Recognition IPCV'17 (2017)
21. Alexnet toolbox at Mathworks [ONLINE] Available at: <https://www.mathworks.com/help/nnet/ref/alexnet.html>. Accessed 13 April 2018



A Two-Fold Machine Learning Approach for Efficient Day-Ahead Load Prediction at Hourly Granularity for NYC

Syed Shahbaaz Ahmed^{1(✉)}, Raghavendran Thiruvengadam¹,
A. S. Shashank Karrthikeyaa², and Vineeth Vijayaraghavan¹

¹ Solarillion Foundation, Chennai, India
{syed.shahbaaz,raghav-thiru,vineethv}@ieee.org
² SSN College of Engineering, Chennai, India
shashank16997@ieee.org

Abstract. Recent surge in electricity requirements has propelled a need for accurate load forecasting methods. Several peak load demand forecasting methods exist which predict the highest load requirement of the day. However, Short Term Load Forecasting (STLF) takes precedence owing to the constant load fluctuation over the day, especially in developed cities, and therefore finds more practical and economical use. While statistical methods have largely been used for STLF, contemporary works involving Machine Learning (ML) have seen more success. Such ML methods have made use of several years of data, focused on testing only for a short duration (few weeks), disregarded federal and public holidays when the load demands are erratic, or utilized simulated and not real-time data. This provokes the need for a solution that is capable of forecasting real-time load accurately for all days of the year. The authors of this paper propose a unique two-fold approach to model the training data used for accurate day-ahead hourly load prediction, which also predicts suitably well for federal and public holidays. The New York Independent System Operator's (NYISO) electrical load dataset is used to evaluate the model for the year 2017 with a Mean Absolute Percentage Error (MAPE) of 3.596.

Keywords: Machine learning · Demand forecasting · Short-term load forecasting · Smart grid · Energy management · Energy efficiency

1 Introduction

With development in technology, the energy consumption and the variation in demand over a day especially in the developed nations has been increasing rapidly over the past few decades [1]. These factors have impressed a need upon the electric utilities to become more competent in estimating the load demand in order to ensure continuity and efficiency in supply to avoid large economic losses, of many others. In 2012, the International Renewable Energy Agency (IRENA),

launched a global roadmap - REMAP 2030, which aims to double the share of renewables in the electricity sector by 2030 [2]. Since, the renewable energy generation is not as abundant as that by the conventional sources and due to the intermittent availability, accurate forecasting of demand becomes necessary for the merging of renewables with the grid. To estimate and supply energy, electric utilities in the past have relied only on the meter data. But the growing use of smart appliances, home area networks and plug-in vehicles along with a surge in demand response programs have served as an impediment to the traditional demand estimation [3].

Smart grid offers a solution to this predicament, primarily by the analytical use of consumer data to understand the transition in consumer behavior [4]. Thus, forecasting of load becomes an important aspect of smart grid and with variation of load - for larger geographical regions like cities - becoming sudden and erratic, efficient forecasting of load for *shorter time intervals* will serve as a key to a better smart grid.

Over the last decade, load forecasting methods have transitioned from statistical analysis towards Machine Learning (ML) algorithms. Researchers have found ML algorithms to provide high efficiency in forecasting and easy handling of data. In this paper, the authors use a machine learning approach to forecast the day-ahead hourly load using the electric load data from NYISO [5].

This paper is organized as follows: Sect. 2 presents the relevant research work carried out in load prediction. The load data from NYISO is analyzed with respect to consumption across weeks, seasons and months in Sect. 3, while the methodology of selecting the features and obtaining the training data is given in Sect. 4. Section 5 describes the results obtained by testing the best performing machine learning model, of those selected, on the data of 2017 and also benchmarks these results with contemporary works.

2 Related Work

Day-ahead electricity load forecasting has been performed primarily using two methods: time-series and causal models [6]. Of the two methods, time-series models have been used widely for moderate success in predicting day-ahead load requirements. Autoregressive Moving Average (ARMA) [7] and regression-based Kalman filtering [8] are some examples of time-series models that are found to have acceptable prediction accuracies on day-ahead electricity demands. All these methods use either the past load demands or environmental and non-environmental factors.

Foster et al. defined a set of variables based on historical load, temperature and other weather parameters, and then performed Forward Selection Regression in order to determine the set of features that gives the best accuracy [9]. With a sliding window model that considers these features from the previous year, they obtain a Mean Absolute Percentage Error (MAPE) as low as 2.4. However, they do not perform prediction for holidays, and have excluded 43% of the year's days.

Fan et al. used a hybrid model of Support Vector Regression, Auto Regression and Differential Empirical Mode Decomposition, termed the DEMD-SVR-AR model [10]. Using intrinsic mode functions obtained from DEMD and suitable parameters from SVR, they obtain a prediction model with an MAPE of 5.37. But these results are reported for a testing period of only 13 days in the NYISO dataset, and hence cannot be used in a real-time scenario for day-ahead prediction over a year.

Neupane et al. used an Artificial Neural Network with 3 hidden layers and selected features such as temperature, calendar days, and load prices trained over a period of four years and tested over one year [11]. They obtained a comparatively low MAPE of 4.06.

The authors of this paper propose the use of a simple Gradient Boosting algorithm for day-ahead hourly load prediction with a unique approach to predict load for federal and public holidays, termed as *atypical* days. Regular workdays and weekends are termed *typical* days. To the extent of the authors knowledge, *atypical* days have not been considered in most works, leading to inaccurate representation of their accuracies. The authors test this approach with appreciable results on the load data of New York City - made publicly available by NYISO - for the year 2017.

3 Dataset

The New York Independent System Operator (NYISO) provides electric load data from June 2001, separately, for each of the 11 load zones in the state of New York for both 5-min and 1-h intervals. Of the 11 load zones present, the authors have chosen the load zone N.Y.C. for prediction primarily due to two reasons - N.Y.C. was found to have the highest load consumption over the years as shown in Table 1. Also, N.Y.C. was the only zone which comprised of a single city and thus had unique weather data, whereas each of the other zones had multiple cities with unique weather data for each city. However, NYISO provides electric load data for only the zones mentioned in Table 1. Since the authors intend to predict the day-ahead hourly load, only the hourly data for N.Y.C. was used. The weather related data for the parameters shown in Table 2 was obtained from World WeatherOnline [12].

For the purpose of experiment and evaluation, six years of load and weather data, starting from 2012, has been considered.

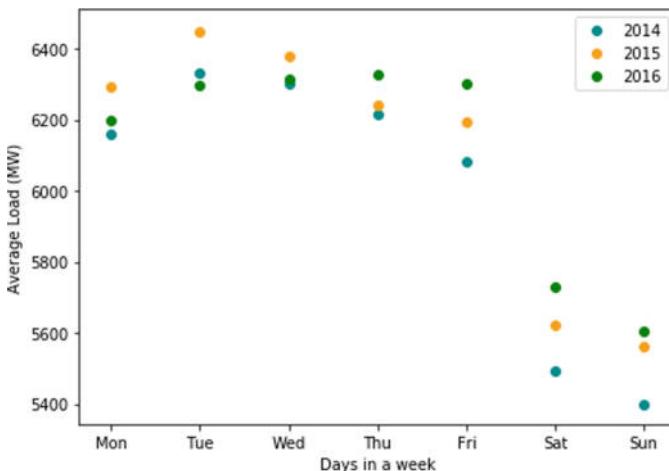
3.1 Load Distribution and Analysis

The load consumption of N.Y.C. exhibits various trends for different days, months and seasons. This section exemplifies these changes and trends with graphical representations for three years from 2014-2016 as these are the years used for validation. The authors would like to highlight that the other years also follow the trends explained further in this section.

Table 1. Average load consumption in MegaWatt (MW) of each load zone in NYISO across different years

Load zones	Avg. load - 2014	Avg. load - 2015	Avg. load - 2016	Avg. load - 2017
CAPITL	1371.0	1418.2	1400.2	1349.8
CENTRL	1866.0	1860.8	1845.0	1806.1
DUNWOD	717.6	719.1	699.0	682.3
GENESE	1130.1	1131.0	1138.0	1116.0
HUD VL	1122.5	1149.1	1135.7	1103.9
LONGIL	2461.8	2501.0	2458.2	2376.4
MHK VL	931.0	929.5	898.8	886.1
MILLWD	307.0	325.0	325.2	329.2
N.Y.C.	5997.1	6106.2	6108.8	5967.1
NORTH	552.0	507.0	499.8	493.4
WEST	1813.6	1799.4	1799.2	1742.3

Day-wise distribution. The average load consumption for each day in a week for the three years is shown in Fig. 1. It can be observed that the average load consumption forms two clusters, with the first containing the weekdays and the second, weekends. A difference of few hundred megawatt (MW) between the average load consumption of each of the clusters accentuates the trend in load consumption across a week, making it a significant indicator of consumption behavior.

**Fig. 1.** Average daily consumption for the years 2014, 2015 and 2016

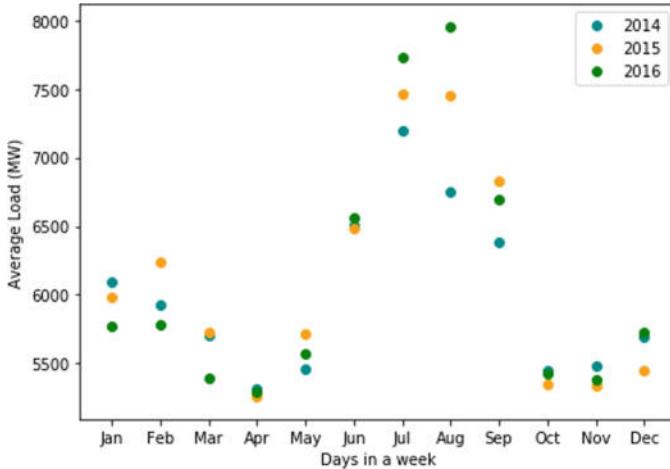


Fig. 2. Average monthly consumption for the years 2014, 2015 and 2016

Seasonal and Monthly distribution. The load consumption also exhibits some macroscopic trends, which are spread across months or seasons over the years. Figure 2 shows the average consumption of every month for the three years. The months in the middle of the year show a digression from the general trend in consumption with an inflection starting from June and peaking in July and August. These months form the entire Summer season and beginning of Fall in New York city. Therefore, it is observed that the average load consumption is similar for the months of the same season and the average consumption of one season is different from the others.

The day-wise trend as explained in Sect. 3.1 is also consistent across different seasons. Figure 3 shows the day-wise trend across four seasons for the three years, where it can be observed that the weekend and weekdays in a particular season form two clusters. Apart from these two clusters, it is evident that the average consumption for a given day is different for every season, thus highlighting the contingency of load consumption on different seasons. Also, these trends were observed to be consistent for the all the years used for testing.

All of these trends have been considered to create the features explained in Sect. 4.

4 Experiment

4.1 Features

The initial list of features selected is shown in Table 2, which was considered to capture the daily, monthly and seasonal trends. Section 4.1 elucidate how the optimal features were selected from the list.

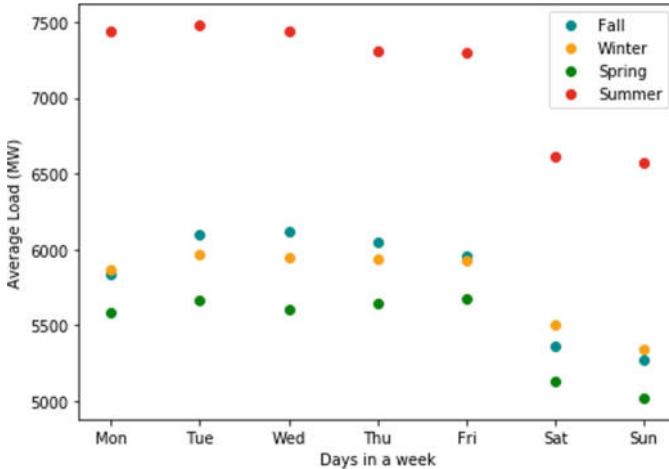


Fig. 3. Average daily consumption for all seasons from 2014 to 2016

Table 2. Feature classification and selection

	Type	All features	Selected features
Calendar variables	Commonly used	Hour, day, month, year	Hour, day, month, year
	Others	Weekends, season	Weekends, season
Weather parameters	Temperature based	Real temperature, wind chill, apparent temperature	Apparent temperature
	Non-temperature based	Humidity, dew point, windspeed	—

Calendar Features. All of the standard features such as *hour*, *day*, *month*, *year* are considered. As described in Sect. 3.1, the double cluster formation indicates a strong trend between the weekdays and weekends which is exploited by adding a feature termed *weekends*. The authors found the inclusion of *weekends* to improve results significantly as observed from Table 5, and hence it was added to the list of commonly taken calendar features. Similarly, a feature - *season*, was also incorporated into the list after observing the seasonal trend in load consumption as explained in Sect. 3.1.

Weather Features. Apart from the dependency of load on time-based features, weather plays a major role on the load consumption behavior. Table 2 shows all the available weather features. The weather feature which was used in the final model, was chosen based on a process of selection and elimination from the initial list. This process was performed based on two factors viz. correlation and impact on the efficiency of the model as explained below.

In order to study the relationship of these features with the load, a correlation matrix, shown as a heat map in Fig. 4 was plotted, showing the correlation coefficients between all the weather features and N.Y.C. load, and those within the features.

Elimination. The first column in Table 3 shows the weather parameters which have low correlation with the load. Apart from low correlation, these feature showed a negative effect on the efficiency when used in the model leading to the direct elimination of these features. Since the primary criteria for elimination or selection of a feature is it's impact on efficiency, the features which had high correlation coefficient but low or no impact on efficiency were also eliminated.

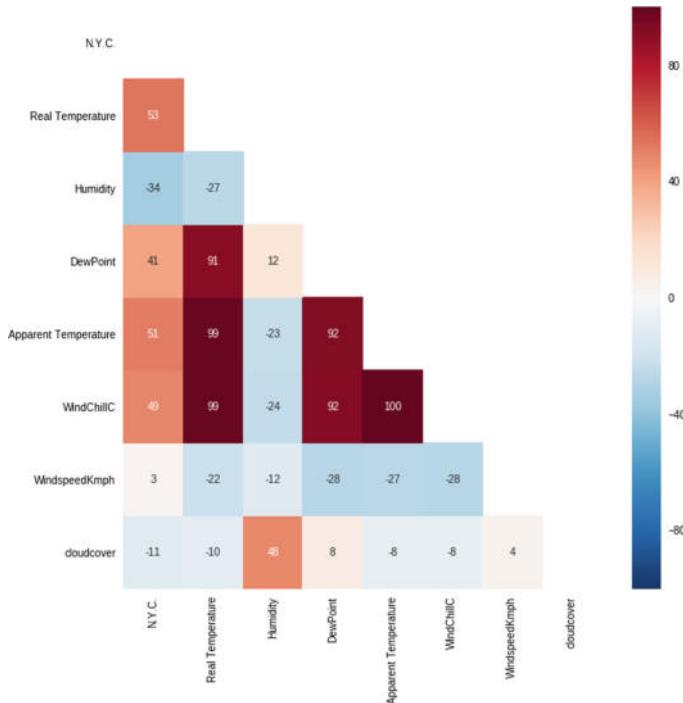


Fig. 4. Correlation between weather features and load

Table 3. Feature selection and elimination

	Low Correlation	High Correlation
Low Efficiency	Cloudcover, wind speed	Wind chill, Humidity, Dewpoint
High Efficiency	-	Apparent, Real Temperature

Selection. The features left after elimination, shown in the green colored cell of Table 3, apart from having a positive impact on the efficiency, have both high correlation coefficient with the load and amongst themselves as is seen from the Fig. 4. This introduces multicollinearity in the model, which can deteriorate the performance of some algorithms. Therefore, to eliminate multicollinearity, each of the two features within the green colored cell of Table 3 were tested individually with the rest of the features for their impact on efficiency. Table 4 shows the MAPE after inclusion of each of these features separately with the other selected features for the years 2016 and 2017. It can be seen that the feature *apparent temperature* gives a lower MAPE and hence it was selected. The authors would like to mention that *real temperature*, which results in a similar MAPE could also be used with a marginal decrease in the performance of the model.

Table 4. Impact of real and apparent temperature on MAPE

Year	MAPE	
	Real temperature	Apparent temperature
2016	3.412	3.409
2017	3.523	3.436

4.2 Training Data

For the purpose of validation and optimization of the model, the data from 2016 is considered. To train any model for load prediction, the study of temporal significance of the data required is necessary. Observing the load trends of typical and atypical days, it can be seen from Fig. 5 that the load for a typical day (Dec 31, 2015) has a high similarity with the load of previous few days (Dec 29 and 30, 2015) and that of an atypical day (Jan 01, 2016) has very low similarity with the load from previous few days (Dec 29, 30 and 31, 2015), which can be attributed to the sudden increase or decrease in electricity consumption on such days when compared to regular workdays. Therefore, using the load data of the previous days for the load prediction of such atypical days would serve only to decrease the effectiveness of the model by a large extent. In order to address this issue, the authors propose a two-fold approach to model the training data used for prediction, which is primarily based on the differing number of training days to be used for typical and atypical days.

Two-Fold Approach A general representation d_k is used to depict the number of training days, where k is the number of consecutive days before the day to be predicted. The day for which the load is to be predicted is represented as d_0 . Here, training with d_k days refers to training the model with all 7 features for each hour of a day for k days.

Micro History. In order to find the optimal number of training days, the authors trained the model for an increasing number of days before the day of

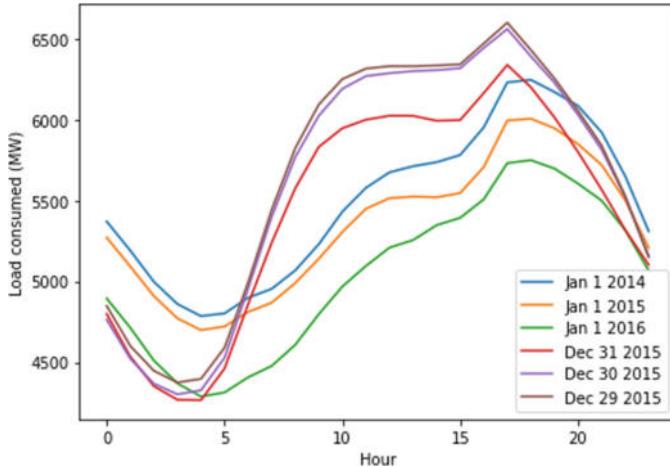


Fig. 5. Load consumption on typical and atypical days

prediction, starting with one previous day. Table 5 shows the average MAPE for different algorithms for select number of training days for the year 2016. It was observed that the Gradient Boosting Regressor, with all features performs the best with a training of d_{10} days as can be observed from Fig. 6. The authors trained the model separately up to d_{14} training days and found that MAPE is higher for $k \neq 10$.

In order to capture the long-term patterns like those spread across seasons or months as explained in Sect. 3.1, the authors used the entirety of the year previous to d_0 (from January to December) as training days, along with d_{10} number of training days. It was observed that this inclusion of training data, reduced the MAPE as seen from Table 5. Since atypical days are erratic and have low correlation with the rest of the days, they are not used in the training of typical days to allow for better prediction.

Thus, the authors define and use *Micro-History* as the previous year's training data along with ten days (d_{10}) prior to the day of prediction that do not contain any atypical days.

Macro History. As explained in Sect. 4.2, the load consumption patterns vary significantly from the previous few days for atypical days. Therefore, using *Micro History* for the prediction of load for an atypical day resulted in erroneous prediction. The prediction for a randomly selected day from the atypical days using *Micro History* is shown in Fig. 7. As shown in Fig. 5, there is a large similarity between the load curve of an atypical day from a year (Jan 1, 2016) with that of the same day from previous years (Jan 1, 2014 and 2015). This trend was also observed for President Day, Labor Day, Columbus Day, Thanksgiving and Christmas.

To capture these patterns, the training days for atypical days comprises of the same day from the previous years. Since there is only one occurrence of a

Table 5. MAPE for various combinations of features

Algorithm	Features	Training period					
		d_2	d_{10}	d_{14}	1 Year + d_2	1 Year + d_{10}	1 Year + d_{14}
Gradient boosting	All features	4.449	3.704	3.638	3.539	3.412	3.464
	Only <i>weekends</i> ^a	4.671	3.710	3.643	3.564	3.433	3.471
	Only <i>season</i> ^a	4.693	4.660	4.706	5.001	4.935	4.810
	No <i>season</i> and <i>weekends</i> ^a	4.719	4.654	4.688	5.078	4.940	4.879
Random forest	All features	4.961	3.862	3.898	3.778	3.631	3.623
	Only <i>weekends</i> ^a	4.952	3.875	3.883	3.723	3.626	3.667
	Only <i>season</i> ^a	5.009	4.545	4.884	4.967	5.059	5.066
	No <i>season</i> and <i>weekends</i> ^a	5.001	4.549	4.885	5.091	5.098	5.078
Extra trees	All features	5.641	3.871	3.813	3.778	3.631	3.623
	Only <i>weekends</i> ^a	5.603	3.833	3.808	3.723	3.626	3.667
	Only <i>season</i> ^a	5.617	4.880	5.020	4.967	5.059	5.066
	No <i>season</i> and <i>weekends</i> ^a	5.680	4.899	5.014	5.091	5.098	5.078
XGBoost	All features	4.696	3.719	3.675	3.552	3.409	3.432
	Only <i>weekends</i> ^a	4.696	3.719	3.675	3.559	3.443	3.439
	Only <i>season</i> ^a	4.696	4.603	4.708	5.015	4.896	4.804
	No <i>season</i> and <i>weekends</i> ^a	4.696	4.603	4.708	5.050	4.932	4.825

^aIncludes the features - hour, day, month, year and apparent temperature, along with the one mentioned in the cell

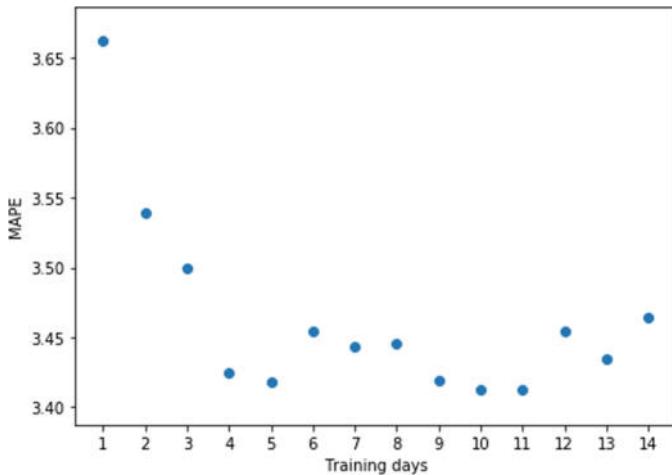


Fig. 6. MAPE for different training days for 2016

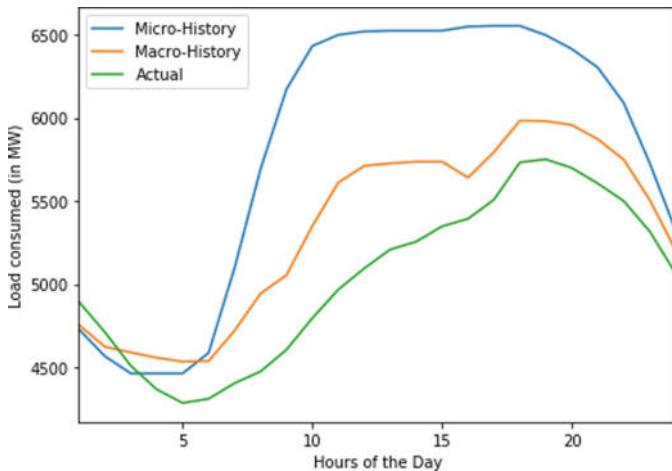


Fig. 7. Load prediction for an atypical day with micro-history and macro-history

In a particular atypical day in a year, the training data for atypical days was the same day from previous four consecutive years. For example, to predict the load demand of New Year 2016, the training data consists of the entire year of 2015, along with the New Years of 2012 to 2015 i.e., a total training data of one year and four days. Figure 7 also shows the predicted load curve for the same atypical day using *Macro-History*, showing a significant improvement in prediction accuracy when compared to prediction using *Micro-History*.

Therefore, for a given day, *Macro-History* is defined and used as the same day from the previous four consecutive years along with the entirety of the previous year.

For both typical and atypical days, addition of more years of training data only reduced the MAPE. The final model uses all the features for the training days as explained in *Micro-History* for typical days and *Macro-History* for atypical days. This shows that only data of the previous year, along with selective days of the current year (for typical days) and selective days from the previous few years (for atypical days) are required for the model.

5 Results

To determine the effectiveness of day-ahead prediction, the authors have considered MAPE as the primary metric. The model was validated and optimized based on the load consumption of the year 2016 as explained in Sect. 4 and was tested for the year 2017. This section shows the results obtained for the year 2017.

5.1 Benchmarking MAPE with Contemporary Works

Using the training data that gave optimal results for 2016, the model is capable of predicting load with an average MAPE of 3.596 across the entire year of 2017. For the *Macro-History*, the same 4 training days (2013–2016) were used instead of (2012–2015). The authors find these results to be better than those obtained by [9–11] in either one or both of the following:

Duration of prediction. The model is capable of predicting everyday of the year suitably well, including federal and public holidays. This is the primary issue with [9, 10] where a significant portion of the year is eliminated or prediction is performed only for a few days.

MAPE. The MAPE obtained by the model is much lower than that presented by [10, 11], but is higher than that of [9]. However, the authors would like to point out that prediction of the proposed model was performed for every day of the year as opposed to predicting for only 57% of the year in [9].

To validate the requirement of one year of data, the model was tested without using the previous year's training data while predicting for 2017. It was found that inclusion of the year's data decreased MAPE by 0.2, thus validating the use of one year of data. Similarly, apparent temperature was found to produce marginally better results, as depicted in Table 4. Hence, the model utilizes only apparent temperature and not actual temperature. Figure 8 shows the actual and predicted load for a typical day, and Fig. 9 shows the actual and predicted load for an atypical day.

Thus, the model performs appreciably well on 2017 after using the two-fold approach on training data and optimizing for 2016. The authors therefore believe that this model can be generalized for any year.

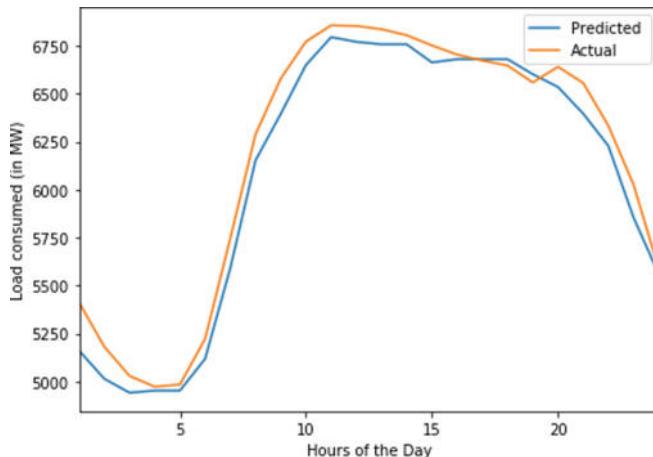


Fig. 8. Actual and predicted load for a typical day in 2017 with an average MAPE of 1.583

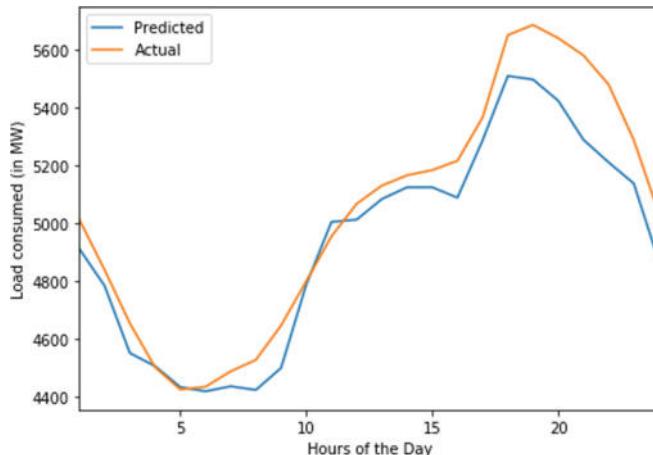


Fig. 9. Actual and predicted load for an atypical day in 2017 with an average MAPE 1.989

6 Conclusion

The authors present an efficient and functional ML model for accurate day-ahead short term load forecasting for all days of a calendar year. The model is capable of predicting the load demands of atypical days with suitable results, which is possible with the implementation of a unique approach in modeling the training data used. The model was validated for the year 2016, and subsequently tested on the entirety of 2017 with appreciable results for both typical and atypical days, thereby validating the approach taken.

Acknowledgements. The authors would like to acknowledge Solarillion Foundation for its support and funding of the research work carried out.

References

1. Energy Information Administration, eia.gov (2018). U.S. Energy Facts- Energy Explained, Your Guide To Understanding Energy- Energy Information Administration. [Online] Available at: https://www.eia.gov/energyexplained/?page=us_energy_home
2. Smart Grid And Renewables: A Guide for Effective Deployment, 2013, p. 47 [Online]. Available: https://www.irena.org/documentdownloads/publications/smart_grids.pdf. Accessed: 23 March 2018
3. How Does Forecasting Enhance Smart Grid Benefits? (White paper), SAS Institute, 2010, p. 9 [Online]. Available: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/how-does-forecasting-enhance-smart-grid-benefits-106395.pdf. Accessed 20 March 2018
4. The Smart Grid: An Introduction. Litos Strategic Communication, 2008, p. 48 [Online]. Available: https://www.smartgrid.gov/files/The_Smart_Grid_Introduction_2008_04.pdf. Accessed 18 March 2018
5. New York Independent System Operator. [Available Online]: <http://www.nyiso.com/public/index.jsp>
6. Hippert, H.S., Pedreira, C.E., Souza, R.C.: Neural networks for short-term load forecasting: a review and evaluation. IEEE Trans. Power Syst. **16**(1), 44–55 (2001)
7. Hafen, R.P., Samaan, N., Makarov, Y.V., Diao, R., Lu, N.: Joint seasonal ARMA approach for modeling of load forecast errors in planning studies. In 2014 IEEE PES T&D Conference and Exposition, pp. 1–5. Chicago, IL, USA (2014)
8. Guan, C., Luh, P.B., Michel, L.D., Chi, Z.: Hybrid Kalman filters for very short-term load forecasting and prediction interval estimation. IEEE Trans. Power Syst. **28**(4), 3806–3817 (2013)
9. Foster, J., Liu, X., McLoone, S.: Adaptive sliding window load forecasting. In: 2017 28th Irish Signals and Systems Conference (ISSC). Killarney pp. 1–6 (2017)
10. Fan, G.-F., Peng, L.-L., Hong, W.-C., Sun, F.: Electric load forecasting by the SVR model with differential empirical mode decomposition and auto regression. Neurocomputing **173**, 958–70 (2015)
11. Neupane, B., Perera, K.S., Aung, Z., Woon, W.L.: Artificial neural network-based electricity price forecasting for smart grid deployment. In: 2012 International Conference on Computer Systems and Industrial Informatics, Sharjah, pp. 1-6 (2012)
12. World Weather Online. [Available Online]: <https://www.worldweatheronline.com>



A Classical-Quantum Hybrid Approach for Unsupervised Probabilistic Machine Learning

Prasanna Date¹(✉), Catherine Schuman², Robert Patton², and Thomas Potok²

¹ Rensselaer Polytechnic Institute, Troy, NY 12180, USA
datep@rpi.edu

² Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA

Abstract. For training unsupervised probabilistic machine learning models, matrix computation and sample generation are the two key steps. While GPUs excel at matrix computation, they use *pseudo*-random numbers to generate samples. Contrarily, Adiabatic Quantum Processors (AQP) use quantum mechanical systems to generate samples accurately and quickly, but are not suited for matrix computation. We present a Classical-Quantum Hybrid Approach for training unsupervised probabilistic machine learning models, leveraging GPUs for matrix computations and the D-Wave quantum sampling library for sample generation. We compare this approach to classical and quantum approaches across four performance metrics. Our results indicate that while the hybrid approach—which uses one AQP and one GPU—outperforms quantum and one of the classical approaches, it performs comparably to the GPU approach, and is outperformed by the CPU approach, which uses 56 high-end CPUs. Lastly, we compare sampling on AQP versus sampling library and show that AQP performs better.

Keywords: Quantum computing · Machine learning · Restricted boltzmann machines · Deep belief networks · MNIST

1 Introduction

Machine Learning and Artificial Intelligence algorithms have shown significant promise in a wide range of applications, and many real-world applications now benefit from them. From smart phone applications to analysis of scientific data sets running on supercomputers, the breadth of benefit from machine learning is nothing short of impressive. Consequently, the push toward enhancing these algorithms and expanding the application usage continues to accelerate. Specifically, increased speed and accuracy are the two key objectives that the machine learning and artificial intelligence community is striving toward.

In this pursuit of increased speed and accuracy, the machine learning and artificial intelligence community has begun searching for computing platforms

that can provide the necessary computational power that is needed for more advanced algorithms. This search has only intensified with the end of Moore's Law, which coincided with the growth of artificial intelligence in real world applications. As a result, from this need for more computational power, the artificial intelligence community experienced a renaissance with the transition of artificial intelligence algorithms to Graphic Processing Units (GPUs) [1], and is now transitioning to GPU-based supercomputers [2, 3].

Despite the renaissance, there are expectations that the limits of GPUs and GPU-based supercomputing will be reached very soon [4]. Thus, interest in extreme heterogenous compute platforms has begun to grow and algorithmic development has started to incorporate the possibility that different aspects of an algorithm may be performed by processing units that are highly tailored to that aspect of the algorithm [5]. Such processing units include neuromorphic processors, Field Programmable Gate Arrays (FPGAs), GPUs, Adiabatic Quantum Processors (AQPs), and gate-based quantum processors. In this work, we focus on AQPs, specifically the D-Wave 2000Q AQP.

For training unsupervised probabilistic machine learning models like Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN), matrix computation and sample generation are the two critical steps. While classical computing platforms (CPUs and GPUs) are good at matrix computations, AQPs are good at generating samples from a probability distribution. We leverage this fact to come up with a hybrid approach for unsupervised probabilistic machine learning. We evaluate our hybrid approach by training RBMs and DBNs on the MNIST (Modified National Institute of Standards and Technology) dataset, and compare it to several classical and quantum approaches. The main contributions of this work are:

1. We present a classical-quantum hybrid approach for probabilistic machine learning, which leverages GPUs for matrix computations and the D-Wave quantum sampling library for generating samples.
2. Using the hybrid approach, we train unsupervised probabilistic machine learning models like Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN) on the MNIST handwritten digits dataset.
3. We compare our hybrid approach to classical and quantum approaches across four performance metrics: matrix computation time, sampling time, training reconstruction error and validation reconstruction error.
4. We compare sample generation on the Adiabatic Quantum Processor (AQP) to sample generation using the D-Wave quantum sampling library [6] for smaller problems, which can be accommodated on the AQP.

Section 2 goes over the literature regarding Quantum Computing, RBMs and DBNs. Section 3 outlines all the classical, quantum and hybrid approaches investigated as part of this work. In Sect. 4, we present our results, and Sect. 5 concludes the paper with directions for future work.

2 Background and Related Work

Quantum Computing is a non-von Neumann computing paradigm which leverages quantum mechanical systems for computation [7]. Owing to its high speed and potential to solve hard problems which conventional computers cannot solve, it is poised to be extremely promising in the near future [8]. Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN) are two machine learning techniques capable of both supervised and unsupervised learning tasks and have been shown to perform extremely well on a myriad of problems [9,10]. We describe relevant literature on quantum computing, RBMs and DBNs here.

2.1 Quantum Computing

There are two major models of quantum computing that have been realized in commercial implementations. The first is “universal” quantum computers or gate-model quantum computers. Universal quantum computers are capable of a broad class of computational problems and are extremely promising for the future of computing [11]. However, though IBM [12], Intel [13], and Google [14] have all pursued implementing quantum computers, these implementations are all fairly small (with the largest being around 50 qubits). Though some powerful computation is available on systems of that size, particularly for machine learning applications, much larger systems will be required to be practically useful.

The second model of quantum computing realized in commercial implementations is the adiabatic quantum computer, most popularly realized in D-Wave systems [15]. Adiabatic quantum computing can perform quantum annealing, and thus operate as quantum annealers. Quantum annealers perform an optimization task to find a minimum of a given objective function using quantum fluctuations. Though experiments on D-Wave have not yet demonstrated quantum supremacy in practice (i.e., the ability for D-Wave to solve problems that classical computers cannot), D-Wave has demonstrated orders of magnitude speed-up over classical solutions like simulated annealing on hard optimization problems [16]. D-Wave systems are much larger than the “universal” quantum computers implemented. In particular, the current implementation of D-Wave (D-Wave 2000Q) supports 2048 qubits.

One of the potential areas for which quantum computers will make a difference in computing is machine learning [17,18]. There are a variety of machine learning approaches that would benefit from quantum approaches like optimization techniques, PCA, and support vector machines [17]. Our key question for both types of quantum computers is what use they currently have for machine learning and data analysis applications. Because of the size limitations of existing “universal” implementations by Google, Intel, and IBM, we focus our attention on quantum annealers, in particular on D-Wave.

2.2 Restricted Boltzmann Machines (RBM)

The Restricted Boltzmann Machine (RBM) is a generative stochastic neural network model that can learn the underlying probability distribution of the data

on which it is trained. RBMs and Boltzmann machines in general are the simplest deep network to quantize [17] and thus are the focus of this work. Data is fed to the RBM through the visible layer (see Fig. 1). For unsupervised learning tasks, the hidden layer learns the underlying features of the data, and for supervised learning tasks, the hidden layer contains the labels or classes.

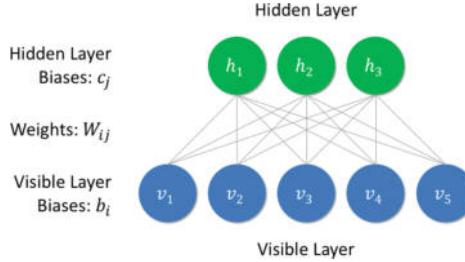


Fig. 1. Restricted Boltzmann Machine (RBM)

We briefly describe the basic theory governing the RBMs here. Let $\mathbb{B} = \{0, 1\}$ be the binary set, \mathbb{R} be the set of real numbers and \mathbb{N} be the set of natural numbers. Given a binary data point $v \in \mathbb{B}^M$ in the visible layer, and its corresponding binary features $h \in \mathbb{B}^N$ in the hidden layer, where $M, N \in \mathbb{N}$, the energy of the configuration (v, h) is computed as follows:

$$\text{Energy } E(v, h; b, c, W) = -b^T v - c^T h - h^T W^T v \quad (1)$$

where, $b \in \mathbb{R}^M$, $c \in \mathbb{R}^N$ and $W \in \mathbb{R}^{M \times N}$ are the RBM parameters to be learned. These are often represented by the set $\theta = \{b, c, W\}$. Given that v is M -dimensional and h is N -dimensional, there are 2^{M+N} possible states of (v, h) . The probability of attaining any of these states is:

$$\text{Probability, } p(v, h; \theta) = \frac{e^{-E(v, h; \theta)}}{Z} \quad (2)$$

$$\text{where, } Z = \sum_{v, h} e^{-E(v, h; \theta)} \quad (3)$$

Because computing Z requires summing over all 2^{M+N} states of (v, h) , it is computationally intractable. Thus, computing the joint probability $p(v, h; \theta)$ is intractable. However, it can be shown that the conditional probabilities can be computed efficiently as follows:

$$p(h|v; \theta) = \sigma(c + W^T v) \quad (4)$$

$$p(v|h; \theta) = \sigma(b + Wh) \quad (5)$$

where $\sigma(s) = \frac{1}{1+e^{-s}}$, represents the sigmoid function. The parameters of RBM (i.e. $\theta = \{b, c, W\}$) can be learned by minimizing the Negative Log Likelihood

Error ($\mathcal{L}(\theta)$) over the training dataset, which is defined as follows:

$$\mathcal{L}(\theta) = \frac{1}{T} \sum_{t=1}^T \left(-\log \sum_h p(v^t, h; \theta) \right) \quad (6)$$

where, $T \in \mathbb{N}$ is the number of datapoints in the training dataset. Computing $\mathcal{L}(\theta)$ requires computing the marginal probability $p(v^t; \theta) = \sum_h p(v^t, h; \theta)$, which requires summing over all 2^N possible configurations of h —this is intractable. However, computing its gradient with respect to θ is slightly better:

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{h|v^t} \left[\frac{\partial E(v^t, h; \theta)}{\partial \theta} \right] - \mathbb{E}_{v,h} \left[\frac{\partial E(v, h; \theta)}{\partial \theta} \right]$$

where $\mathbb{E}_x[y]$ denotes the expected value of y under the condition x . While the first term in the above equation depends only on the data and is tractable, the second term depends on the model, and is intractable. In practice, the second term is approximated by Gibb's sampling.

Now, since computing the Negative Log Likelihood Error ($\mathcal{L}(\theta)$) is intractable, we cannot use it to monitor the progress of training. Instead, we use the Reconstruction Error. It measures how well the learned model is able to reconstruct the input ($\mathcal{R}(\theta)$) by first sampling h^t from the probability distribution $p(h^t|v^t; \theta)$ for a given datapoint v^t , then sampling its reconstructed version v_r^t from the probability distribution $p(v_r^t|h^t; \theta)$, and finally computing the mean squared error:

$$h^t \sim p(h^t|v^t; \theta) \quad (7)$$

$$v_r^t \sim p(v_r^t|h^t; \theta) \quad (8)$$

$$\mathcal{R}(\theta) = \frac{1}{M} \|v_r^t - v^t\|_2^2 \quad (9)$$

RBMs were introduced in the 1980s [19], and popularized with the introduction of the contrastive divergence learning algorithm by Hinton in 2002 [20]. RBMs have been successfully used on a variety of applications, including network anomaly detection [21], collaborative filtering [9], and phoneme recognition [22]. RBMs, like traditional artificial neural networks, can be shown to be universal approximators [23].

2.3 Deep Belief Networks (DBN)

Deep Belief Networks (DBNs) are a type of deep learning network that are composed of stacks of RBMs (see Fig. 2). The weights in a DBN are learned layer-by-layer [24]. DBNs can be used as classifiers or as generative models. Though RBMs individually are universal approximators, deep belief networks (stacks of RBMs) can more efficiently represent complex distributions than a single RBM layer and may also generalize better [23]. Variants of DBNs (e.g., convolutional

deep belief networks) can be used to improve performance for certain data types (e.g., images) [25]. DBNs and their variants have been used successfully on a variety of tasks, including speech recognition [26], natural language understanding [10], and sentiment classification [27].

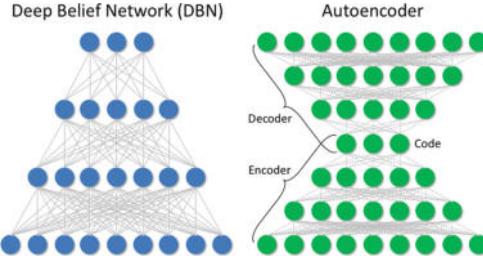


Fig. 2. Deep belief network (DBN) and autoencoder

DBNs as classifiers are comparable to deep learning models such as convolutional neural networks. However, DBNs can also be used as generative models. In this respect, DBNs can also be used as autoencoders as shown in Fig. 2, though there are other types of autoencoders, such as those based on more traditional artificial neural networks or convolutional neural networks that are trained using reconstruction error instead of a classification error [28].

3 Approaches

The larger question that we are trying to ask is how do modern day quantum computers perform on machine learning and deep learning tasks? Does their performance on such tasks compare with classical computing platforms (CPUs and GPUs)? In this work, we evaluate the performance of quantum sampling libraries (which sample from the same probability distribution as Adiabatic Quantum Processors (AQP)) on unsupervised probabilistic machine learning tasks. Specifically, we use the quantum sampling libraries of the D-Wave 2000Q quantum computer to train RBMs and DBNs on the MNIST dataset. From a machine learning point of view, we investigate if RBMs and DBNs trained using this approach can extract features from MNIST images and segregate them into clusters. To explain our approach, we first need to describe various libraries at our disposal. The process of training RBMs and DBNs can be broken down into two tasks: (i) matrix computations; and (ii) sample generation from a known probability distribution, and there are libraries for both.

The libraries used for matrix computations in this work are: (i) Numpy (NP) [29], which is an open source Python library written in Python and C that leverages CPUs for large-scale multi-dimensional array computations; (ii) Google TensorFlow (TF) [30], which is an open source Python library written in Python,

C++ and CUDA that leverages both GPUs (whenever available) and CPUs for dataflow programming; and (iii) D-Wave Matlib (ML) [31], which is a Python library provided by D-Wave to its users, which leverages both GPUs (whenever available) and CPUs for matrix computations. By bundling the Matlib library in their Solver Application Programming Interface (SAPI) 3.0 [32], D-Wave enabled both training and sampling operations for probabilistic machine learning models on CPUs and GPUs without the need for third party libraries.

The libraries used for generating samples from probability distributions are: (i) Native Libraries, i.e., using inbuilt pseudo random number generators in computing libraries; (ii) D-Wave Classical Sampling Library, which generates samples from the classical Boltzmann distribution and runs on both GPUs (when available) and CPUs; and (iii) D-Wave Quantum Sampling Library, which generates samples from the quantum Boltzmann distribution [6], and runs on CPUs.

Table 1. Permutations of compute and sampling libraries

Compute libraries	Sampling Libraries		
	Native samplers	D-Wave Classical (C)	D-Wave quantum (Q)
Numpy (NP)	NP	×	×
TensorFlow (TF)	TF	×	TF-Q
D-Wave Matlib (ML)	×	ML-C	ML-Q

With three computing libraries and three sampling libraries, there are nine permutations for training machine learning models as shown in Table 1. We investigate five of them in this work. These are labeled as follows: (i) NP: Numpy for both computation and sampling; (ii) TF: TensorFlow for both computation and sampling; (iii) ML-C: D-Wave Matlib Library for computation and D-Wave Classical Sampling Library for sampling; (iv) ML-Q: D-Wave Matlib Library for computation and D-Wave Quantum Sampling Library for sampling; (v) TF-Q: TensorFlow for computation and D-Wave Quantum Sampler for sampling. These five permutations are segregated into three approaches: (i) Classical Approach; (ii) Quantum Approach; and (iii) Hybrid Approach.

Classical Approaches: The classical approaches comprise of NP, TF and ML-C approaches. NP is a CPU-only approach that leverages parallel processing on all available cores and has been optimized to great extent. It is one of the most efficient libraries available for CPU computations. In this work, NP serves as the absolute base case to which we compare all other approaches. Because we wanted NP to only serve as the base case from a computation standpoint, we decided not to proceed with the other two Numpy approaches (NP-C and NP-Q). The TF

approach primarily uses GPUs for training and CPUs for auxiliary tasks such as weight initialization, printing output to console, etc. From a computation standpoint, the TF approach is a GPU-only approach. The ML-C approach runs its jobs on a small number of cores throughout the training process (unlike Numpy, which uses all available cores) and uses the classical sampling library for sampling. It must be mentioned that although the Matlib library is capable of running on GPUs, it failed to detect our NVIDIA Quadro P6000 GPU. As a result, in this work, the ML-C approach is categorized as a CPU-only approach that uses the D-Wave classical sampling library for sampling. In this sense, it is different from both NP and TF and is a ‘non-traditional’ classical approach. Furthermore, because the Matlib library lacks parallelism, it was seen as no match to Numpy during our preliminary runs. This is the reason we left out the ML approach for subsequent runs. Thus, under the Classical umbrella, we investigate a CPU-only approach (NP), a GPU-only approach (TF), and a non-traditional classical approach (ML-C).

Quantum Approach: Our quantum approach consists of ML-Q, i.e., Matlib library for computation and quantum sampling library for sampling. The D-Wave 2000Q quantum computer cannot accommodate more than 2048 qubits and not more than 5600 inter-qubit connections, which are further constrained by the Chimera graph that prevents all-to-all connectivity. So, it is not possible to perform large-scale matrix computations entirely on the quantum computer. In this sense, ML-Q is not a *purely* quantum approach, but, we still label it as a quantum approach because it quantifies the gains obtained by using a quantum sampler, when all other computations are carried out using D-Wave’s Matlib computation library. The quantum sampling library stores the RBM/DBN model parameters, which can be updated as training progresses. Given a probability distribution, the quantum sampling library samples from a quantum Boltzmann distribution, which is the same distribution used by the D-Wave 2000Q quantum processor to generate samples.

Hybrid Approach: Our hybrid approach (TF-Q) uses TensorFlow for computation and the quantum sampling library for sampling. TensorFlow leverages GPUs for data flow operations and is fast for matrix computations. D-Wave’s quantum sampling library samples from quantum Boltzmann distribution and is expected to generate samples quickly. Hence, combining these two libraries into a single approach must have some advantages. We objectively study the hybrid approach and compare it to classical and quantum approaches.

We now describe the machine learning task undertaken in this work. Our focus is on unsupervised probabilistic machine learning techniques, specifically Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN). We choose an unsupervised machine learning task because it is the most natural machine learning task for the D-Wave. Given a probability distribution—specifically, a quantum Boltzmann distribution—the D-Wave processor generates accurate samples swiftly. Moreover, unlike traditional compute libraries that use pseudo-random number generators, quantum computers in general, and the D-Wave in particular, use quantum mechanical processes to generate samples, making them *truly* random, thus suitable for security and cryptography applications [33].

The specific unsupervised learning task that we target is to extract features from the MNIST handwritten digits dataset. The MNIST dataset consists of images of handwritten digits and their labels divided into training set (60,000 images), and testing set (10,000 images). The size of each MNIST image is 28×28 pixels. We chose the MNIST dataset as it is the canonical benchmark problem to demonstrate any machine learning or deep learning approach. So, it naturally lends itself for this work, where we want to demonstrate the classical-quantum hybrid approach for unsupervised probabilistic machine learning.

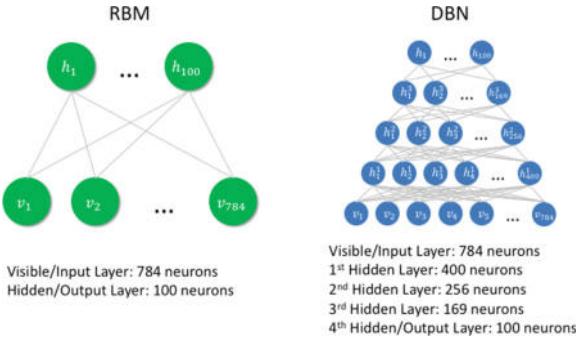


Fig. 3. RBM and DBN architecture

Figure 3 shows the architecture of RBM and DBN models used in this work. All of our RBM models consist of 784 ($= 28 \times 28$) neurons in the visible layer and 100 neurons in the hidden layer. All of our DBN models consist of a visible layer having 784 neurons and four hidden layers containing 400, 256, 169 and 100 neurons respectively.

It must be pointed out that at the time this paper was written, the largest quantum computer available (D-Wave 2000Q) could not accommodate an RBM trained on the MNIST dataset. The reason being, training an RBM with 784 neurons in visible layer and 100 neurons in the hidden layer incurs 78,400 weights ($W \in \mathbb{R}^{784 \times 100}$), and 884 biases ($b \in \mathbb{R}^{784}$ and $c \in \mathbb{R}^{100}$)—a total of 79,284 learning parameters. While the D-Wave 2000Q quantum computer has 2048 qubits, it can only accommodate 5600 inter-qubit connections, making it impossible to accommodate 79,284 learning parameters. So, instead of using the quantum processor directly, we use the D-Wave quantum sampling library. The library samples from the quantum Boltzmann distribution, which is the same distribution that the quantum processor uses to generate samples. Thus, theoretically, if there was a big enough quantum computer that could accommodate an RBM trained on the MNIST dataset, it would sample from the same probability distribution as the quantum sampling library. In order to project how such a future quantum processor would perform, and to compare its performance to that of quantum sampling library, we present sampling results for a smaller set of problems that could be accommodated on the D-Wave 2000Q in Sect. 4.

We evaluate the performance of all approaches mentioned above using the following performance metrics: (i) Matrix Computation Time (MCT): Time spent on matrix computations during training, like forward pass, gradient computations, backward pass, weight updates etc.; (ii) Sampling Time (ST): Time taken to generate samples of MNIST-like images using Eqs. 4 and 5. (iii) Training Error: Reconstruction error (see Eq. 9) computed on the training dataset; and (iv) Validation Error: Reconstruction error computed on the validation dataset. Note that MCT and ST together constitute the total time taken for all computations during training. We do not account for the time taken for non-compute tasks during training, for example, checkpointing the training process, reading data from file, writing weights to a file etc.

RBM and DBN models require binary data for training. However, the MNIST dataset contains values in the range [0, 1]. So, as a preprocessing step, we used rounding to convert the MNIST dataset into a *binary*-MNIST dataset. Since we did this step for all our RBM and DBN models, it is possible to do a fair comparison across all of them. However, rounding the MNIST dataset can accentuate noisy pixels in the images, for example a gray pixel with a value of 0.52 located at the top right corner in a one-digit image would be rounded up to 1.00. In this regard, the *binary*-MNIST dataset is a fundamentally different dataset from the original MNIST, and it would be unfair to compare the accuracies of the trained RBMs and DBNs in this work to the accuracies obtained with state-of-the-art models trained on original MNIST.

4 Results and Discussion

All our experiments were run on a machine that contained 56 Intel Xeon Processor E5-2690 v4 CPUs running at 2.60 GHz base frequency, each of which contained 14 cores, could accommodate 28 threads, and had 35 MB Intel Smart-Cache. The machine had 64 GB RAM, 896 KB L1 cache, 3584 KB L2 cache, and 35 MB L3 cache. The machine also contained an NVIDIA Quadro P6000 GPU, which contained 3840 cores connected via PCI Express 3.0×16 , and 24 GB of GDDR5X memory operating at 432 GB/s memory bandwidth.

Tabel 2 shows the hyperparameters of all RBM and DBN models. We trained the RBM models for 3000 epochs and used a batch size of 128 images. While the base learning rate for NP, TF and TF-Q RBM models was 1.0, it was 0.009 and 0.008 for ML-C and ML-Q respectively. To update the learning rate as training progressed, we used a step policy, in which we decrease the learning rate by multiplying it with a factor called ‘decay’ after every certain number of training iterations called the ‘step size’. For the RBM runs, we used a decay of 0.5 for NP, TF, and TF-Q RBMs and 0.9 for ML-C and ML-Q RBMs. We chose a step size of 500 for NP, TF, ML-Q and TF-Q RBMs, and 600 for the ML-C RBM. Since all RBM models were trained for 3000 epochs, we trained all DBN models for 12,000 epochs as they essentially contained 4 RBM layers each. Furthermore, we used a batch size of 128 for all DBN models. We used a learning rate of 0.3 for NP and TF DBNs, 0.05 for ML-C DBN, 0.03 for ML-Q DBN and 0.4 for

Table 2. RBM and DBN hyperparameters

Model	Hyperparameters	Classical			Quantum	Hybrid
		NP	TF	ML-C		
RBM	# Training epochs	3000	3000	3000	3000	3000
	Batch size	128	128	128	128	128
	Learning rate	1.0	1.0	0.09	0.08	1.0
	Decay	0.5	0.5	0.9	0.9	0.5
	Step size	500	500	600	500	500
DBN	# Training epochs	12,000	12,000	12,000	12,000	12,000
	Batch size	128	128	128	128	128
	Learning rate	0.3	0.3	0.05	0.01	0.4
	Decay	0.9	0.9	0.9	0.9	0.9
	Step size	1000	1000	1000	1000	1000

TF-Q DBN. Finally, we used a decay of 0.9 and a step size of 1000 for all DBN models. The hyperparameters in Table 2 were selected after several rounds of training and were found to be the best for the respective learning models.

4.1 RBM Results

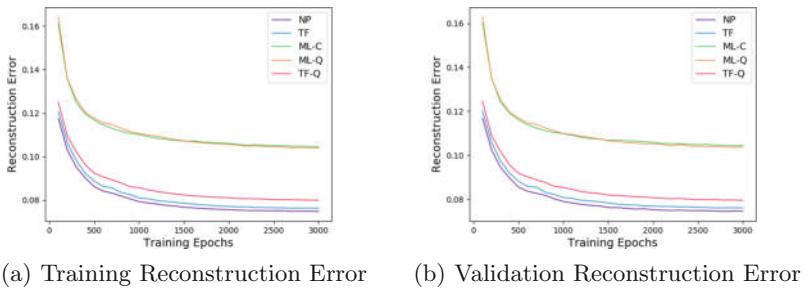
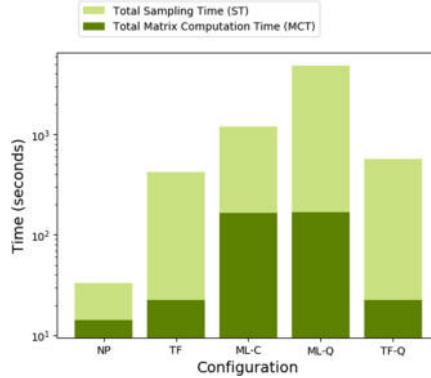
Table 3 shows the performance metrics for all the RBM models. In terms of matrix computation time (MCT), the NP RBM was the fastest (14.19 s), followed by TF-Q (22.25 s), TF (22.41 s), ML-C (163.81 s) and ML-Q (166.61 s). The NP RBM was fastest at generating samples and took 18.65 s, while the others took 395.41 s (TF), 546.05 s (TF-Q), 1020.89 s (ML-C) and 4654.79 s (ML-Q). The training reconstruction error for the NP RBM was the least (0.0748), followed by TF (0.0764), TF-Q (0.0799), ML-Q (0.1040) and ML-C (0.1046). The validation reconstruction error followed a similar pattern: NP (0.0745), followed by TF (0.0760), TF-Q (0.0794), ML-Q (0.1038) and ML-C (0.1044).

Figure 4a and b respectively show the training and validation reconstruction errors as the training progressed. The NP, TF and TF-Q reconstruction errors follow each other very closely, while those of ML-C and ML-Q RBMs are separated by considerable distance. We observed this trend even after training a large number of ML-C and ML-Q RBMs—trying to get their reconstruction errors closer to those of NP, TF and TF-Q. Figure 5 shows MCT and ST for all RBM models on a log-scaled stacked bar graph. While NP RBM took the least ST, ML-Q and TF-Q took similar ST followed by TF. ML-C RBM was the slowest to generate the samples. In terms of MCT, while Numpy and TensorFlow RBMs (NP, TF and TF-Q) took less than 30 s, the Matlab RBMs took over 2.5 min—the Matlab RBMs were about 8X slower than the Numpy RBM, and about 7X slower than TensorFlow RBMs.

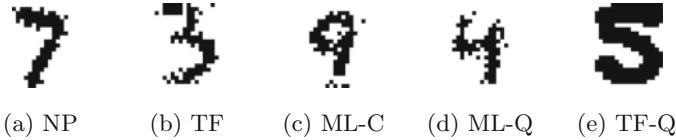
Figure 6 shows sample images generated by all the RBM models. Recall that we used a binary version of the MNIST dataset for training our RBMs. In the

Table 3. RBM and DBN performance metrics

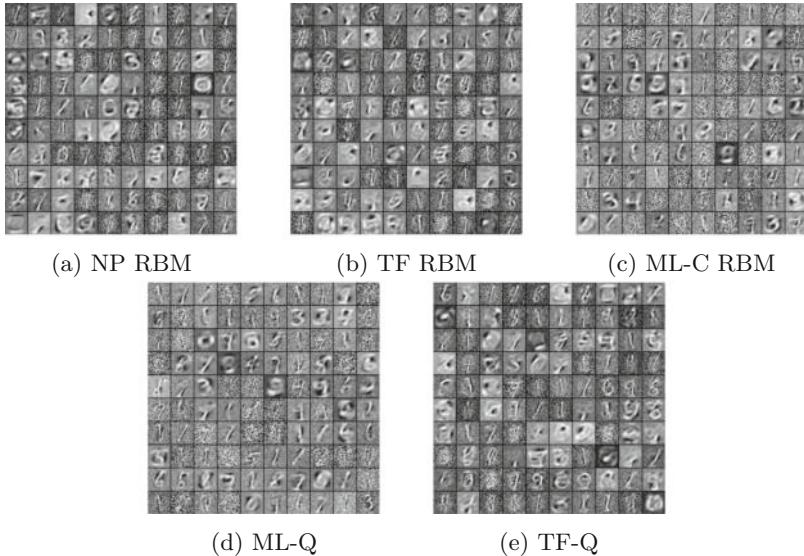
Model	Performance metrics	Classical			Quantum	Hybrid
		NP	TF	ML-C		
RBM	MCT (s)	14.19	22.41	163.81	166.61	22.25
	ST (s)	18.65	395.41	1020.89	4654.79	546.05
	Training error	0.0748	0.0764	0.1046	0.1040	0.0799
	Validation error	0.0745	0.0760	0.1044	0.1038	0.0794
DBN	MCT (s)	190.25	341.30	4301.15	4501.59	340.10
	ST (s)	52.20	1023.2	1932.64	1362.59	1256.66
	Training error	0.0730	0.0721	0.1138	0.1171	0.0715
	Validation error	0.0729	0.0720	0.1140	0.1171	0.0715

**Fig. 4.** Reconstruction errors for RBM models**Fig. 5.** Time analysis for RBM models

binary-MNIST dataset, each pixel is either a zero or a one. This is why the images in Fig. 6 appear pixelated. Furthermore, these images have been generated from Boltzmann probability distributions, which gives the probability of a pixel being equal to one. If this probability is not zero, there is always a chance

**Fig. 6.** Samples generated by RBM models

of bit flips, which produces noisy binary-MNIST images. The hybrid approach (TF-Q), which used the D-Wave quantum Boltzmann sampling library, is seen to produce very robust images. These images, by themselves are a testimony to prefer quantum Boltzmann samplers over classical ones. Figure 7 shows the weights (W matrix from Eq. 1) learned by all the RBM models. Each of the 100 columns of W , having a length of 784, has been first resized into a 28×28 image and then displayed in a 10×10 grid. A quick visual examination of these weights suggests that all RBMs have learned the underlying features of binary-MNIST dataset.

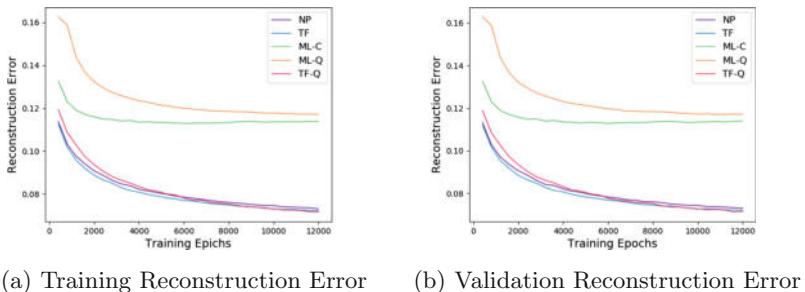
**Fig. 7.** Learned weights for all RBM models

4.2 DBN Results

Table 3 shows performance metrics for DBN models, and a trend similar to the RBM performance metrics is seen. The NP DBN is the fastest DBN in terms of

MCT (190.25 s), followed by TF-Q (340.10 s), TF (341.30 s), ML-C (4301.15 s) and ML-Q (4501.59 s). In this case, while NP takes just over 3 min, TF and TF-Q take just under 6 min and ML-C and ML-Q take over an hour for training. The NP DBN, clocking at 52.2 s, is fastest for sampling as well, followed by TF (1023.2 s), TF-Q (1256.66 s), ML-Q (1362.59 s) and ML-C (1932.64 s). TF-Q is the most accurate DBN with training and validation errors 0.0715 and 0.0715 respectively. This is followed by TF (0.0721 and 0.0720), NP (0.0730 and 0.0729), ML-C (0.1138 and 0.1140) and ML-Q (0.1171 and 0.1171). The TF-Q DBN is also the best model among all RBM and DBN models in terms of reconstruction error.

Figure 8 shows plot of training and validation reconstruction errors. Once again, we see a trend similar to RBMs, the only difference being, at around 6000 and 8000 training epochs, the TF-Q model gets better than NP and TF respectively. Both the Matlab DBNs (ML-C and ML-Q) were outperformed by other models on reconstruction error. Figure 9 shows the sampling time and training time as a log-scaled stacked bar graph. The total time taken by NP is the least, followed by TF and other models. Figure 10 shows the samples generated by the DBN models, which appear less noisy than the RBM models.



(a) Training Reconstruction Error (b) Validation Reconstruction Error

Fig. 8. Reconstruction errors for DBN models

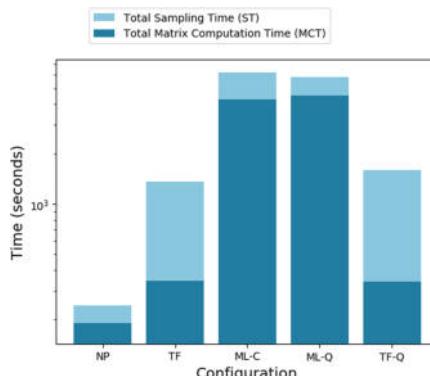
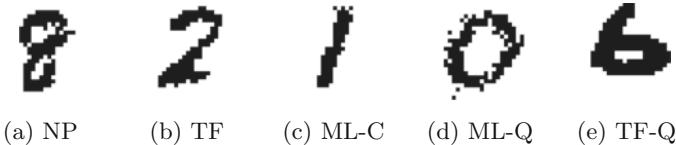
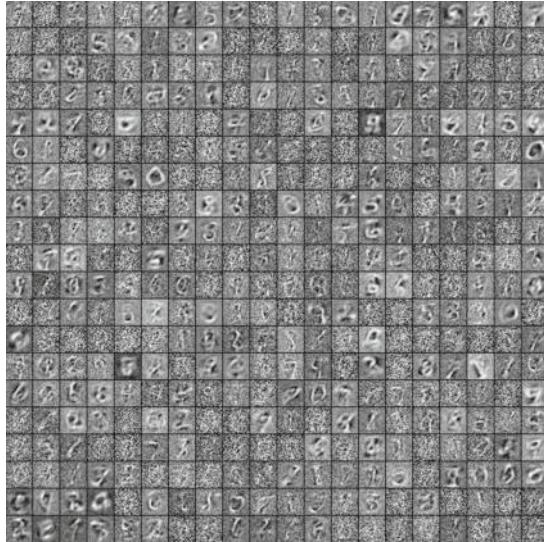


Fig. 9. Time analysis for DBN models

**Fig. 10.** Samples generated by DBN models**Fig. 11.** Learned weights from first layer of ML-C DBN

The samples generated by the quantum sampler still seem more robust than all other samples. Figure 11 shows the weights from the first hidden layer of the ML-C DBN. In order to preserve space and maintain resolution of the image, we present the weights of ML-C DBN only. The weights from first hidden layer of all other DBNs had similar features.

Unlike the RBM weights in Fig. 7, we do not see number images learned by the DBN weights, barring a few. Instead, we observe rather arbitrary lines and curves in Fig. 11. This observation can be explained by the way RBMs and DBNs learn. RBMs have a single hidden layer, and try to recognize the entire digit in one go, resulting in weights resembling numbers. DBNs on the other hand, have multiple hidden layers, and learn features hierarchically. The first hidden layer, as shown in Fig. 11, learns to recognize different lines and curves that make up the digits. In some way, it learns the different ‘pen strokes’ used to write the digits by hand. The subsequent layers aggregate low-level features from preceding layers into more complex features. This hierarchical method of DBN learning produces better samples (Fig. 10) than the RBM samples.

4.3 Discussion

Numpy models consistently outperformed TensorFlow models on MCT and ST. We think this was because Numpy was running on very fast high end CPUs, and that too, 56 of them, containing 14 cores each—a total of 784 cores running at 2.6 GHz. GPU computation with TensorFlow, on the other hand, used 3840 cores running at 1.5 GHz, and incurred a lot of overheads, for instance, copying data to the GPU memory etc., which may have compromised its performance—especially on a small dataset like MNIST, where overhead costs may outweigh any speedup gained. The biggest highlight of this study was that, with just one simulated Adiabatic Quantum Processor (AQP) through the quantum sampling library, we could get a performance comparable to the GPU approach (TF). This is a huge motivation to move towards hybrid approaches for machine learning, as with future improvements in quantum processor technology, even greater speedups could be achieved, and we may even get performance comparable to, or even better than the CPU approach (NP).

Numpy and TensorFlow models significantly outperformed the D-Wave Matlab models on computation tasks. We attribute this difference in performance to lack of parallelism in the implementation of Matlab library. Numpy and TensorFlow use all available cores for computation, while Matlab only uses a handful. Furthermore, while Numpy and TensorFlow have been immensely optimized and are numerically stable, we suspect this may not be the case for Matlab. Matlab can only handle matrices unlike Numpy, which is optimized for multi-dimensional array computation, and TensorFlow, which is optimized for data-flow computation. While Numpy and TensorFlow were introduced in 2006 and 2015 respectively, Matlab was introduced in 2018. With algorithmic and implementation improvements in the future releases, the performance of Matlab could resemble that of Numpy and TensorFlow.

We mentioned in Sect. 3 that an RBM trained on MNIST is too big for the D-Wave 2000Q. To get around this problem, we used D-Wave’s quantum sampling library, which samples from the same probability distribution as the one used by a quantum computer, big enough to accommodate the MNIST RBM. In a way, the quantum sampling library simulates the AQP and when used for training RBMs and DBNs, it must compute $\log(Z)$ from Eq. 3, which requires exponential time. D-Wave’s classical and quantum sampling libraries solve this problem heuristically and end up taking significant amounts of compute time – up to tens of hours. A real AQP on the other hand, could do it in a fraction of second. In order to avoid an unfair comparison between models that use these sampling libraries (ML-C, ML-Q and TF-Q) and have to compute $\log(Z)$, and the ones that don’t (NP and TF), we refrain from reporting the time taken to compute $\log(Z)$ and only report matrix computation time (MCT) and sampling time (ST).

4.4 Sampling Comparison: AQP Versus Quantum Library

In Sect. 3 we pointed out that the D-Wave 2000Q Adiabatic Quantum Processor (AQP) cannot accommodate an RBM or DBN trained on the MNIST dataset.

Here we ask the question: theoretically, if there was an AQP large enough to accommodate these RBMs and DBNs, would it be faster than the quantum Boltzmann sampling library? To answer this question, we perform a scaling study and compare the time taken to generate samples on AQP to sampling using quantum sampling library. We vary the problem size, defined by total number of variables (i.e. number of visible neurons + number of hidden neurons) from 2, 4, 8, ..., 64. 64 is the maximum problem size that can be accommodated on the D-Wave 2000Q AQP. The number of visible and hidden neurons were two-thirds and one third of the problem size respectively.

Figure 12 shows the analysis of various times involved in this comparison. While generating samples on the AQP, we have the sampling time, which is the time spent on the actual quantum annealing process—shown in dark red; AQP overhead time, which is the time taken for post-processing of results obtained from the quantum annealing process at D-Wave’s (server) end—shown in bright red; and, other overhead time, which includes all the other overheads like communication from the client machine to the D-Wave machine and back, time taken to transfer data over the network etc.—shown in pale red. For generating samples using the quantum sampling library, we have the sampling time (dark yellow) and time taken to compute $\log(Z)$ (light yellow) as defined in Eq. 3.

In Fig. 12, we observe that the total time spent in generating samples from the quantum sampling library scales linearly with problem size on a log graph. We observe a similar trend for both sampling time (dark yellow) and $\log(Z)$ time

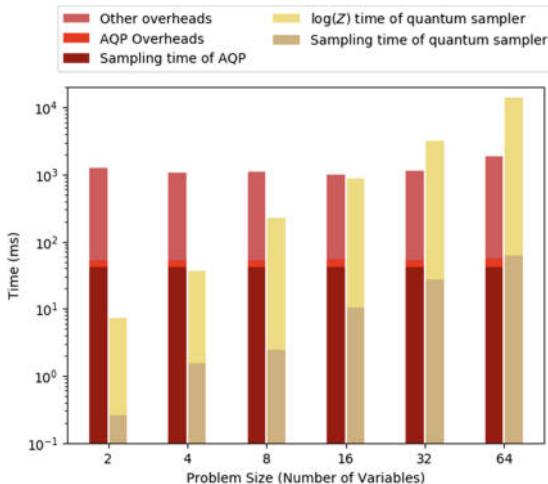


Fig. 12. Sampling time analysis. Red bars show total time taken by AQP, and yellow bars show total time taken by quantum sampling library. Dark red bars show sampling time of AQP, bright red bars show time taken by AQP overhead computations, and pale red bars show time taken for other overheads like communication over network etc. Dark yellow bars show sampling time of quantum sampling library and light yellow bars show time taken to compute $\log(Z)$ as per Eq. 3

(light yellow). On the other hand, the time taken to generate samples on the AQP (sampling time in dark red + AQP overheads in bright red) seems to be constant regardless of the problem size. The other overheads component is seen to increase slightly for the 64 variable case, and this is expected because, as we transfer more data to and from the D-Wave machine, the communication and networking time would increase.

The AQP sampling time (dark red) refers to the time spent in quantum annealing, which lasts for a few microseconds because it is difficult to maintain the qubit states on the AQP for any longer periods of time. Ideally, we would like to have greater annealing time to get more accurate results, but doing so jeopardizes the qubit states, resulting in even worse results. To overcome the problem of short annealing times, quantum annealing is usually done multiple times (up to 10,000 times on D-Wave 2000Q) and the best solution is ultimately picked. The AQP overheads and post-processing time is constant because post-processing is done on the final states of qubits, for example, reading out the qubit state etc. Since number of qubits on the AQP are constant, i.e. 2048 qubits, the AQP overheads time is also constant.

AQP sampling and overheads together consume only about 5% of the total time and about 95% of the total time is spent on communication and networking overheads. Assuming this trend continues, we could generate the MNIST samples in about 50 ms, which roughly is the sum of sampling time and AQP overheads time. This is in the ballpark of sampling time for one RBM training epoch using the NP CPU approach ($14.17 \text{ s}/3000 \text{ epochs} = 4.7 \text{ ms}$). On the other hand, it took several hours to compute $\log(Z)$ using the quantum sampling library as mentioned in Sect. 4. So, the AQP convincingly outperforms the quantum sampling library, especially on larger problem sizes. This provides even more motivation for using hybrid approaches for machine learning.

5 Conclusion

In this work, we proposed a classical-quantum hybrid approach for unsupervised probabilistic machine learning using the D-Wave quantum sampling library in conjunction with GPUs. The hybrid approach (TF-Q) leverages GPUs for matrix operations and uses the quantum sampling library for generating samples. We tested the hybrid approach on the MNIST handwritten digits dataset in an unsupervised learning setting by training Restricted Boltzmann Machines (RBM) and Deep Belief Networks (DBN). We further compared our hybrid approach to classical (NP, TF and ML-C) and quantum (ML-Q) approaches. Our results indicate that the hybrid approach is faster and more accurate than ML-C and ML-Q approaches, performs competitively against the TF approach, and is outperformed by the NP approach. Ideally, we would like to use the D-Wave Adiabatic Quantum Processor (AQP) for generating samples, but size of the current AQP limits the size of problems that can be solved on it. So, to get an estimate of the performance of the AQP, we compared its sampling time to that of the quantum sampling library on smaller problems, which can be accommodated

on the AQP. Our results show that AQP sampling outperforms sampling using quantum sampling library, especially on larger problems. With bigger quantum computers, which can accommodate larger real-world problems, the hybrid approach may outperform or perform competitively against the NP approach. In the future, we would like to do a comparative analysis of all nine approaches in Table 1. We would also like to analyze the performance of convolutional RBMs and convolutional DBNs and compare their performance to the results obtained in this work.

References

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436 (2015)
2. Iandola, F.N., Moskewicz, M.W., Ashraf, K., Keutzer, K.: Firecaffe: near-linear acceleration of deep neural network training on compute clusters. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2592–2600 (2016)
3. Young, S.R., Rose, D.C., Karnowski, T.P., Lim, S.-H., Patton, R.M.: Optimizing deep learning hyper-parameters through an evolutionary algorithm. In: Proceedings of the Workshop on Machine Learning in High-Performance Computing Environments, p. 4, ACM (2015)
4. Kish, L.B.: End of moore’s law: thermal (noise) death of integration in micro and nano electronics. *Phys. Lett. A* **305**(3–4), 144–149 (2002)
5. Potok, T.E., Schuman, C.D., Young, S.R., Patton, R.M., Spedalieri, F., Liu, J., Yao, K.-T., Rose, G., Chakma, G.: A study of complex deep learning networks on high performance, neuromorphic, and quantum computers. In: Machine Learning in HPC Environments (MLHPC), Workshop on, pp. 47–55, IEEE (2016)
6. Amin, M.H., Andriyash, E., Rolfe, J., Kulchytskyy, B., Melko, R.: Quantum boltzmann machine arXiv preprint [arXiv:1601.02036](https://arxiv.org/abs/1601.02036) (2016)
7. Gruska, J.: Quantum computing, vol. 2005. McGraw-Hill London (1999)
8. Rabitz, H., de Vivie-Riedle, R., Motzkus, M., Kompa, K.: Whither the future of controlling quantum phenomena? *Science* **288**(5467), 824–828 (2000)
9. Salakhutdinov, R., Mnih, A., Hinton, G.: Restricted boltzmann machines for collaborative filtering. In: Proceedings of the 24th international conference on Machine learning, pp. 791–798 ACM (2007)
10. Sarikaya, R., Hinton, G.E., Deoras, A.: Application of deep belief networks for natural language understanding. *IEEE/ACM Trans. Audio, Speech, and Lang. Process.* **22**(4), 778–784 (2014)
11. Watrous, J.: Quantum computational complexity. In: Encyclopedia of Complexity and Systems Science, pp. 7174–7201. Springer (2009)
12. Frisch, A.: Ibm qintroduction into quantum computing with live demo. In: System-on-Chip Conference (SOCC), 2017 30th IEEE International, pp. 1–2, IEEE (2017)
13. 2018 CES: Intel advances quantum and neuromorphic computing research’ 2018. <https://newsroom.intel.com/news/intel-advances-quantum-neuromorphic-computing-research/>
14. Terhal, B.M.: Quantum supremacy, here we come. *Nat. Phys.* p. 1 (2018)
15. Johnson, M.W., Amin, M.H., Gildert, S., Lanting, T., Hamze, F., Dickson, N., Harris, R., Berkley, A.J., Johansson, J., Bunyk, P., et al.: Quantum annealing with manufactured spins. *Nature* **473**(7346), 194 (2011)

16. Denchev, V.S., Boixo, S., Isakov, S.V., Ding, N., Babbush, R., Smelyanskiy, V., Martinis, J., Neven, H.: What is the computational value of finite-range tunneling? *Phys. Rev. X* **6**(3), 031015 (2016)
17. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S.: Quantum machine learning. *Nature* **549**(7671), 195 (2017)
18. DeBenedictis, E.P.: A future with quantum machine learning. *Computer* **51**(2), 68–71 (2018)
19. Smolensky, P.: Information processing in dynamical systems: foundations of harmony theory. COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, Tech. Rep. (1986)
20. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14**(8), 1771–1800 (2002)
21. Fiore, U., Palmieri, F., Castiglione, A., De Santis, A.: Network anomaly detection with the restricted boltzmann machine. *Neuro Comput.* **122**, 13–23 (2013)
22. Jaity, N., Hinton, G.: Learning a better representation of speech soundwaves using restricted boltzmann machines. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884–5887, IEEE (2011)
23. Le Roux, N., Bengio, Y.: Representational power of restricted boltzmann machines and deep belief networks. *Neural Comput.* **20**(6), 1631–1649 (2008)
24. Hinton, G.E., Osindero, S., Teh, Y.-W.: A fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2006)
25. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: *Proceedings of the 26th annual international conference on machine learning*, pp. 609–616, ACM, 2009
26. Mohamed, A.-R., Yu, D., Deng, L.: Investigation of full-sequence training of deep belief networks for speech recognition. In: Eleventh Annual Conference of the International Speech Communication Association (2010)
27. Zhou, S., Chen, Q., Wang, X.: Fuzzy deep belief networks for semi-supervised sentiment classification. *Neuro Comput.* **131**, 312–322 (2014)
28. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: International Conference on Artificial Neural Networks, pp. 52–59. Springer (2011)
29. Oliphant, T.E.: A guide to NumPy, vol. 1. Trelgol Publishing USA (2006)
30. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. *OSDI* **16**, 265–283 (2016)
31. D-Wave Systems Inc.: Training probabilistic models using d-wave sampling libraries (2018)
32. D-Wave Systems Inc.: Developer guide for python (2018)
33. Bierhorst, P., Knill, E., Glancy, S., Zhang, Y., Mink, A., Jordan, S., Rommal, A., Liu, Y.-K., Christensen, B., Nam, S.W., et al.: Experimentally generated randomness certified by the impossibility of superluminal signals. *Nature* **556**(7700), 223 (2018)



Comparing TensorFlow Deep Learning Performance and Experiences Using CPUs via Local PCs and Cloud Solutions

Robert Nardelli¹✉, Zachary Dall¹, and Sotiris Skevoulis²

¹ Pace University, Pleasantville, NY 10570, USA

{rn03490p, zd90667w}@pace.edu

² Pace University, New York, NY 10038, USA

sskevoulis@pace.edu

Abstract. Deep Learning is a multilevel learning tool that helps better understand information from various sources. For those who are newly entering the Deep Learning domain, the use of TensorFlow is an optimal way to enhance and learn about neural network designs and a way to deep dive into machine learning. Even though TensorFlow is early with its inception, it is however quickly becoming the direction in which many researchers are using for their open source machine-learning framework. The research, we take advantage of, is the capabilities of TensorFlow to enhance the processing of a substantial dataset that shows the growth and accuracy of cells of embryos in incubation. A typical approach to execute such a massive dataset, today's researchers would first go to directly to the cloud to accomplish this output. From the experiments that ran in this paper, we concluded that an individual afresh to Deep Learning will not always need to proceed to a large GPU cloud solution setup for optional results. The research presented, garners the usage of local PC CPUs, in which we conclude, introduce a beginner to deep learning, less frustration, easier setup time, and better cost initiatives, than one would get if adopting a cloud solution.

Keywords: Machine learning frameworks · Convolutional neural network (CNN) · Deep learning · Artificial intelligence

1 Introduction

In the following sections, we present an introduction to Machine and Deep Learning. TensorFlow, a deep learning framework, was chosen for the experiment to run our deep neural network model tests. Furthermore, the data selected, and setup of our experiment utilized a larger dataset as opposed to the standard MNIST dataset to detect cell generation from month to month.

1.1 Machine Learning

Prior to Machine Learning, it was virtually impossible for humans to study statistical information in debt as compared to the potential of computer processing. It was up until recently we are able to explore unsupervised learning with the power of machines.

By using today's frameworks, machines are able to "learn" without being programmed [1]. This type of learning progressively improves the performance of specific tasks.

Machine learning aids in the improvement of a variety of complex layers of mathematical models and improves various technological advances in the areas of healthcare and security recognition [2].

Artificial Intelligence is where we are moving towards. Machine Learning was a step towards that goal. More recently Deep Learning advanced as a subset of Machine Learning. Deep Learning now moves the needle one step closer towards Artificial Intelligence.

1.2 Deep Learning

Learning is conducted in a few ways such as task-oriented learning or representation of data. Deep Learning represents learning from collecting data, processing it to determine the similarities of the data within a defined parameter of a dataset.

Deep Learning can break down an image in the most simplified form, examine and do comparisons to identify similarities. A deep learning model can map values based on a multilayer perceptron. This multilayer perceptron function maps information read in and values to values already processed. Figure 1 represents this process in an image. The more layers of mathematical functions used, the better the representation outcome can be [3].



Fig. 1. Deep learning image representation model

Artificial neural networks are the backbone of deep learning processing. You can think of this as a waterfall like method in that the top layer is processed and then passes down the information from one layer to the next layer. This is a continuous process becoming more complex as each hierarchy layer is processed until completed [4].

As a result, for the need of more computation power, the benefits of tensors are required to streamline and coordinate rules, so that immense arrays can be utilized to indoctrinate further and use exceedingly complex dimensional data.

1.3 Tensor in TensorFlow

Tensors are objects which, are linear, and used to define relations between various other objects, in addition to other components. TensorFlow utilizes tensors as part of its framework to represent computation of tensors produce values. TensorFlow processes many tensor objects in order to produce these values. As a result, images, for example, can be read as input, captured, processed, converted and compared [5].

1.4 TensorFlow

One of the most common open source Machine Learning frameworks available is TensorFlow. There are many others but TensorFlow prides itself on its vast library of mathematical algorithms, such as neural networks [6]. The user-friendly and power of TensorFlow clearly appeals to developers that want to tap into the power of deep learning and the many computational processes available within the ML framework. Additionally, how it can adapt to other platforms.

TensorFlow, a product of Google, is supported across many platforms, which appeals to more users. Presently there has been a large effort to incorporate TensorFlow within the cloud space. However, many cloud providers are having trouble with the performance of the framework. Presently, the chosen platforms TensorFlow supports are on Windows, Linux and Mac [7].

Architecture within TensorFlow follow the below three operations:

- (1) Types—Design of data utilized by the consumer
- (2) Shapes—The way in which TensorFlow defines dimensionality
- (3) Rank—Corresponds to a mathematical entity (vector, tensor, scalar, etc.).

1.5 Advantages of TensorFlow

The following is a list of advantages using the TensorFlow framework:

- Flexibility of representations.
- Lineage marked from Google.
- Scalable across a multitude of large datasets and machine datasets.
- Can be run against both C++ and Python interfaces.
- CPU and GPU can be used to performance calculations.
- Support for asynchronous and parallel computation.
- Has advanced visualization to view data for Neural networks.
- Benefits huge computational problems.

1.6 Frameworks of Competitors

The competition of TensorFlow has been increasing. These competitors include:

- PyTorch.
- Keras.

- Lasagne.
- Blocks.
- Pylearn2.

1.7 The Experiment

There are many platforms today which software frameworks can run on. These include PC's, Mac's, and cloud. This paper examines and compares each framework by conducting an experiment training the computer to identify mouse embryo cells at different stages of development. The experiment was conducted on a personal computer, MacBook Pro, and two cloud services—Amazon Web Services and Floyd-Hub. Additionally, the user experience was noted, by conducting interviews, on the setup and conducting the experiment, on each platform, using TensorFlow.

1.8 Limitations

The experiment was limited to the frameworks mentioned in the paper using CPU's. GPU's were not used in this study and the results may vary conducting the experiment using them.

2 Problem

There is an increasing number of frameworks being developed that can process machine learning algorithms, specifically Convolutional Neural Networks or CNN. These frameworks, specifically TensorFlow, can run on multiple platforms, such as personal computers as well as the cloud. Additionally, it can be a daunting task for someone to be able to get up and running with these frameworks, not to mention expensive.

This research is aimed to study the performance of TensorFlow running on both cloud and personal computing platforms. In addition, determine the impact on the novice user setting up and executing on these platforms.

2.1 Dataset

There are many datasets out there which are being processed using TensorFlow. One of the most common vetted datasets utilized today for deep learning is the MNIST dataset. As many in the field know there has been plenty of deep research performed utilizing this dataset. However, this dataset is not large enough to really test and compare the performance of various frameworks of deep learning.

This research was conducted to run datasets and compare multiple platforms that exist with TensorFlow. This includes comparisons of cloud performance, in addition to personal computers. To really get a good comparison, the dataset used for comparison was much larger in size than the MNIST dataset. The dataset used for this research was published by Cicconet et al. [3], which is comprised of embryonic cells from mice.

2.2 Image Processing

The pixels of the images are depicted in (Fig. 2a). However, these images were too large and needed to be further processed to be within an acceptable size to be trainable [2]. Additionally, for CNN networks, there needs to be a larger amount of data to be processed for more accuracy. We used a dataset of 3064 images acceptable to train. This is a small amount for training CNN models. One of the popular things to do is to increase the data to be trained. This can be accomplished by making identical copies of the data images. This will provide enough data to ensure the most accurate training of the data.

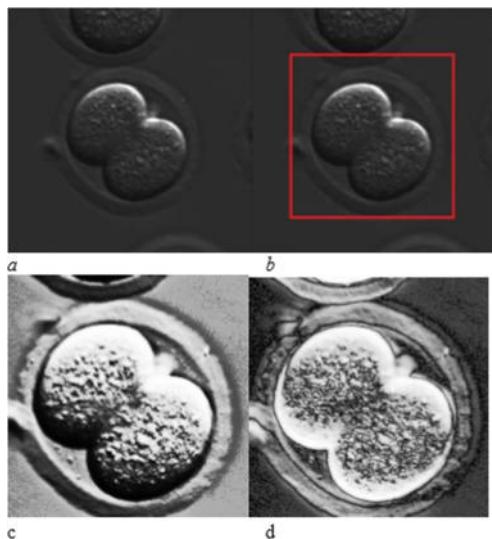


Fig. 2. **a** Original image size. **b** Reposition of the image. **c** Tonal distribution. **d** Lighting enhanced image

The dataset housed irregular shapes and sizes of images, which the dataset creator needed to perform some rotations and manipulation of lighting sequencing. There was no manipulation of the images themselves. They were left intact. By doing this enhancement allowed the training images to be increased by a multiple of 20 [5]. By doing the manipulations and duplicating of the images significantly increased the size of both the testing, training and processing time of the images (Fig. 2). This now would be an acceptable dataset to be used in TensorFlow using the Convolutional Neural Networks (CNN) [2]. This was the main reason this dataset was used for this research as opposed to the much smaller MNIST dataset. The images were in need of a size reduction to enable quicker processing.

2.3 Construction of Convolution Neural Network

The dataset used in this research used six layers, which is depicted in Fig. 3 as opposed to Convolution Neural Network which utilizes three layers for processing [2]. The first layer is convolution, which reads in an image and processes it by a factor of thirty-two. Those output images then go through the first pooling layer, which further refines the image size. The images pass through the next fully connected layer which further reduces and normalizes the image. Finally, the images are at the number ready for producing output.

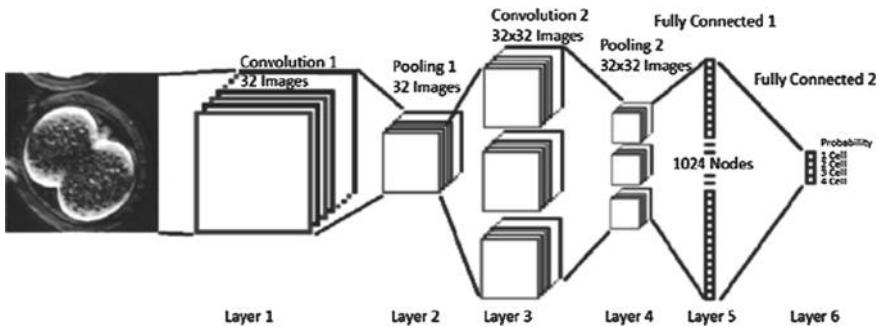


Fig. 3. Construction of convolutional neural network

2.4 Environment Setup

The new technology frameworks that are available for use, vary on their ease of use or user-friendly setups. Performance of these frameworks is an important factor as well. Users can be impacted tremendously based on a variety of deep learning software available. The user needs to be comfortable and gain adequate knowledge of these environments in order to maintain the parameters, such as trainable, pooling, and fully connected layers necessary to produce successful output.

3 Research Approach

3.1 Local Machine Setup

For this research, TensorFlow was used for comparisons on both the personal computing environment, as well as other cloud environments. For the local setups, Mac and PC were used to conduct the experiment and for comparison.

3.2 Environments Utilized

These following Table 1 are the environments used for setup:

Table 1. Experiment environments used

Platform	Microsoft laptop	MacBook Pro laptop	Amazon AWS cloud	FloydHub cloud
Operating system	Microsoft Windows 2010	Mac 10.13.1 High sierra	Linux Ubuntu	Linux Ubuntu
Processor	Intel i5 CPU @ 2.40 GHz	Intel i7 CPU @ 2.5 GHz	High-frequency Xeon processor	High-frequency Xeon processor
RAM	8 GB	16 GB	16 GB	8 GB
System	64-bit	64-bit	64-bit	64-bit

3.3 Local Laptop Setup

TensorFlow 1.5 was used to perform the experiment. The TensorFlow framework was set up on two machines, one is a MacBook Pro and the other a Windows-based machine. Both machines were loaded with the embryo dataset. This dataset contained all the images which will be used to both tests and train the computers. The process would be to read in each image and process it through TensorFlow. In order to accomplish this, there needs to be software to pass the information to the API. To do this Python 2.7 was downloaded to each machine. Python language is used to communicate with the TensorFlow API. Python is the language of choice for many data science applications needing API connections. It is a more simplistic language as opposed to JAVA [8].

The number of epochs was captured and recorded. It is the timing of processing each epoch to test the speed of which TensorFlow was able to train the model.

The local experiment results were captured and presented in Figs. 4 and 5.

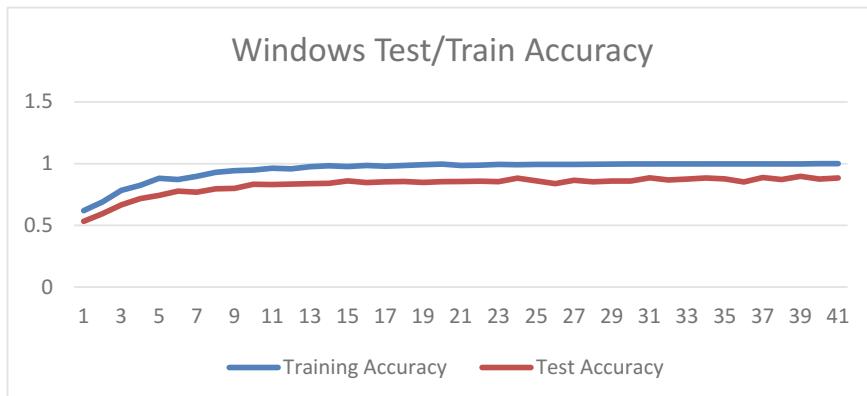


Fig. 4. Results of testing and training accuracy as a result of processing the embryo dataset on local window computer

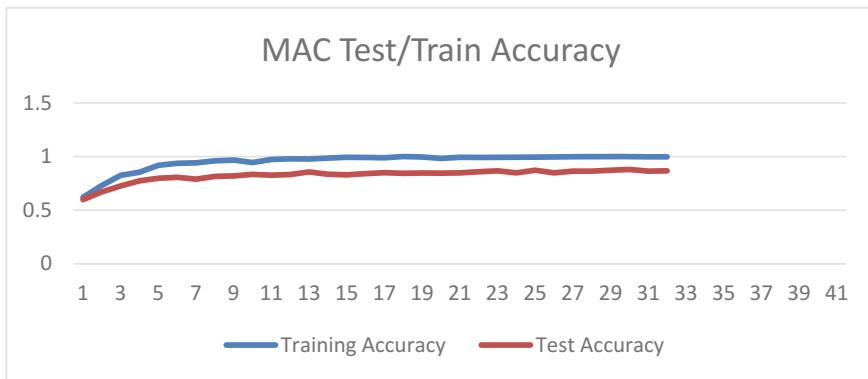


Fig. 5. Results of testing and training accuracy as a result of processing the embryo dataset on local MacBook Pro OS operating system

3.4 Cloud Setup

Cloud computing has been gaining in popularity. There are many different cloud providers today and all are rapidly enhancing their environment to attract customers. Amazon Web services are one of the larger cloud providers in the industry and are very popular with consumers. Amazon has some free services but can get very expensive depending on what you are using and how much processing is needed. Another not so well-known cloud provider is FloydHub. FloydHub offers many of the same services as Amazon, but some of the services Amazon offers for a cost are free with FloydHub. Both FloydHub and Amazon both offer deep learning frameworks along with offering TensorFlow.

As mentioned, both cloud services included both TensorFlow and Python. However, during the setup, it was discovered that both of the packages the user needed to select what version is needed to install. Once installed the packages needed to be updated in the library. For both services in the cloud, there was a need to set up virtual servers following the upload of the dataset to those server spaces. Also, the necessary coding files needed to be uploaded to be executed on the servers.

Figure 6 illustrates both the testing and training accuracy of the dataset in Amazon Web Services. Figure 7 was an attempt to show the same using FloydHub, but as noted the experiment on FloydHub would not run due to the fact it had issues with the amount of RAM that was allocated as it was processing.

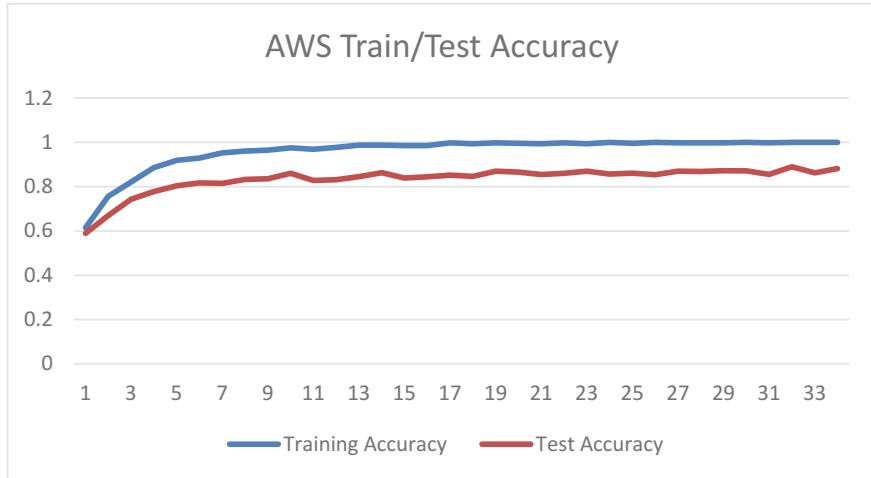


Fig. 6. The accuracy of both testing and training using AWS

```

2018-03-15 16:36:00 PST Waiting for container to complete...
2018-03-15 16:36:01 PST WARNING: This job ran out of memory and was killed. You can lower the memo
Alternatively you can also use a different instance. You can see the list of instances on FloydHub
2018-03-15 16:36:01 PST Creating data module for output...
2018-03-15 16:36:01 PST Data module created for output.
2018-03-15 16:36:01 PST [failed] Task execution failed in 1569 seconds for TaskInstance <TaskInsta
CyW6NbFhrciAXJBhEmhmtN>

```

Fig. 7. Illustration of FloydHub's memory issues

3.5 Experiment Outcomes

The results of the experiment using the four environments were surprising. The expectation is that cloud providers would process the data much more quickly and efficiently.

One of the most surprising outcomes was the performance comparison of the MacBook Pro and AWS. Figure 8 illustrates the performance of training for the PC, MacBook Pro, and AWS. Reminder, FloydHub failed to complete the experiment.

Additionally, the shocking comparison of the number of epochs needed to train, between the MacBook Pro and AWS, were amazingly close as depicted in Fig. 9.

A final observation of performance was the amount of time the PC took to complete as compared to the other platforms in the comparison. Figure 10 illustrates these differences in performances. As expected, the PC took the longest to process, approximately four to seven times longer. More surprising was that the MacBook Pro tested and trained quicker than AWS.

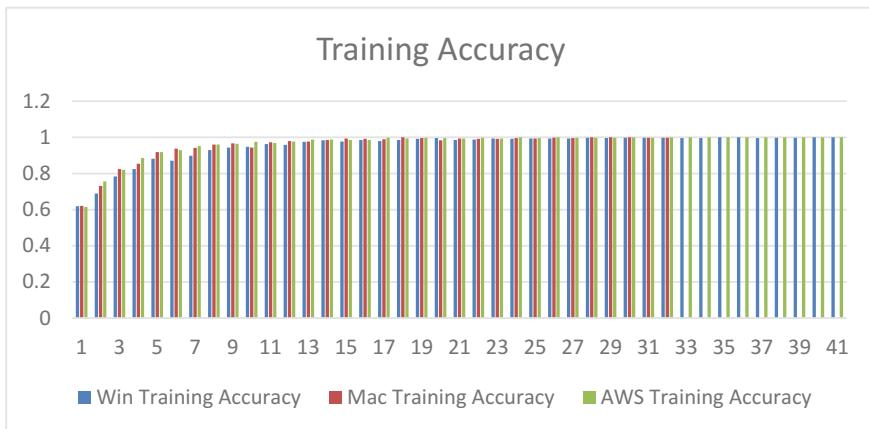


Fig. 8. PC, MacBook Pro, and AWS training accuracy

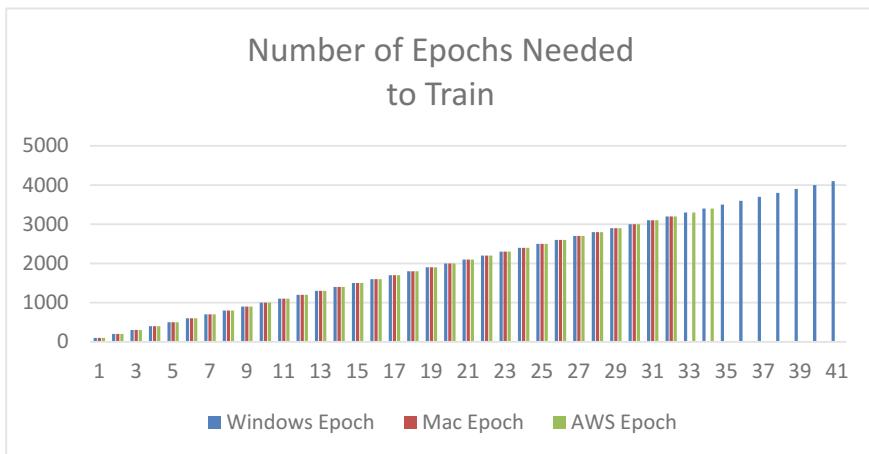


Fig. 9. The number of epochs needed to train for all platforms

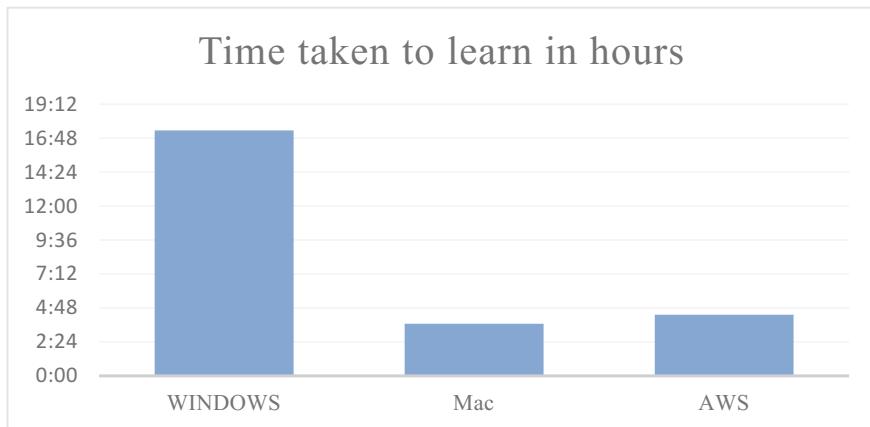


Fig. 10. Execution time

4 User Experience

Part of this research was to measure the impact on the user and their experiences with dealing with all of these different platforms. Especially the newer cloud setups and configurations.

There were many steps to consider when installing frameworks on any platform being used and they are usually not totally the same process to get all components up and running smoothly. The cloud services pride themselves with their solutions because by moving to the cloud, it is cheaper, faster, fewer problems as opposed to dealing with taking on solutions by yourself, or a small, even a larger company. However, this research unveils the real truth about cloud setups versus the personal computing option.

The researcher's, along with other novice users, conducted the experiment on all four platforms. The results were surprising with the various setup options and the process that one would need to go through to be in a position to accurately execute and process a dataset using CNN's. Setting up and installing the frameworks, including TensorFlow, Python and loading the dataset was much more of an easier process on the personal computers as opposed to setting up the same components in the cloud solutions. Both cloud solutions needed to have virtual servers created, set up the memory configurations, ensure the correct versions and the most up to date versions of the packages are installed.

One big hurdle that was unexpected was the ease of loading the dataset on the personal computers as opposed to the cloud. It was a struggle to get this larger dataset loaded into the cloud space. There was no upload file feature in the cloud to just upload it straight from the personal computer to the drive created in the cloud. The best solution was to set up an online free share drive and upload the dataset. The following step was to evaluate the commands on how to move the dataset from the online share drive into the cloud space. This was the same process needed to move the source code into the environment. This was not the easiest of tasks for a novice.

The cloud solution was much more of a challenge to start up and to run. Additionally, in the cloud it was difficult to determine how much executing the experiment would actually cost. You can get some general idea by multiplying the costs by the CPU time it took on the personal computers.

The bulleted items below summarize common pinpoints of setting up and configuring a cloud solution. This is only for the two tested companies of AWS and FloydHub.

- Although there was documentation about the cloud environments, AWS was better documented than FloydHub, it didn't go into the depth needed for a novice computer person to set up and execute an experiment.
- The initial setup and configuration.
- SSH.
- Importing various files.
- Not stable.
- Hidden costs.

Stability and being able to execute an experiment fully are vital to any project. One would not believe in the cloud environment you would have any issues regarding performance in comparison running against a personal computing solution. This research experiment uncovered that cloud performance is not always reliable. In this particular research, the MNIST dataset was proven to execute on FloydHub. We were able to confirm this by conducting our own experiment using the MNIST dataset in the cloud using FloydHub. However, when we used the much larger embryo dataset, there were many performance issues. The first time the experiment ran there were memory issues causing the execution to fail. This was the case after multiple attempts to execute. The coding was changed to limit the number of batches read in at any one time. The execution may be better, but eventually, we saw it crashed due to the same memory issues. The experiment was never able to be completed using FloydHub.

5 Conclusion

This research was an attempt to compare different platforms by conducting a deep learning Convolutional Neural Network experiment using different platforms including personal computers and two cloud computing companies' platforms. This experiment used Python programming language and TensorFlow as the framework.

The result of these experiments supports the claim that using cloud computing is not always necessarily the best solution for performance and ease of use. This research proved that in similar processing and hardware specifications, the personal computers outperformed their cloud counterparts. In fact, the one cloud company, FloydHub, was unable to carry out the experiment. There were many attempts to execute the experiment—by changing parameters to limit the amount of data loaded into memory at one time, but all failed with the same memory issues. In the end, the MacBook Pro stood out to be the best performer of the experiment with both testing and training the dataset images.

This experiment was limited to two personal computing platforms and two cloud computing companies. Future work exists to extend this research by including more cloud companies and possibly using Linux as another operating system for personal computing performance comparison.

References

1. Koza, J.R., Bennett, F.H., Andre, D., Keane, M.A.: Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In: Gero, J.S., Sudweeks, F. (eds.) *Artificial Intelligence in Design '96*. Springer, Dordrecht (1996)
2. Lawrence, J., Malmsten, J., Rybka, A., Sabol, D.A., Triplin, K.: Comparing TensorFlow Deep Learning Performance Using CPUs, GPUs, Local PCs and Cloud. 2017: Pace University Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
3. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. <http://www.deeplearningbook.org>: MIT Press (2016)
4. Murnane, K.: What is deep learning and how is it useful?.2016: Forbes.com
5. Tensor. [cited 2018 April 20]; <http://mathworld.wolfram.com/Tensor.html>
6. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng X.: TensorFlow: large-scale machine learning on heterogeneous systems, 2015. Soft-ware available from tensorflow.org
7. Unruh, A.: What is the TensorFlow machine intelligence platform? Opensource.com (2017)
8. Lukaszewski, L: What Is Python? [cited 2018 April 20]; <https://www.thoughtco.com/what-is-python-2813564>



An Efficient Segmentation Technique for Urdu Optical Character Recognizer (OCR)

Saud Ahmed Malik, Muazzam Maqsood^(✉), Farhan Aadil,
and Muhammad Fahad Khan

Department of Computer Science, COMSATS University Islamabad, Attock
Campus, Islamabad, Pakistan
muazzam.maqsood@cuiatk.edu.pk

Abstract. In Cursive languages like Urdu, segmentation of handwritten text lines is such a problem because of context sensitivity, diagonality of text etc. In this work, we presented a simple and robust line segmentation algorithm for Urdu handwritten and printed text. In the proposed line segmentation algorithm, modified header and baseline detection method are used. This technique purely depends on the counting pixels approach. Which efficiently segment Urdu handwritten and printed text lines along with skew detection. Handwritten and printed Urdu text dataset is manually generated for evaluating algorithm. Dataset consists of 80 pages having 687 handwritten Urdu text lines and printed dataset consist of 48 pages having 495 printed text lines. The algorithm performed significantly well on printed documents and handwritten Urdu text documents with well-separated lines and moderately well on a document containing overlapping words.

Keywords: Urdu OCR · Text line segmentation · Skew detection · Header · Baseline detection

1 Introduction

As in this modern era, computers are overcoming the humans workload in every field now it is very important that computers can read, search and edit the documented text. Unlike humans which are capable of recognizing text very easily from any document/image, machines are not enough intelligent to perceive information from the image. Therefore, a lot of work has been put forward so that to change the document in a form which is understandable for the machine. This need emerges the idea of Optical Character Recognizer.

Cursive Optical Character Recognizer (OCR) remains a challenging problem because of several languages, their variants, and different writing styles. To handle these complexities in the text, a lot of techniques are employed from areas of image processing, Pattern recognition, and artificial intelligence. In the Cursive text, segmenting handwritten text lines is such a problem because of context sensitivity, diagonality of text etc.

In the field of artificial intelligence and Pattern recognition, English language OCR are having good accuracy. Now in English printed OCR perspective, recognizing

English characters is not a big deal. OCRs are too accurate now researchers are only focusing on noisy documents, skew documents etc. But this case is not for other language text script OCRs. Segmentation of Urdu language text is such a hurdle is attaining OCR recognition accuracy. In segmentation of text, line segmentation is the first and most important step as it directly affects word segmentation which further affects recognition of text. To solve this problem a lot of techniques are employed from areas of Pattern recognition and image processing.

In the last two decades, OCR system appeals a lot of researcher's attention towards cursive scripts. But still, OCRs are not accurate for languages like Urdu [1]. Challenges are different for each language problems are more challenging for cursive languages (Arabic, Urdu etc.) because of their writing pattern.

In this work, we presented a line segmentation algorithm for Urdu handwritten and printed text. In the proposed line segmentation algorithm, modified header and baseline detection method are used. This technique purely depends on the counting pixels approach. Which efficiently segment Urdu handwritten and printed text along with skew detection. Handwritten and printed Urdu text dataset is manually generated for evaluating algorithm. Dataset consists of 80 pages having 687 handwritten Urdu text lines and printed dataset consist of 48 pages having 495 printed text lines. Proposed algorithm gives more accurate results as compared to segmenting techniques proposed till date.

1.1 Challenges in Urdu like Scripts

Challenges are different for each language problems are more challenging for cursive languages (Arabic, Urdu etc.) because of their writing pattern, this problem becomes even more complex when it comes to the handwritten document. In OCRs, mostly image segmentation is meant to be the most tricky and error generating step that's why mostly Urdu OCRs use segmentation free approaches [2]. In [3] Discuss the (implicit and explicit) segmentation and their problems for English cursive writing. Explicit segmentation segments the text based on character shape while in Implicit segmentation text is segmented after recognizing a particular character for better results, but it is difficult to train large vocabulary for true segmentation. Mainly segmentation algorithms lacking accuracy because of over-segmentation, under segmentation and inaccurate partition of characters. Some English characters lack in accuracy In [4] use a neural network as post segmentation step, which determines accurate recognition which increases segmentation accuracy.

Characters are context sensitive in Nastalik script. As one character has no particular shape, it changes according to its place in ligature and its next character. Even in one word, the shape of the same character varies according to context, Example of 'ب' is given in Figs. 1, 2 and 3.

Due to diagonal nature of the script, inter and intra-ligature overlapping is a hurdle in attaining good accuracy as shown in Fig. 4.

Many diacritics (secondary components) are merged with ligatures (primary components), It's difficult to identify/segment these components from a single body (Fig. 5). Even placement of diacritics is very important as 17 out of 39 characters are differentiated because of their diacritic placement (Fig. 6).



Fig. 1. character ‘ڻ’ using in mid of ligature



Fig. 2. starting and ending with ‘ڻ’ but having a different shape



Fig. 3. The isolated shape of ‘ڻ’



Fig. 4. Intra-Ligature (Left) and Inter-Ligature (Right) overlapping



Fig. 5. Mergence of diacritics with a ligature



Fig. 6. Different characters having a same secondary component with a difference of diacritics placement

- (1) *Diagonality*: In Urdu text (Nastalik script) when characters join, they form diagonal shapes. Ligature bend towards the left bottom with an angle which depends on the particular character. It generates a problem, when ligature extends to adjacent below the line and touches characters of the next line.
- (2) *Bidirectional nature*: Urdu and Arabic script is written in both directions. Urdu text is written from right to left whereas Urdu numerals are written from left to right direction.
- (3) *Non-Monotonic behavior*: Urdu script is non-monotonic in nature. When it comes to the writing of characters, all character is not fully written in one direction, many characters go back to already written text.
- (4) *Overlapping characters*: In Urdu like scripts overlapping is a difficult task during segmentation. As this script has inter-ligature and intra-ligature overlapping problem. Characters of ligature overlap each other and another ligature vertically.

The main motivation of this work is to make robust text line segmentation algorithm which can handle both printed and handwritten text. OCR has five main stages pre-processing, segmentation, feature extraction classification, and post-processing stage. This paper is focusing on two modules of OCR (Pre-processing and segmentation). A method is proposed for line segmentation and skew correction of printed/handwritten documents. The method consists of the following steps: Pre-processing, Skew correction, Text line segmentation, and ligature\sub-word segmentation phases.

For preprocessing, algorithm separates the background from the foreground of the image with nonuniform brightness. Skew detection algorithm works by extracting the area between text lines and fit it to a straight line. For line segmentation algorithm, modified header and baseline detection method [5] is used. This technique purely depends on the counting pixels approach. And in last, the projection profile method is applied on segmented text lines for extracting ligature.

Till now we have discussed problems in Urdu cursive text script and motivation. Next SECTION consists of an overview of Related work and text segmentation techniques, the further paper is organized as SECTION III we discussed the proposed methodology followed by results and finally in last SECTION, we conclude the paper.

2 Related Work

In purpose of OCR system is to recognize the text with the same accuracy is the human brain is capable of. Recognition stage is directly depending on the segmentation of text. The outcome of preprocessing is not directly fed into the recognition stage as the resultant units are not the basic building blocks of that language. So, in this stage text is converted into the smallest units known as segments. For text, segmentation is classified in three main steps. Firstly, the page is divided into lines. Then divide the line into words/ligatures (One word may contain one or more ligatures and ligature can be a complete word or character) and in last step segment words into smallest units (characters).

Sometimes in some language OCR (having cursive nature script e.g. Urdu), this phase can be omitted in the case when it's hard to segment the smallest unit of text. And segmentation is suffered by the over-segmentation problem, then this stage drops the accuracy of the whole system. In these cases, a holistic approach is used in which instead of segmenting words, recognition is directly done against a dictionary. But in the case of languages with large vocabulary and a large set of characters require too much class in recognition. As the Urdu language contains character set 39 basic characters, mainly consist of letters from the Persian and Arabic character set.

In holistic, several classes are trained for each dictionary word. Because of this drawback holistic approach is not so practical in languages like Urdu, Arabic. Necessary segmentation is required for scripts having many separable units and having the same stroke sequences.

There are two types of segmentation approaches for cursive-like script shown in Fig. 7.

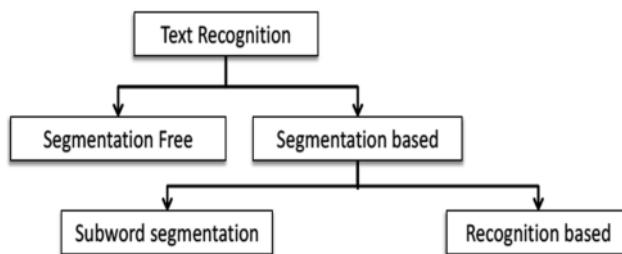


Fig. 7. Types of text segmentation

- (1) Segmentation free: Segmentation approach, which not segment the text to basic character. In this approach, text is divided into word or sub-word and then directly recognized word/sub-word.
- (2) Segmentation-based: Segmentation in which segment the text till smallest recognized piece(character). It is further divided into two parts (Explicit and Implicit segmentation).

Explicit segmentation: Segment word/sub-word into smallest unit of a word (character) by using rules/algorithms.

Implicit segmentation: In this recognition-based approach lines or words are segmented into the smallest unit(character). Segmentation is done at training stage through RNN or hidden Markov model (HMM).

2.1 Literature Review

Latin script is the most widely used script in the world. In the field of OCRs, this script is the most mature as work on Latin script was started from the 1940s [6]. Here we are presenting the latest literature review on state of art methods. Mandal et al. [7]

proposed algorithms for the line, words, and character segmentation. The line is segmented into words by using a contour tracing algorithm from the extreme left of the paragraph. Words are segmented by tracing white pixels in between words. Character segmentation algorithm use Baseline Pixel Burst Method for remaining character segmentation. Some noise removal methods are used to increase accuracy up to 95%.

Inaccurate line segmentation problem is highlighted as a hurdle in the segmentation of Urdu text [8]. Israr Ud Din et al. [8] proposed a segmentation of lines and ligatures as a necessary step in preprocessing. The proposed scheme successfully segments overlapping and touching lines. The 30 documents comprised a total of 310 text lines 306 of which were correctly segmented using the proposed technique reporting a line segmentation accuracy of 98.7%. Naz et al. [9] presented an Urdu Nastalik line recognizer using a multidimensional approach in which it handles diagonal text (Nastalik style). It uses an output layer of the neural network to recognize Nastalik text lines. This approach gives 98% accuracy on printed text. For line segmentation, In addition to some heuristics rules, the projection profile method gives good accuracy [10, 11].

In many Urdu OCR, studies ligature is considered as a basic component to be segmented. Javed and Hussain [12] segmented the thinned strokes in the window and Discrete Cosine Transform (DCT) features are computed for each window. The DCT vectors from the sequence of windows for every segment is being used to teach the HMMs for identification and sequenced to produce the ligature. The system has 92.73% base form recognition accuracy. Diacritics are rarely handled in Urdu segmentation and cause wrong segmentation [13]. Rana et al. [14] use pre-segmented data and then segmented the Urdu words in two segmented categories of primary ligatures and secondary ligatures (consisting of diacritics).

Its difficult task to extract a feature of handwritten text in even English language (in which each character is written separately), when it comes to cursive language like Urdu than its hard to proper segment each ligature. In this section, we review segmentation of handwritten cursive scripts and different methods that are used to overcome complex segmentation problem in the last few years.

Saabni et al. [15] proposed language independent line segmentation techniques. Two different segmentation approaches are proposed for grayscale and binary images of historical Arabic, English and Spanish text [15]. Energy map constructs using Signed Distance Transform for the image to identify text line and upper and lower limit of the text line. Proposed technique tested on four different datasets (ICDAR2009 Contest, Private Collection, ICDAR2009 and Evaluation protocol) and get above 98% accuracy on average. Brodić [16] modifies linear water flow algorithm, by changing its linear function by power function. Waterflow deals linearly straight lines, in this modified algorithm bounding boxes, are added to handle angular dimensions of the text. Bounding boxes contain characters or words, each line of a text document is enclosed in bounding boxes. This results in extracted required line regions.

In [17] Abhishek et al. proposed a segmentation method for handwritten cursive English text. The proposed method uses modified horizontal and vertical projection for line and word segmentation. It even accurately segments multi-skewed lines. The method was tested on IAM dataset giving promising results for line and word (95 and 92% respectively). In [18] Horizontal projection approach is used for text line extraction of Handwritten Kannada Historical Script Documents after the preprocessing

stage and connected component labeling. Projection profile is used to collect text line information and additionally neighborhood search is used to assign text to a line.

To segment palm leaf manuscripts of Dai, Peng et al. [19] use algorithm based on HMM, to evaluate all segmentation paths. Afterward, optimal segmentation paths are computed by projection properties of the corpus. The system evaluated the historical collection of Dai manuscripts gives an accuracy of 89.9%. Projection profile is the most extensively used technique when there is a sufficient gap between lines. In [20] Pastor-Pellicer et al. used the Conventional neural network for text line extraction of historical English documents. Initially, extract line layout analysis and estimates the text area between corpus line and baseline. Further, the Watershed transform is used to extract text lines. This approach is tested on two datasets Saint Gall (98.74%) and Parzival corpora (94.24%). In another study Quang Nhat et al. [21] proposed a multilingual text line segmentation approach. In which trained fully conventional network (FCN) is used to figure out text lines pattern. Through FCN, line map is extracted through which initial segmentation is done and after that line, adjacency graph is used to handle overlapping words between lines. This gives 98.6% accuracy on ICDAR-2013.

3 Proposed Methodology

In this section methodology of the proposed work is discussed. A modified algorithm for solving the line segmentation problem along with skew detection is introduced. It shows that the proposed algorithms outperform (or give comparable performance to) the state-of-the-art algorithms. Furthermore, this section discusses pre-processing and word segmentation techniques used in this work.

This paper is focusing on Preprocessing and segmentation the of OCR. This methodology reflects a simple and efficient line segmentation technique. While the projection profile approach is employed for ligatures/words segmentation. In this paper, pixel counting based simple and robust algorithm is proposed for text line segmentation. Graphical representation of the method is shown in Fig. 8.

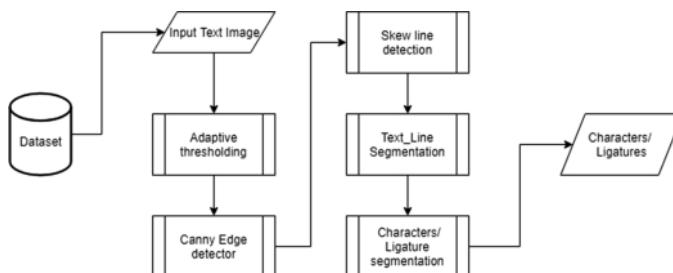


Fig. 8. Proposed methodology

In line segmentation, input image should be noiseless and skew less. The main idea of text line segmentation algorithm is that all rows of the image are scanned. As the

image is a binary image (only containing pixel value as 0 or 1) the intensity of the image is tested for pixels of all rows. In the case of white background, Rows having dark pixels greater than standard deviation are extracted. So, this threshold varies for all images according to their pixel's standard deviation. As line contains several consecutive rows (having dark pixels in case of white background). Consecutive rows having text (dark pixels) are extracted as a line. Two Lines are separated when black pixels are less than the adaptive threshold in rows. These white pixel rows act as the boundary line between two lines. This line segmentation technique is also known as “head and baseline detection”. For extracting a line, this algorithm uses two parameters of line, the head of the line and the end row of the line (baseline). Both are determined by several pixels in rows. In Fig. 9 Consecutive black rows are determined as text line while consecutive white rows are considered as space between two lines. Start of the line is considered as the head of the line and white row after text line is considered as baseline.

ہوتے ہوئے تمام لوگ اپنے اپنے گھر وہ میں بیٹھ جاتے۔ گناہ کی یہ خوشیاں ختم ہوتے ہی اب جو کی فصل کے کام نہیں اور اسے گاہنے کا انتظار رہتا اور اسی کے ساتھ ہی کام کا جن کا باقاعدہ آغاز ہو جاتا تھا اور زمیندار لوگ اپنے کمر کس کے رجیے تھے تھے تاکہ وقت پر فصل اٹھانیا جائے اور اس دن کے بعد ہر گھر کا سر پر سست اپنے آلات کشاد رزی مٹلا پڑھے کے قابلیں (جنہیں بروشکی میں سر مرکب کہا جاتا ہے) کی مرمت کرنے، بھوس اچھانے والے کامتوں (برو شکسکی میں ہر ہڑتے) اور دیگر ضروری آلات کی تیاری میں لگا رہتا تھا اور ان صروفیات کے ساتھ ہی ایک دو تین بد جوکی فصل پک کر تیار ہوتی تھی ہے جو جانشی کے ایک دو ٹھنڈوں کے اندر رکھ کر اپنے اپنے بھتیوں کے خون کے اردو گرد لالج بنا لے اور پھر وقت پر گاہنے کا عمل شروع ہوتا تھا۔ اس طرح جوکی فصل اکٹھی کرنے اور سینچنے کے فور ایعد ہی گندم کی فصل تیار ہوتی ہے اور ماہ اگست میں اس فصل کو اسی طرح سینچنا جاتا ہے اور یوں گری کا یہ موسم گزر جاتا ہے۔

Fig. 9. Printed Urdu text

As in the case of Urdu language diacritics are above the line and they may occupy fewer pixels. The algorithm is designed in such a way that it is lenient to a certain minimum number of black pixels (text) in a white row. But in some cases, the row having less than a minimum threshold, this may affect accuracy by not detecting dot/diacritics in a line. The output of original Urdu printed text image is shown in Fig. 10.

Projection profile is applied on the segmented text line. Vertical axis profile is constructed to segment lines into words. The boundary of each connected dark region in profile is extracted as a separation region. In this way ligatures are segmented.

Proposed method works in a way that Urdu text page is inserted as input, as a line is segmented from the page it directly fed into word segmentation algorithm which divided lines into smallest possible ligatures. In this way, ligatures are automatically arranged in sequence. As shown in Fig. 11a, b.

ہوتے ہوئے تمام لوگ اپنے اپنے کھروں میں قیچ جاتے۔ گناہی کی یہ خوشیاں دھرم ہوتے ہیں اب جو کی فصل کے کام نہیں اور سے گاہنے کا انتظار رہتا اور وہی کے ساتھ ہی کام کرنے کے باقاعدہ آغاز ہو جاتا تھا اور زمیندار لوگ اپنے کمر کس کے راستے چکر کر دلت پر فصل اٹھایا جائے اور اس دن کے بعد ہر گھر کا سر پرست اپنے آلات کشاور زی مثلا پیڑے کے قیلوں (جنہیں بروڈ شکل میں سر ہم کہا جاتا ہے) کی مرمت کرنے، بھوس، اچانے، والے کامتوں (بروڈ شکل میں ہر ہنڑ) اور دیگر ضروری آلات کی تاری میں لگا رہتا تھا اور ان مصروفیات کے ساتھ ہی ایک دو طبقہ بعد جو کی فصل لیکر کر تارہ ہوتی ہی ہے جو لالگ کے ایک دو ہمتوں کے اندر اندر رکھت کر اپنے اپنے کھیتوں کے خرمن کے اوپر گرد لالگ بناتے اور پھر وقت پر گاہنے کا عمل شروع ہوتا تھا۔ اس طرح جو کی فصل اسکھی کرنے اور سینے کے غرائب بعد ہی گندم کی فصل تیار ہوتی ہے اور ماہاگست میں اس فصل کو اس طرح سینا جاتا ہے اور یوں گری کا یہ موسم گزر جاتا ہے۔

Fig. 10. The result of proposed line segmentation algorithm

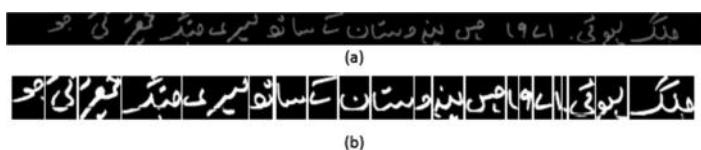


Fig. 11. a Original image b ligature segmented image

3.1 Dataset Generation

Unconstraint Urdu handwritten dataset is collected. Twenty-four writers contribute to the formation of this dataset. Each participant writes a different number of pages (three pages on average). And a total of 80 pages having 687 lines. Several lines in a page and number of words in a line vary throughout the dataset. Some pages have 14 lines, and some have less than 10 lines. Sample data contain almost all types of issues that dataset is having different writing styles vary from writer to writer, each page has a different text size so the database works like a real database. Skewed images are also part of the dataset which makes it more diverse and challenging. The images were captured using a high-resolution digital camera. Then images are scanned and stored in jpg format.

4 Experimental Results

The algorithm is tested on scanned images of all formats PNG, jpg, jpeg, grayscale etc. Above generated dataset (handwritten and printed) is tested for evaluating proposed method. Here we discuss the results of handwritten and printed text. MATLAB 2017a is used for evaluating results. The results are reported in terms of accuracy [22–25].

Dataset is evaluated on a proposed line segmentation algorithm. The algorithm is tested on both printed and handwritten Urdu text. For handwritten data, Urdu handwritten dataset having 80 pages (687 lines) is used to evaluate the line segmentation algorithm. Proposed algorithm segments 687 lines, in which 9 lines are under

segmented and 4 lines are affected by over-segmentation. The algorithm gives 98.1% line accuracy by truly detecting 674 handwritten text lines. For printed data total 48 pages are tested. 48 pages contain 495 lines. Algorithm truly detects 491 lines giving 99.19% accuracy.

The presented algorithm gives 99.42% accuracy on a collection of online pages. On Newspaper paragraphs, the algorithm gives 98.8% accuracy. It detects 85 lines from a collection of 86 lines. The algorithm is also evaluated on 11 scanned pages of the digest. Digest documents and Newspaper both written in Nastalik script.

As the size of text and difference between lines per page varies in digest and Newspaper. That's why several lines in the digest is 131 while in Newspaper it decreases to 86. The algorithm gives 99.23% accuracy on digest paragraphs. It truly segments 130 lines out of 131 lines.

5 Conclusion and Future Work

In this research, the algorithm is proposed which is a script, font size, and font style-independent. This algorithm does not use any specific script knowledge. This paper proposed a line segmentation methodology using a top-down approach which was based on image text pixels and their density. The algorithm performed significantly well on printed text documents and handwritten text documents with well-separated lines and gives good results document containing overlapping words. The main advantage of this algorithm is its ability to detect lines across varying samples.

We make this approach more flexible for handwritten text so that dot/diacritics will remain in concern line and not be the part of adjacent lines.

References

1. Ganai, A.F., Lone, F.R.: Character segmentation for Nastaleeq URDU OCR: a review. In: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT). IEEE (2016)
2. Hussain, S., Ali, S.: Nastalique segmentation-based approach for Urdu OCR. Int. J. Doc. Anal. Recogn. (IJDAR) **18**(4), 357–374 (2015)
3. Rehman, A., Saba, T.: Off-line cursive script recognition: current advances, comparisons and remaining problems. Artif. Intell. Rev. **37**(4), 261–288 (2012)
4. Saba, T., Rehman, A., Sulong, G.: Cursive script segmentation with neural confidence. Int J. Innov. Comput. Inf. Control (IJICIC) **7**(7), 1–10 (2011)
5. Palakollu, S., Dhir, R., Rani, R.: A new technique for line segmentation of handwritten hindi text. Spec. Issue Int. J. Comput. Appl. 0975–8887 (2011)
6. Amin, A.: Segmentation of printed Arabic text. In: International Conference on Advances in Pattern Recognition. Springer (2001)
7. Mandal, R., Manna, N.: Handwritten english character segmentation by baseline pixel burst method (BPBM). Adv. Model. Anal. B **57**(1), 31–46 (2014)
8. Din, I.U., et al.: Line and ligature segmentation in printed Urdu document images. J. Appl. Environ. Biol. Sci. **6**(3S), 114–120 (2016)

9. Naz, S., et al.: The optical character recognition of Urdu-like cursive scripts. *Pattern Recogn.* **47**(3), 1229–1248 (2014)
10. Lehal, G.S.: Ligature segmentation for Urdu OCR. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR). IEEE (2013)
11. Adiguzel, H., Sahin, E., Duygulu, P.: A hybrid for line segmentation in handwritten documents. In: 2012 International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE (2012)
12. Javed, S.T., Hussain, S.: Segmentation based urdu nastalique OCR. In: Iberoamerican Congress on Pattern Recognition. Springer (2013)
13. Muaz, A.: Urdu optical character recognition system MS thesis. Diss. National University of Computer & Emerging Sciences
14. Rana, A., Lehal, G.S.: Offline Urdu OCR using ligature based segmentation for Nastaliq Script. *Indian J. Sci. Technol.* **8**(35) (2015)
15. Saabni, R., Asi, A., El-Sana, J.: Text line extraction for historical document images. *Pattern Recogn. Lett.* **35**, 23–33 (2014)
16. Brodić, D.: Text line segmentation with water flow algorithm based on power function. *J. Electr. Eng.* **66**(3), 132–141 (2015)
17. Bal, A., Saha, R.: An improved method for handwritten document analysis using segmentation, baseline recognition and writing pressure detection. *Procedia Comput. Sci.* **93**, 403–415 (2016)
18. Vishwas, H., Thomas, B.A., Naveena, C.: Text line segmentation of unconstrained handwritten kannada historical script documents. In: Proceedings of International Conference on Cognition and Recognition. Springer (2018)
19. Peng, G., et al.: Text line segmentation using Viterbi algorithm for the palm leaf manuscripts of Dai. In: 2016 International Conference on Audio, Language and Image Processing (ICALIP). IEEE (2016)
20. Pastor-Pellicer, J., et al.: Complete system for text line extraction using convolutional neural networks and watershed transform. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS). IEEE (2016)
21. Vo, Q.N., Lee, G.: Dense prediction for text line segmentation in handwritten document images. In: 2016 IEEE International Conference on Image Processing (ICIP). IEEE (2016)
22. Ateeq, T., et al.: Ensemble-classifiers-assisted detection of cerebral microbleeds in brain MRI. *Comput. Electr. Eng.* (2018)
23. Kalsoom, A., et al.: A dimensionality reduction-based efficient software fault prediction using Fisher linear discriminant analysis (FLDA). *J. Supercomputing*, 1–35 (2018)
24. Khan, S., et al.: Optimized gabor feature extraction for mass classification using cuckoo search for big data e-healthcare. *J. Grid Comput.* 1–16 (2018)
25. Nazir, F., et al.: Social media signal detection using tweets volume, hashtag, and sentiment analysis. *Multimedia Tools and Appl.* 1–34 (2018)



Adaptive Packet Routing on Communication Networks Based on Reinforcement Learning

Tanyaluk Deeka^(✉), Boriboon Deeka, and Surajate On-rit

Ubon Ratchathani Rajabhat University, Ubonratchathani 34000, Thailand
{tanyaluk.d,boriboon.d,surajate.o}@ubru.ac.th

Abstract. An adaptive approach to routing packets on a communication network using machine learning has been reported on our empirical study. We show that the approach of Q-routing previously demonstrated on small toy networks can be expanded to large networks of realistic sizes. The performance of such a routing approach on synthetic networks of three different topology has been studied: random connections, preferential attachment (PA) and a specific architecture known as highly optimized topology (HOT), specifically designed to mimic the Internet's router level topology. Our simulations show that in terms of discovering alternate paths under high loads, the HOT topology is able to offer significant advantage over a PA network which is characterized by hubs at which communication bottlenecks form.

Keywords: Adaptive routing · Preferential attachment · Highly optimized topology · Reinforcement learning

1 Introduction

Routing is a common optimization problem with networked traffic systems, ranging from systems for delivery of goods over road networks to communicating packets of data over electronic networks. In early communication networks, router level algorithms are usually designed to be static, with routing tables stored at every node fixed by computing shortest paths between pairs of source and destination nodes [1]. Dijkstra's shortest path algorithm and the distance vector routing algorithm are common variants of optimization algorithms to compute such routing tables.

With the need to operate networks at ever increasing traffic loads, and the increasing use of more flexible network environments such as wireless communication systems (and the recently popularized notion of the 'Internet of Things' [2]), there is a need for more adaptive (or dynamic) approach to routing in which traffic and network connectivity changes can dynamically determine the routing table. Such thinking leads to more complex optimization problems.

Over the years, interest in the use of artificial intelligence techniques to solve complex and adaptive optimization problems has attracted much interest. Early

work in this area includes the use of the Hopfield network as a basis to formulate the Travelling Salesman problem [3–5]. Reinforcement learning, a branch of machine learning [6], has been a particularly successful technique for formulating and solving difficult optimization problems. An example of this is the elevator arrival optimization formulated as a learning problem in [7]. More recently, approach has also been applied to group image elements on recurrent neural networks for providing a relationship between contour linking and curve-tracing [8].

In the context of routing in communication networks, Boyan and Littman introduced Q-routing, an application of reinforcement learning, in its original formulation as Watkin's Q-learning [9]. This work demonstrated that an adaptive routing table could be learnt and the performance of the network as measured by the average time delay to deliver packets can be improved under heavy loads. However, Boyan et al.'s demonstration was on a very small network of 36 nodes the topology which resembled a grid. Subsequent work in the field has considered networks of around 250 nodes and studied path energy cost and throughput [10–16].

While Boyan et al.'s work [17] is over two decades old, subsequent work on the topic by several authors neither addressed larger networks nor topologies with different connectivities. Since, communication networks have to increase its sizes, and develop its connectivity structures in order to support massive number of users. Hence an empirical evaluation of the performance of Q-routing on networks of realistic sizes and connectivity properties as seen in the Internet is in order. This is the task undertaken in the present study.

In this paper, we consider several network topologies with the number of nodes set at 500 and the number of connections in the network set at 5000. For network of this size, we designed it based on IBM red book [18]. We construct different network topologies with random connections between nodes and connections formed sequentially by preferential attachment [19]. We also consider a novel network architecture, known as a heuristically optimized topology due to Li et al. [20] which is designed to be more reflective of the Internet's router level topology than a preferential attachment network. Since, all traffic from the network edge has to transmit via interconnected routers. By doing these, we show in this paper that the Q-routing approach scales to larger problems of adaptive routing. Our comparison also shows the effect of the different topologies in how much Q-routing can help improve performance when the networks are subject to increasing amounts of traffic.

The paper is organized as follows. Section 2 presents notations and concepts of the Q-routing approach, and how to apply it to get optimal paths for forwarding packets on the network. In Sect. 3, describes how we construct network topologies, and present our implementation and simulation results. Finally, in Sect. 4, we draw conclusions.

2 Q-Routing: The Basics

Routing protocols are designed to find optimal paths for packet transmission in the network which aim to reduce delay time, simplify, and stable. Moreover,

the routing protocols can communicate among nodes in the network by using packet routing which has to specify its neighboring node on the way to destination. Since, the communication networks have always changed in terms of number of users, and its connectivity structures. Hence, the packet routing with static routing tables might not be suitable for these networks. In addition, the packet routing should change its routing tables based on routing information in the network which has an effect on network performance. For example, the routing tables can be changed based on delay time to find destination routes with consuming a minimum time.

A famous adaptive routing is Q-routing which is applied for finding good neighboring nodes to send packets with avoiding traffic congestion [17, 21, 22]. The Q-routing is applied on various network contexts, however it has not applied on large scale networks like Internet network for packet transmission in order to avoid congestion on popular paths. In addition, Internet networks have been growing rapidly, and they also need to find optimal routing algorithm to guarantee quality of services and resilience. Hence, this is a good chance to apply the Q-routing algorithm on the Internet network models which should be resilient when the network traffic is increased, and select optimal paths for avoiding traffic congestion which guarantee quality of services of the networks.

Since, the Q-routing is improved from Q-learning which is one method of Reinforcement Learning (*RL*) for routing proposed, and it is designed based on a *RL* framework in order to achieve its goal. In addition, a number of packets has to be generated continuously until traffic congestion occurs on the Internet networks. Hence, each node on the network has to find the best next neighbouring node to sent packet to its destination, and should avoid traffic congestion.

Let $N = \{1, 2, \dots, 500\}$ is a set of nodes or states (s) in the Internet network which can make routing decisions by observing the delay time between source and destination on the network (environment), and then assigned as a routing table in terms of Q-values in order to control routing policy. The set of actions (a) is the set of neighbouring nodes which can forward packets to its destination. According to Boyan et al.'s work [17], the best estimated delivery time of packet P between source node x and destination node d via neighboring node y can be represented as $Q_x(d, y)$ which should take minimum delivery time between node y and node d for forwarding packets. In addition, node x has to receive routing information feedback which is estimated time between neighboring of node y and destination node d to make a routing decision for forwarding packet P via node y . As the following equation represents estimated minimum time of node y .

$$t = \min_{z \in y} Q_y(d, z) \quad (1)$$

where z is neighboring of node y . Consider, if the packet P cannot serve immediately and it has to spent in node x 's queue before serving which the queueing time of packet P represented as q . In addition, the transmission time between node x and node y represented by s which the estimated delivery time between node x and node y shown as Eq. 2.

$$\Delta Q_x(d, y) = \eta(q + s + t - Q_x(d, y)) \quad (2)$$

where η is a learning rate, and terms of $(q + s + t)$ and $Q_x(d, y)$ represent new estimated time and old estimated time respectively.

Due to the Q-routing has to change its routing table based on estimated delivery time of neighbouring nodes. Finally, we will get the $Q_x(d, y)$ tables which each node uses to select its neighboring node for avoiding traffic congestion in the network.

3 Internet Network Model

In this paper, three different network topologies are compared to study how connectivity has an effect on network performance such as queueing delay time. Furthermore, OMNET++ 4.3.1 network simulation program is employed to build these networks, and runs on the University super computer (Iridis 4) which is high performance computing, and it has 750 compute nodes with dual 2.6 GHz Intel Sandybridge processors. Each compute node has 16 CPUs per node with 64 GB of RAM.

3.1 Network Topologies

In this section, three representative sample of structural network models namely random network, random network with preferential attachment and heuristically optimal topology are considered.

Random Network The random network as shown in Fig. 1 is a basic network model which is given a fixed number of nodes and connected each link between pairs of node with probability p . A connectivity process of random network creates a giant component which has attracted a lot of networking research to study its phase transition properties [23, 24].

Random Network with Preferential Attachment In most real networks such as the collaboration and citation networks continually grow a network size by adding nodes and edges according to a power-law distribution [25–27]. In these networks, new nodes prefer to connect with an existing node which has high number of connections as new nodes are added to the network depending on probability proportional to the current node number of connections. This process is called preferential attachment as shown in Fig. 2, and the pseudo code of these network construction is given in [19].

Heuristically Optimal Topology The heuristically optimal topology (*HOT*) as shown in Fig. 3 is designed based on combining the technological and economic issues in order to apply for the network core and the network edge planning [20]. Due to all traffic from the network edged has to be transmitted through the network via interconnected routers which leads to have heavy congestion on core of the network. In addition, the transmission delay will be increased

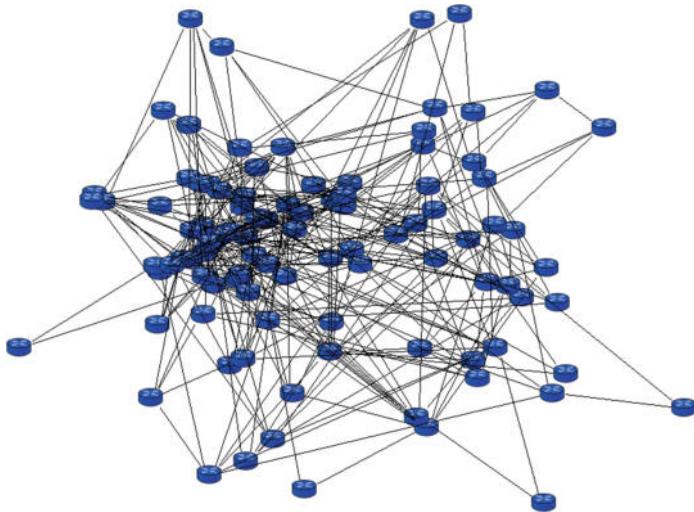


Fig. 1. Random network

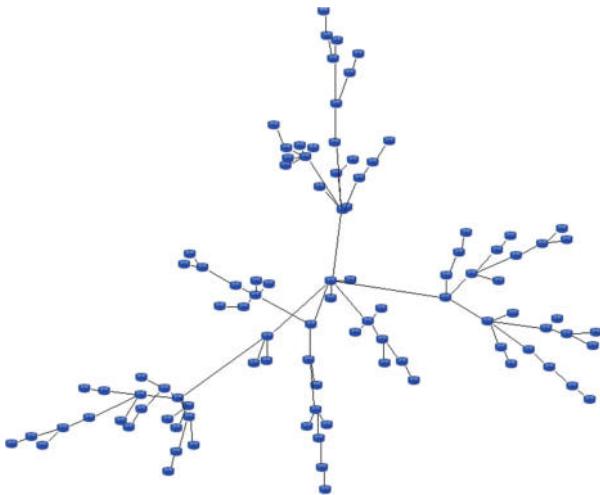


Fig. 2. Random network with preferential attachment

if the network edges far from its destination. Hence, the HOT topology is also designed to minimize the distance between the network core and edge in order to minimize transmission time. Li et al. [20] suggested that the HOT topology is structural three network layers: core, gateway and edge routers. Furthermore, the HOT topology should represent a power-law distribution which shows relationship in the connectivity between AS-level and router-level. Hence, the first step to create HOT topology is to generate a random network with preferential

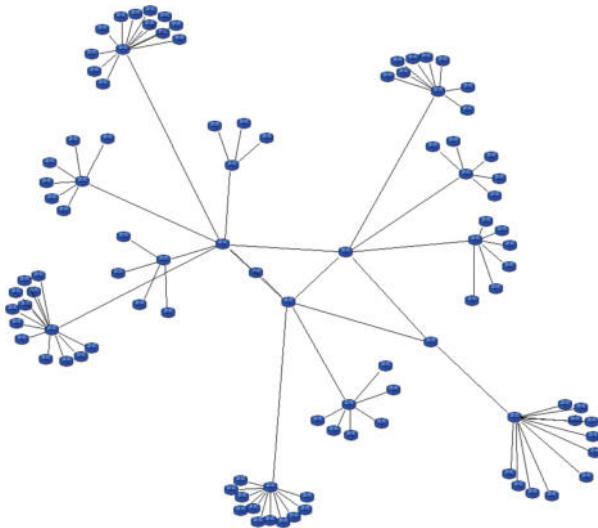


Fig. 3. Heuristically optimal topology

attachment, and then rewire the network connectivity in order to create three structural layers. Due to the core of network has to contain heavily congestion, so it should have low connectivity which its speed can be increased to improve network performance, and it also save cost to maintenance. The gateway routers are connected with the core of network by selecting the other higher-degree nodes, and then connected the edge of network according to the degree of each gateway.

4 Experiments on Synthetic Networks

This section presents the experimental studies which introduces experimental setting, and provides experimental results and analysis.

4.1 Experimental Settings

These experiments are intended to demonstrate the ability of the Q-routing algorithm for packet transmission in term of average delay time and distribution of queue lengths, and how they are tolerant of traffic congestion under different load levels on three network topologies.

In this paper, we set a size of packet based on Ethernet jumbo frames which expanded frame sizes from the original standard IEEE 802.3 in order to reduce the effect of TCP frame overhead [28]. The frame size of packet starts from 1526 bytes and should less than 11,455 bytes because of limit of Ethernet's error checking. However, size of packet frames has an effect on transmission delay in Ethernet link [28]. Since, we consider number of packets which is generated to the network as a result of traffic congestion, and it is called load levels. The

load levels are also increased based on size of packet frames. For example, load level 1 and 6 contain different frame sizes which are 1526 bytes and 9156 bytes, respectively. Hence, a load level 6 generates a sixfold of 1526 bytes at a time.

Furthermore, each node generates packets are periodic which are sent to all over nodes in the network. Each packet specifies its destination, and it is sent out following a routing table. Moreover, the simplest queueing model M/M/1 is embedded in each node to store multiple packets with unbounded FIFO queue. In this paper, we observed delay time which can tell us how long the packet has to spend time in the queue until it can be transmitted over the link in the network. The performance of using Q-routing is compared with the shortest path algorithm.

4.2 Experimental Results and Discussions

Figure 4 is a comparison of average delivery time between Q-routing and shortest path algorithm while the number of packets is increased to make traffic congestion. It can be clearly seen that the Q-routing can decrease maximum delay time at load level 6 on three network topologies, and it has slightly different on queueing delay time at load level 1 because of no traffic congestion. In addition, the Q-routing can find the same routes as the shortest paths after convergent time which is the reason why the average delay time at low load level is not different. Moreover, it can decrease delay time 60.33 and 58.30% at load level 6 when the PA and the HOT are compared with the random network, respectively. In addition, the PA network contains highest delay because some nodes on the network connected with large number of connections, contrasting with some nodes has only a single way to transmit packets as a result of traffic congestion.

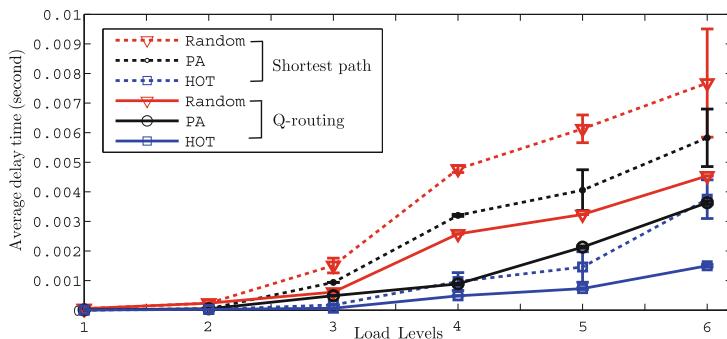


Fig. 4. Comparing average delay time between shortest path and Q-routing

Furthermore, comparing average delay time at load level 6 as shown in Fig. 5 which compared a high load level between the shortest path and the Q-routing algorithms on three network topologies, it can be clearly seen that the Q-routing algorithm can decrease average delay time 59.46, 37.93, and 40.78% on the Random, PA, and HOT networks respectively because the Q-routing algorithm is

embedded on each node which reflects current traffic condition by using its Q-values table for making routing decision, and then selects optimal paths for reducing traffic congestion.

Figure 5 observed distributed of queue lengths between load levels 1 and 6 where the Q-routing algorithm is employed for packet transmission on three network topologies, and it is clearly seen that distribution of queue length for each link on random network holds smallest number of queue length at both of load levels because the random network is built by connected each node with the same probability 0.04, and leads it has the same number of node degree connection. However, the PA network is different from the random network with new node prefers to connect existing nodes with higher degree of connection, so it leads this network has been wildly growing up only one side, and this is cause why this network holds highest number of queue length at both load levels when compared with the rest of networks. In addition, the HOT network is constructed from rewiring node degree connection of the PA network, so it can reduce traffic congestion and contains lower queue lengths when compared with the PA network, but it holds higher queue length than the random network because the traffic will congest at core of routers which connected with lowest node degree connection.

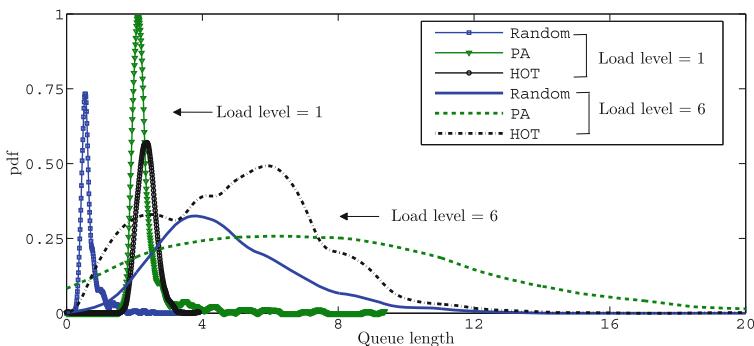


Fig. 5. Distribution of queue lengths between load levels 1 and 6 on three networks

In summary, the shortest path algorithm is not appropriate for forwarding packets if there are a large number of packets would like to be sent into the network because it uses static routing table for packet transmission, and also leads to easily get traffic congestion since it always used the same path for packet transmission. Moreover, the Q-routing can find the same routes as the shortest paths if it learns until convergent time. Hence, the Q-routing algorithm is appropriate for forwarding packets especially if the number of packets is steadily increasing because its routing table can be updated to select suitable it's neighboring node in order to avoid congested paths to send the packet to its destination which can decrease average delay time, and contain large number of packets as shown good results on previous section.

5 Conclusion and Future Work

In this paper, an adaptive routing strategy has been explored for communication networks based on the machine learning methodology of reinforcement learning. The idea, known as Q-routing was introduced over two decades ago. However that work and all work citing it has been on small toy example networks. In this work, we have shown that the Q-routing approach can scale up to realistic router level networks with 500 nodes and 5000 links between them. While the preferential attachment construct is seen as the popular model of several natural and man-made networks including the Internet, a recent suggestion of router level topology is the HOT (Highly Optimized Topology [20]) topology. In comparing networks of random preferential attachment and HOT network topology with respect to adaptive routing, we demonstrate how a random network achieves the best improvement in reducing average delay at high loads because it is easier to find alternated routes. The HOT topology, being a more realistic model of Internet routing is able to outperform the PA architecture significantly, suggesting adaptive routing is a strategy that may be deployed on real networks operating under heavy loads.

In our current work, we are interested in exploring adaptive routing strategies for ad hoc mobile networks including routing in the context of the Internet of Things'. We are also interested in more efficient algorithms in the class of RL, such as the SARSA algorithm [29] and performance optimization strategies with resource limitations (e.g. finite buffer sizes at nodes).

References

1. Tanenbaum, A.S., Wetherall, D.J.: Computer Networks. Pearson (2010)
2. Atzori, Iera, A., Morabito, G.: The internet of things: a survey. *Comput. Netw.* **54**(15), 2787-2805, (2010)
3. Hopfield, J.: Neurons with graded response have collective computational properties like those of two-state neurons. *Proc Nat. Acad. Sci USA* **84**, 3088–3092 (1984)
4. Aiyer, S.V., Niranjan, M., Fallside, F.: A theoretical investigation into the performance of the Hopfield model. *IEEE Trans. Neural Netw.* **1**(2), 204–215 (1990)
5. Smith, K.A.: Neural networks for combinatorial optimization: a review of more than a decade of research. *INFORMS J. Comput.* **11**(1), 15–34 (1999)
6. Sutton, R.S., Barto, A.G.: Reinforcement Learning: an introduction. The MIT Press (2018)
7. Crites, R.H., Barto, A.G.: Improving elevator performance using reinforcement learning pp. 1017–1023 (1996)
8. Brosch, T., Neumann, H., Roelfsema, P.R.: Reinforcement learning of linking and tracing contours in recurrent neural networks. *PLOS Comput. Biol.* **11**(10), 1–36 (2015)
9. Watkins, C.J., Dayan, P.: Q-learning. *Mach. Learn.* **8**(3–4), 279–292 (1992)
10. Haraty, R.A., Traboulsi, B.: MANET with the Q-routing protocol. In: ICN The Eleventh International Conference on Networks, pp. 187–192 (2012)

11. Maleki, M., Hakami, V., Dehghan, M.: A reinforcement learning-based bi-objective routing algorithm for energy harvesting mobile ad-hoc networks. In: IST The Seventh International Symposium on Telecommunications, pp. 1082–1087 (2014)
12. Bhorkar, A.A., Naghshvar, M., Javidi, T., Rao, B.D.: Adaptive opportunistic routing for wireless ad hoc networks. *IEEE/ACM Trans. Networking (TON)* **20**(1), 243–256 (2012)
13. Lin, Z., van der Schaar, M.: Autonomic and distributed joint routing and power control for delay-sensitive applications in multi-hop wireless networks. *IEEE Trans. Wirel. Commun.* **10**(1), 102–113 (2011)
14. Santhi, G., Nachiappan, A., Ibrahim, M.Z., Raghunadhan, R., Favas, M.: IEEE. Q-learning based adaptive qos routing protocol for manets. In: 2011 International Conference on Recent Trends in Information Technology (ICRTIT), pp. 1233–1238 (2011)
15. Hu, T., Fei, Y.: Qelar: a machine-learning-based adaptive routing protocol for energy-efficient and lifetime-extended underwater sensor networks. *IEEE Trans. Mobile Comput.* **9**(6), 796–809 (2010)
16. Dowling, J., Curran, E., Cunningham, R., Cahill, V.: Using feedback in collaborative reinforcement learning to adaptively optimize manet routing. *IEEE Trans. Syst. Man, Cybern.-Part A* **84**, 3088–3092 (1984)
17. Boyan, J.A., Littman, M.L.: Packet routing in dynamically changing networks: A reinforcement learning approach. *Adv. Neural Inf. Process. Syst.* 671–678 (1994)
18. Murhammer, M.W., Lee, K.K., Motallebi, P., Borgi, P., Wozabal, K.: IP Network Design Guide. IBM (1999)
19. Batagelj, V., Brandes, U.: Efficient generation of large random networks. *Phys. Rev. E* **71**(3), 1–13 (2005)
20. Li, L., Alderson, D., Willinger, W., Doyle, J.: A first-principles approach to understanding the internet's router-level topology. *ACM SIGCOMM Comput. Commun. Rev.* **34**(4), 3–14 (2004)
21. Chiochetti, R., Perino, D., Carofiglio, G., Rossi, D., Rossini, G.: ACM. Inform: a dynamic interest forwarding mechanism for information centric networking, pp. 9–14 (2013)
22. Paul, S., Banerjee, B., Mukherjee, A., Naskar, M.K.: Priority-based content processing with Q-routing in information-centric networking (ICN). *Photonic Netw. Commun.* 1–11 (2016)
23. Chakrabarti, D., Faloutsos, C.: Graph mining: laws, tools, and case studies, *Synthesis Lectures on. Data Mining Knowl. Discovery* **7**(1), 1–207 (2012)
24. Newman, M.E., Watts, D.J., Strogatz, S.H.: Random graph models of social networks. *Proc. National Acad. Sci.* **99**(1), 2566–2572 (2002)
25. Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. *Physica A: Stat. Mech. Appl.* **311**(3), 290–614 (2002)
26. Dorogovtsev, S.N., Mendes, J.F.: Evolution of networks. *Adv. Phys.* **51**(4), 1079–1187 (2002)
27. Newman, M.E.: The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003)
28. Ethernet Jumbo Frames, <http://www.ethernetalliance.org/wp-content/uploads/2011/10/EA-Ethernet-Jumbo-Frames-v0-1.pdf>. Accessed 6 October 2016
29. Rummery, G.A., Niranjan, M.: On-line Q-learning Using Connectionist Systems. University of Cambridge, Department of Engineering, (1994)



ScaffoldNet: Detecting and Classifying Biomedical Polymer-Based Scaffolds via a Convolutional Neural Network

Darlington Ahiale Akogo¹(✉) and Xavier-Lewis Palmer^{1,2}

¹ MinoHealth AI Labs, 00233 SCC Inside, Weija, Accra, Ghana
darlington@gudra-studio.com

² Biomedical Engineering Institute, Old Dominion University,
Norfolk VA23529, USA

Abstract. We developed a Convolutional Neural Network model to identify and classify Airbrushed (alternatively known as Blow-spun), Electrospun and Steel Wire scaffolds. Our model ScaffoldNet is a 6-layer Convolutional Neural Network trained and tested on 3043 images of Airbrushed, Electrospun and Steel Wire scaffolds. The model takes in as input an imaged scaffold and then outputs the scaffold type (Airbrushed, Electrospun or Steel Wire) as predicted probabilities for the 3 classes. Our model scored a 99.44% Accuracy, demonstrating potential for adaptation to investigating and solving complex machine learning problems aimed at abstract spatial contexts, or in screening complex, biological, fibrous structures seen in cortical bone and fibrous shells.

Keywords: AI · Machine learning · Tissue engineering

1 Introduction

Convolutional Neural Networks have been an important aspect of Deep Learning in recent years. They were mainly responsible for the re-emergence and popularity of Neural Networks. The work of Alex Krizhevsky and Ilya Sutskever which won the ImageNet Large Scale Visual Recognition Competition in 2012 (ILSVRC-2012) was disruptive in the Artificial Intelligence, Machine Learning and Computer Vision community [1]. Since then Convolutional Neural Networks have been heavily applied to all sorts of problems, from various Object Detection and Image Segmentation problems and to specific domains like Medical Image Analysis [2–8].

Convolutional Neural Networks themselves aren't new, they were developed initially in the 1980s and were called Neocognitron [9–11]. They are broadly part of a wide set of models called Multi-Stage Hubel-Wiesel Architectures [12]. Hubel and Wiesel in the 1950s and 1960s identified orientation-selective simple cells with local receptive fields in the cat's primary visual cortex, whose role is similar to the Convolutional Neural Network's filter bank layers, and complex cells, whose role is similar to the pooling layers. The Neocognitron was the first model to simulate them on a computer. It used a layer-wise, unsupervised competitive learning algorithm for the filter banks, and a separately-trained supervised linear classifier for the output layer.

And then in 1989, LeNet-5 was introduced which simplified the architecture and used the Backpropagation algorithm to train the entire architecture in a supervised fashion [13]. The architecture was successful for tasks such as Optical Character Recognition and Handwriting Recognition.”

Neural Networks including Convolutional Neural Networks were unfortunately generally abandoned in the late 1980s. The reason why they were abandoned and why they re-emerged can both be attributed to Computational Power and Amount of Data. Deep Neural Networks require a lot of processing power to be effectively trained and they only perform well when trained on a lot of Data. Both were lacking then and have grown a lot in recent years. With lots of powerful GPUs and Big Data, Neural Networks finally could be applied to complex problems.

Convolutional Neural Networks are very effective for Computer Vision problems. After winning the ILSVRC-2012 Competition, Convolutional Neural Networks have been applied to multitudes of Computer Vision problems. Their effectiveness can be attributed to their ability to handle translation invariances in images by relying on shared weights and exploit spatial locality by enforcing a local connectivity pattern between neurons of adjacent layers. We chose them for this reason, knowing we wanted a model that could visually detect and differentiate between different types of scaffolds.

In this example, we utilized our convolutional neural network towards demonstrating that they can be used for distinguishing between different scanning electron microscope images of polymer-based scaffolds manufactured for research and application towards tissue regeneration. Three scaffold types were used. One was from an airbrushed set, and two others were from a training study, composed of an electrospun fiber set and a control set composed of steel wires [14]. We set out to test if we could develop a Convolutional Neural Network model that could identify and classify among the sets in hopes of producing a tool that would be useful in biomedical manufacturing, forensics, and perhaps more.

2 Problem Formation

Our problem was framed as a Classification problem, given a 128×128 pixels image of a scaffold, our model has to classify it as either Airbrushed, Electrospun or Steel Wire. Our model’s objective during training is to optimize the Cross-entropy loss:

$$-\sum_{c=1}^M y_{o,c} \log(p_{o,c})$$

where M—number of classes (3: Airbrushed, Electrospun and Steel Wire) log—the natural log—binary indicator (0 or 1) if class label c is the correct classification for observation o—predicted probability observation o is of class c.

3 Model Architecture: Convolutional Neural Network

We used a 6-layer Convolutional Neural Network trained on our 3043 scaffold images dataset. As shown in Fig. 1, our network ScaffoldNet starts with two 2-Dimensional Convolutional layers with a 3×3 kernel size and 32 output filters with the first as its input layer. Then followed by a single 2-Dimensional Convolutional layers also with a 3×3 kernel size and 64 output filters. The receptive fields of the Convolutional layers' filters (equivalently this is the filter size) captures various features across local regions in our input images, which is why they are powerful. During the forward propagation, our network slides (convolves) each filter across the width and height of the input images and computes dot products between the entries of the filter and the input at any position. As our network slides the filter over the width and height of the input images, it will produce a 2-dimensional activation map that gives the responses of that filter at every spatial position. During training, our model network will learn filters that activate when they see some type of visual element such as an edge of some orientation or a blotch of some color on the first layer, or eventually entire honeycomb or wheel-like patterns on higher layers of the network.

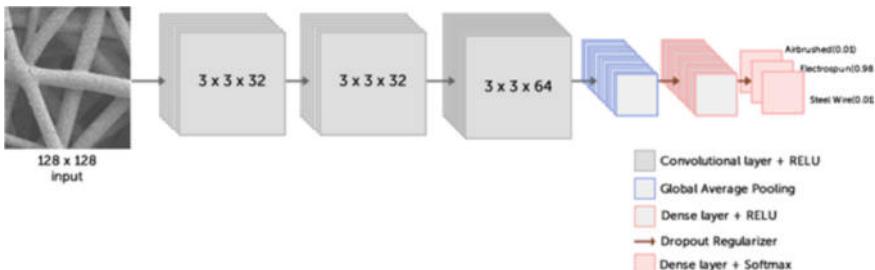


Fig. 1. The architecture of our Convolutional Neural Network: ScaffoldNet. This shows a graphical demonstration of ScaffoldNet at work, it takes a 128×128 grayscale image of an Electrospun scaffold in its input layer (the 1st $3 \times 3 \times 32$ Convolutional layer), propagates it through all its 4 hidden layers then finally outputs predicted probabilities for the 3 classes in its 3 unit output layer (the last Dense layer). It outputs the following probabilities for the 3 classes respectively; Airbrushed (0.01), Electrospun (0.98) and Steel Wire (0.01). ScaffoldNet accurately identified and classified the scaffold image as Electrospun by assigning it the highest probability among the classes (Electrospun: 0.98).

We then introduce a 2-Dimensional Global Average Pooling layer to reduce the spatial dimensions of our tensor [15]. Global Average Pooling performs Dimensionality Reduction to minimize overfitting by turning a tensor with dimensions $h \times w \times d$ into $1 \times 1 \times d$ which is achieved by reducing each $h \times w$ feature map to a single number simply by taking the average of all hw values. To further prevent overfitting, we then add a Dropout Regularizer with a fraction rate of 0.5 which is a Regularization technique that prevent Neural Networks developing complex coadaptation on the training data [16]. We then introduce a 32 unit densely-connected Neural Network

layer into our network architecture, followed by another Dropout Regularizer with a 0.5 fraction rate. Our final output layer is 3 unit densely-connected Neural Network layer.

All Convolutional and Densely-connected layers except the output layer use the Rectified Linear Unit (RELU) activation function: $f(x) = \max(0, x)$ where x is the input to a neuron [17]. RELUs are currently the most popular and successful activation function because they allow Deep Neural Networks to be more easily optimized than sigmoid and tanh activation functions due to the fact that gradients are able to flow when the input to the ReLU function is positive [18]. Our final densely connected output layer uses a Softmax activation function:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}}$$

where z is a vector of the inputs to the output layer (we have 3 output units, so there are 3 elements in z). And again, j indexes the output units.

Softmax function squashes the raw class scores into normalized positive values, then outputs them as separate probabilities for each of our classes (Airbrushed, Electrospun and Steel Wire), where all the probabilities add up to 1.

ScaffoldNet is trained end-to-end with Adam optimization algorithm using the standard parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) [19]. Adam is an algorithm for first-order gradient-based optimization of stochastic objective functions which computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. We train our model using mini batches of 32. We use a learning rate (α) of 0.001 and pick the model with the lowest validation loss.

4 Data

The dataset used for training, validating and testing our model is a collection of 3043 grayscale images of Airbrushed, Electrospun and Steel Wire scaffolds. We can see sample images from the 3 classes in Fig. 2. The dataset contains 1013 Airbrushed scaffold images, 960 Electrospun scaffold images and 1070 Steel Wire scaffold images as shown in Fig. 3. The airbrushed dataset is derived from scanning electron microscope images of airbrushed scaffolds under multiple polymer admixture settings taken at 200x magnification via a JEOL 1990 scanning electron microscope. Details regarding the electrospun and steel wire set can be found in Hotaling et al. 2015 [14]. In total, the three classes are Airbrushed, Electrospun, and Steel Wire. Each of the sets are SEM images taken at 200x magnification, giving them suitable plane for comparison.

The images of each class differ largely with respect to two factors, porosity and fiber diameter. In terms of appearance, the airbrushed class is quite easy to distinguish from the electrospun and steel wire class, but this ease disappears when comparing between the electrospun and steel wire classes without prior training. A useful AI with regards to manufacturing and forensic utilities would need to be able to differentiate between scaffold classes, ideally at extremely high rates for bulk processing. The impact of such a model is wide in terms of machine optimization and analysis in bulk

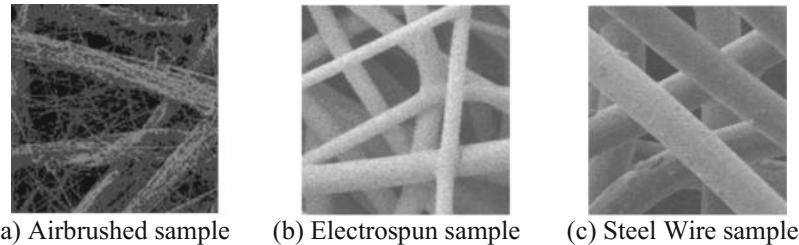


Fig. 2. Samples of the imaged Scaffolding materials belonging to the 3 classes used in training ScaffoldNet: Airbrushed, Electrospun and Steel Wire.

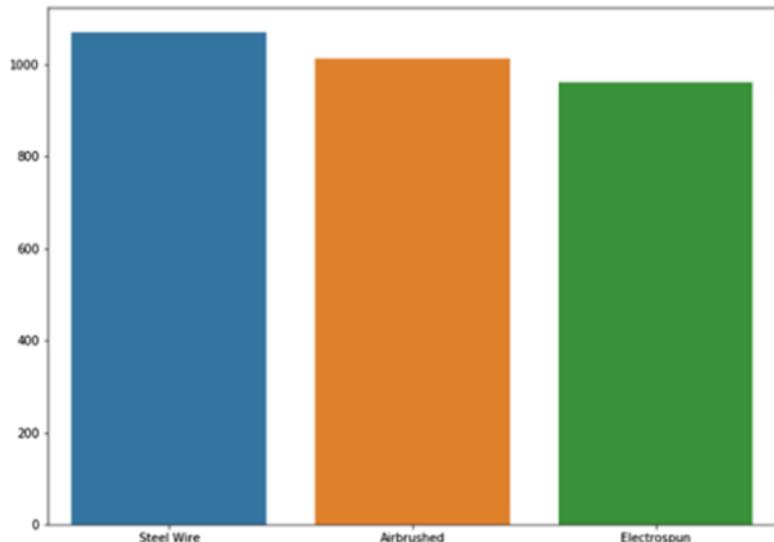


Fig. 3. The dataset contains 1070 Steel Wire scaffold images, 1013 Airbrushed scaffold images and 960 Electrospun scaffold images.

studies. As per Hotaling et al. 2015 and 2016, processes capable of batch analysis can vastly improve workflow and the pace of nanofiber scaffold research [14, 20].

5 Data Preparation

The collected set of images was then prepared and divided into subsets to best suit Machine Learning. The dataset was split into Training set, Validation set, and Testing set with an 8.8: 1.2: 1.0 ratio (2376, 368, 301), respectively. The splits were stratified so each Scaffold image class still gets equally split with an 8.8: 1.2: 1.0 ratio between the Training set, Validation set and Testing set. This is to prevent under training on some classes compared to others. The dimensions of all the images were reshaped

to 128×128 pixels. The images were then transformed by Standardization, which entails setting each image sample's mean to 0 and dividing its pixel values with its Standard Deviation. We do this so our image pixel values that would act as inputs to our model would have a similar range in order to have more stable gradients during training. The images were then further augmented with random horizontal flips, 5° rotations, width shifts, height shifts and zooms. This is to increase the variation in our dataset in order to make our trained model generalize.

6 Model Training and Validation

Using the Training set (2376 images), our Convolutional Neural Network was trained with Adam optimization algorithm and Cross-Entropy loss function and the Accuracy Classification Score was used as metrics. The Accuracy Score formula is as follow;

$$\text{accuracy}(y, \hat{y}) = \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i)$$

where the \hat{y}_i is the predicted output for the i -th sample \hat{y}_i is the (correct) target output computed over $n_{samples}$.

ScaffoldNet was trained in 11 epochs and its hyperparameters tuned using the Validation set (368 images). After just the first epoch, our model had the following performance results on the Validation set:

Accuracy score: 64.44%

Cross-Entropy loss: 0.6376.

Our model performance was drastically improving as training continued. After the 11th epoch, our model's final performance results on the Validation set were:

Accuracy score: 100%

Cross-Entropy loss: 0.0100.

7 Model Testing and Results

After all 11 epochs of training with the Training set, validation and hyperparameter tuning with the Validation set, our model was finally evaluated on the Test set (368 images). With this final evaluation, we get a better picture on how well our model generalizes and performs on unseen data since the Training set is used during the model's training and the Validation set is used to observe the performance of the model and to tune its hyperparameters to improve such performance. The Test set is the final evaluation for a model and changes are not made to the model after the results.

ScaffoldNet's final performance results on the Test set were:

Accuracy score: 99.44%

Cross-Entropy loss: 0.0997.

From the results of ScaffoldNet's final evaluation, we can tell that our model generalizes well and doesn't overfit, the performance on the Validation set after the 11 epochs and hyperparameter tuning is consistent with the evaluation results on the Test set. Our Convolutional Neural Network consistently demonstrates extremely high accuracy performance beyond even our expectations.

To further evaluate ScaffoldNet's output quality, we use the Receiver Operating Characteristic (ROC) metric and its Area Under Curve (AUC) score. We use the ROC curve to plot our model's true positive rate on the Y axis, and false positive rate on the X axis. ROC curves are mainly for Binary Classification and since our Convolutional Neural Network is Multiclass Classifier, we extended the ROC curve by drawing the ROC curve per class. We also drew another ROC curve by considering each element of the label indicator matrix as a binary prediction (micro-averaging). And we also used macro-averaging, which gives equal weight to the classification of each label. The top left corner of the plot is the "ideal" point—a false positive rate of zero, and a true positive rate of one. The ideal point has an AUC score of 1.00, the closer a classifier's score is, the better the classifier.

Our Convolutional Neural Network classifier ScaffoldNet has near perfect AUC score as shown in the plot above in Fig. 4. Both Class 0 (Airbrushed) and Class 1 (Electrospun) have an ideal score of 1.00 and Class 2 (Steel Wire) has a 0.99 score with only a few false positive. The micro-average AUC score of all 3 ROC curves being 0.99 and macro-average AUC score being 1.00. The results from our ROC curve and AUC score are also consistent with our model's evaluation results on the Test set.

The state-of-the-art performance and results of ScaffoldNet across multiple metrics can be explained by the clear and strong visual dissimilarities across our 3 scaffold classes as can be seen in Fig. 2. With such visible visual differences between our classes, classification especially with a powerful algorithm like Convolutional Neural Network becomes an easy task. Our model easily found the right visual features in the form of weights to use in identifying and classifying the 3 scaffold types.

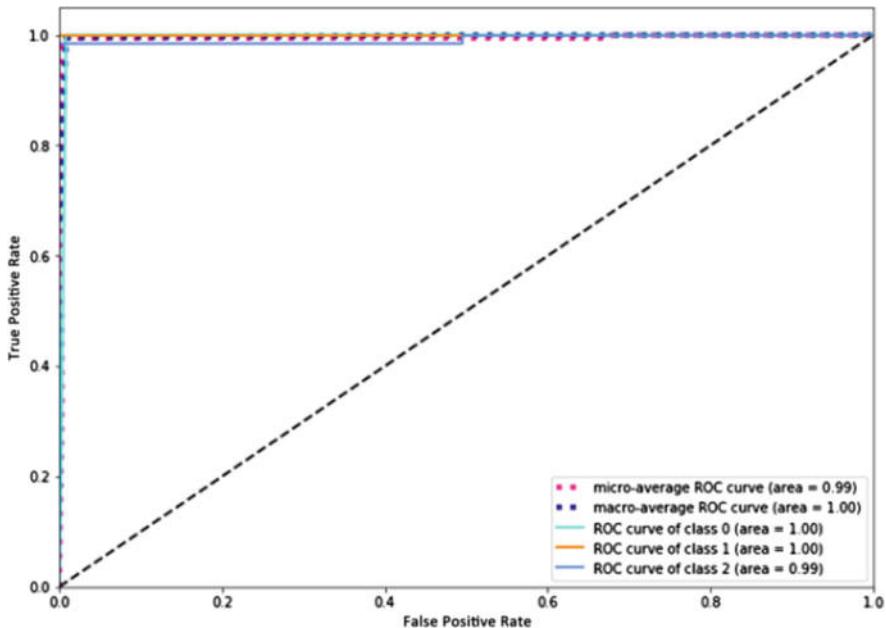


Fig. 4. Class 0, Class 1 and Class 2 are Airbrushed, Electrospun and Steel Wire classes respectively. As seen on the curve, our Classifier has an extremely optimal performance and score, a micro-average AUC score of 0.99, and a macro-average AUC score of 1.00 which is the highest possible AUC score.

8 Regarding Related Works

A close related work can be found in Chen et al. 2016 in which machine learning is used to identify cell shape phenotypes that have micro-environmental cues [21]. Our paper can be used to address a gap in that we are demonstrating a feasible focus on the extracellular matrix mimicking biocompatible polymer-based scaffolding, which would directly affect the microenvironment, which is found to be increasingly important in cell behavior and means of tissue engineering [22]. Our techniques and algorithms can be seen as an updated follow up in spirit, given the dated techniques and algorithms used in theirs, lacking true computer vision according to current standards. Our Deep Learning model is end to end, with the ability to learn to analyze and classify images on its own without need for excessive manual feature engineering. Their model requires Image Processing and Shape Quantification with Snake algorithm and Branching analysis, followed by “Filtering” Feature Selection and Heterogeneity Reduction before running it through a classic Machine Learning classifier such as a Support Vector Machine, whereas our Deep Neural Network model ScaffoldNet is trained to analyze and classify raw images on its and be able to improve with more training data. Our system uses a Convolutional Neural Network in contrast, which is state of the art in Computer Vision and Image/Object Detection.

9 Conclusion and Outlook

We developed a Convolutional Neural Network called ScaffoldNet that accurately classifies Airbrushed, Electrospun, and Steel Wire scaffolds after being trained on 2376 scaffold images, validated with 368 scaffold images and tested on 301 scaffold images. Our model demonstrates state of the art results including 99.44% Accuracy score, 0.99 micro-average AUC score and a 1.00 macro-average AUC score. We've shown the potentials of using Convolutional Neural Network, Deep Learning and Artificial Intelligence as Classification tools for differentiating between 3D Biomedical Polymer-based scaffolds. Our research is also meant to demonstrate the many possibilities and potentials of combining two of arguably the most revolutionary technologies of time, Artificial Intelligence and Biotechnology. We believe that plenty opportunities exist in which we can combine the two fields, mostly by using Artificial Intelligence to automate and optimize the many processes involved in Biotechnology, especially within automating identification and minimization of defects in manufacturing processes and forensics for material on the microscale. Specifically, we are interested in expanding the useful parameters by which the explored classes can be further distinguished and analyzed by our Convolutional Neural Network. Broadly, we intend to further explore many more of such ways in which we can combine Artificial Intelligence and Biotechnology at different measurement scales.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Systems* **25** (2012)
2. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.: Semantic image segmentation with deep convolutional nets and fully connected CRFs (2014)
3. Redmon, J., Divvala, S., Girshick, R., Farhadi A.: You only look once: unified, real-time object detection (2015)
4. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks (2015)
5. Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N.: AggNet: deep learning from crowds for mitosis detection in breast cancer histology images (2016)
6. Van Grinsven, M., Van Ginneken, B., Hoyng, C., Theelen, T., Sánchez, C.: Fast convolutional neural network training using selective data sampling: application to hemorrhage detection in color fundus images (2016)
7. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.: Suggestive annotation: a deep active learning framework for biomedical image segmentation (2017)
8. Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks (2017)
9. Fukushima, K.: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**(4): 193–202. <https://doi.org/10.1007/bf00344251> (1980)
10. Fukushima, K., Miyake, S., Ito, T.: Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Trans. Syst. Man, and Cybern. SMC* **13**(3), 826–834 (1983)

11. Fukushima, K.: A hierarchical neural network model for selective attention. In: Eckmiller, R., Von der Malsburg, C. (eds.) *Neural Computers*, pp. 81–90. Springer-Verlag (1987)
12. Hubel, D., Wiesel, T.: Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**(1), 215–243 (1968)
13. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Backpropagation applied to handwritten zip code recognition. *Neural Comput.* (1989)
14. Hotaling, N., Bharti, K., Kriel, H., Simon, C., Diameter, J.: A validated opensource nanofiber diameter measurement tool. *Biomaterials* **61**(August), 327–338 (2015)
15. Lin, M., Chen, Q., Yan, S.: Network in network (2013)
16. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting (2014)
17. Hahnloser, R., Sarpeshkar, R., Mahowald, M., Douglas, R., Seung, H.: Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* **405**, 947–951 (2000)
18. Ramachandran, P., Barret, Z., Quoc, L.: Searching for activation functions (2017)
19. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization (2014)
20. Hotaling, N., Jeon, J., Wade, M., Luong, D., Palmer, X-L., Bharti, K., Simon Jr, C.: Training to improve precision and accuracy in the measurement of fiber morphology. *PLOS One* **11**, e0167664 (2016)
21. Chen, D., Sarkar, S., Candia, J., Florczyk, S., Bodhak, S., Driscoll, M., Simon, C., Dunkers, J., Losert, W.: Machine learning based methodology to identify cell shape phenotypes associated with microenvironmental cues. *Biomaterials* **104**, 104–118 (2016)
22. Patrick, S., Mollica, P., Bruno, R.: Tissue specific microenvironments: a key tool for tissue engineering and regenerative medicine. *J. Biol. Eng.* **11**(1) (2017)



Transfer Learning for Cross-Domain Sequence Tagging Tasks

Meng Cao¹, Chaohe Zhang¹, Dancheng Li^{1(✉)},
Qingping Zheng², and Ling Luo²

¹ Northeastern University, Shenyang, China
ldc@mail.neu.edu.cn

² IBM China Development Lab, Beijing, China
lingluo@cn.ibm.com

Abstract. Neural network has been proved to be effective in sequence annotation task. Since it does not require task-specific knowledge, the same network structure can be easily applied to a wide range of applications. However, domain sequence tagging tasks still suffer from lack of available data. First, there is fewer available domain annotated data to train the recurrent neural network adequately. Second, the corpus maybe not available for domain-specific word embedding training. In this paper, we explore the problem of transfer learning of domain name entity recognition task. We proposed a modified skip-gram model for training cross-domain word embeddings, and we use source task with a large number of annotations (e.g. NER on CoNLL2003) to improve the performance on target task with fewer available annotations (e.g. NER on biomedical dataset). We evaluate our approach on a range of sequence tagging benchmarks, and the results show that significant improvement can be achieved using our approach.

Keywords: Sequence tagging · Transfer learning · Word embeddings

1 Introduction

Sequence tagging, such as part-of-speech (POS) tagging, text chunking and named entity recognition (NER), is an important problem in natural language processing. The task sequence tagging is to predict a linguistic tag for each word in a given context, which can be considered as a classification problem. Recently, lots of researches have shown that neural networks, specifically recurrent neural networks, achieve state-of-the-art performance on sequence tagging tasks. Sequence tagger based on neural network requires no task-specific feature engineering, which means one model structure can be applied to a wide range of sequence tagging tasks with little or no modification.

However, sequence tagging in specific domain will be affected by insufficient annotated data, such as biomedical corpora [6] and Twitter [15]. Consider the

task of named entity recognition (NER) on biomedical abstracts, several problems may limit the performance of the sequence tagging model. First, since there is only a small number of available annotations, the neural network may not be fully trained, especially for deep neural network when which may have lots of parameters.

Second, current neural models generally make use of word embeddings, which has shown great improvement over count-based models. However, they still have weaknesses and the most obvious problem arises when dealing with out-of-vocabulary (OOV) words. Generally, all OOV words in the dataset are represented using a generic vector representation, which will cause loss of information in the context. When the sequence tagging model is applied to a specific domain, the OOV problem will become more obvious. Besides, some domain-specific words, which that have been seen very infrequently in general corpus, they are likely have low quality due to lack of training data.

In order to address the above problems, we propose two approaches for the sequence tagging system to makes use of data of the source domain to improve the performance in the target domain. First, we apply transfer learning between different sequence tagging tasks, train the model to learn common feature representations of sequence tagging task. This is achieved by sharing specific hidden layers of the model between different tasks. Second, we propose a modified skip-gram model for learning cross-domain word embeddings, where the source domain information in the pre-trained embeddings is selectively incorporated for learning the target domain word embeddings in a principled manner.

Experimental results show that our approach can significantly improve the performance of the target task when the target task has few labels and is more related to the source task. Furthermore, we show that transfer learning can improve performance over state-of-the-art results even if the number of labels is relatively abundant.

2 Base Model for Sequence Tagging

In this section, we describe the basic neural network architecture we use for sequence tagging. This architecture is similar to the ones presented by Lample et al. [9].

2.1 Bidirectional LSTM for Sequence Tagging

Our sequence tagging model is based on bidirectional recurrent neural networks (RNNs) and conditional random fields (CRF). More specifically, we apply long short-term memory (LSTM) for networks implementation.

Figure 1a illustrate the general architecture of the sequence labeling network. The model receives a sequence of tokens (w_1, w_2, \dots, w_T) as input, which are mapped to a distributed vector space, resulting in a sequence of word embeddings (x_1, x_2, \dots, x_T) . Next, the word embeddings are given as input to two LSTM

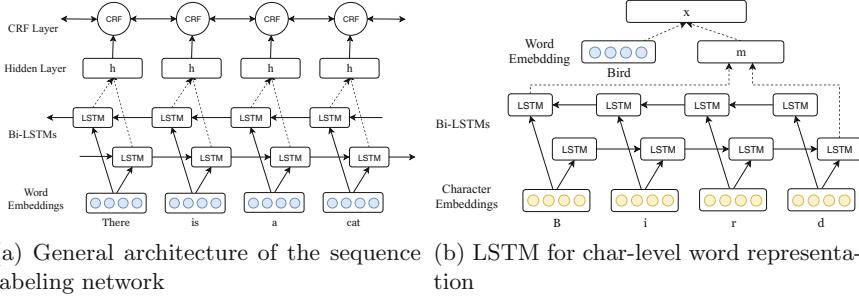


Fig. 1. The architecture of our base model for sequence tagging.

components moving in opposite directions through the sequence:

$$\vec{h}_t = \text{LSTM}(x_t, \vec{h}_{t-1}) \quad \overleftarrow{h}_t = \text{LSTM}(x_t, \overleftarrow{h}_{t+1})$$

At each word position, the hidden states of both forward and backward LSTM component are concatenated, resulting in a new feature representation that incorporates the contextual information. And we add an extra hidden layer on top of the LSTM, which allows the model to detect higher-level feature combinations [14]:

$$h_t = [\vec{h}_t : \overleftarrow{h}_t] \quad d_t = \tanh(W_d h_t)$$

where W_d is a weight matrix between layers.

Finally, we use a conditional random field (CRF) as the output layer, which conditions each prediction on the previous scores for the word having each of the possible labels. A separate weight matrix is used to learn transition probabilities between different labels, and the Viterbi algorithm is used to find an optimal sequence of weights. Given that y is a sequence of labels $[y_1, \dots, y_T]$, then the CRF score for this sequence can be calculated as:

$$s(y) = \sum_{t=1}^T A_{t,y_t} + \sum_{t=0}^T B_{y_t, y_{t+1}}$$

$$A_{t,y_t} = W_{o,y_t} d_t$$

where A_{t,y_t} shows how confident the network is that the label on the t -th word is y_t . $B_{y_t, y_{t+1}}$ shows the likelihood of transitioning from label y_t to label y_{t+1} , and these values are optimized during training. In order to optimise the CRF model, the loss function maximises the score for the correct label sequence, while minimising the score for all other sequences:

$$E = -s(y) + \log \sum_{\tilde{y} \in \tilde{Y}} e^{s(\tilde{y})}$$

where \tilde{Y} is the set of all possible label sequences.

2.2 Character-Level Word Representation

One problem of word-level model is that there are lots of words in the sequence which are out of vocabulary. Using a word-level model can only represents these words with a generic unknown word vector, which is shared between all other unseen words. One way to solve this is to combine word-level model with character-level word representations. Each word in the input sequence is split into individual characters, which are mapped to a sequence of character embeddings (c_1, c_2, \dots, c_R) using a randomly initialized character look up table. The character embeddings are given in direct and reverse order to a bidirectional LSTM, which outputs the representation that encodes the morphological information of the word:

$$\overrightarrow{h}_i^* = \text{LSTM}(c_i, \overrightarrow{h}_{i-1}^*) \quad \overleftarrow{h}_i^* = \text{LSTM}(c_i, \overleftarrow{h}_{i-1}^*)$$

We then use the last hidden vectors from each of the LSTM components, concatenate them together, and pass the result through a separate no-linear layer.

$$h^* = [\overrightarrow{h}_R^*; \overleftarrow{h}_l^*] \quad m = \tanh(W_m h^*)$$

where W_m is a weight matrix mapping the concatenated hidden vectors from both LSTMs into a joint word representation m , built from individual characters.

We now have two feature representations for each word, one character-based word representation and one pre-trained word representation learned from unannotated corpora. Following Lample et al. [9], one possible approach is to concatenate the two vectors and use this as the new word-level representation for the sequence labeling model:

$$x = [\tilde{x}; m]$$

3 Our Approach

In this section, we introduce our transfer learning architecture for sequence tagging. We first discuss our transfer learning architecture and parameter sharing schemes, then we introduce how we transfer pre-trained embeddings into the target domain.

3.1 Transfer Learning Architecture

Sequence tagging has a couple of applications including POS tagging, chunking, and named entity recognition. Similar to the motivation in, it is usually desirable to exploit the underlying similarities and regularities of different applications, and improve the performance of one application via joint training with another. Moreover, transfer between multiple applications can be helpful when the labels are limited.

Sequence tagging in different domains have different tags. For example, in the CoNLL2003 task, the entities are LOC, PER, ORG and MISC for locations,

persons, organizations and miscellaneous. The no-entity tag is O. But things are not the same when refer to a specific target domain. In the BC2GM benchmark, the entities are just NE and O. NE for the biological entities and O for others. Because of the difference in tagging through different domains, we cannot simply share all the parameters between the models. Thus, the tagging layers need to be trained by target corpus in order to make the model have the correct format while tagging on target domain and the parameters in the other parts of the model can be shared.

Word and char embedding, the word- and char- level RNN layer are shared between tasks. Since label sets in two domains are different, the source and target task have separate CRF layers.

3.2 Word Embeddings Transfer

Word embeddings were shown effective in many NLP tasks such as named entity recognition [17], sentiment analysis and syntactic parsing. A common practice is to initialize the model with publicly available pre-trained word embeddings. However, these word embeddings are often trained on corpus like Wikipedia, Google News or data from web crawling, which contain only a few domain-specific knowledge and words. It is also feasible to train the word embeddings on the target domain corpus from scratch, but the target corpus is usually small which restricts the quality of the word vectors. In order to address this problem, we propose a simple yet effective method for word embeddings transfer learning.

Let us first state the objective function of the skip-gram model [12] as follows:

$$\mathcal{L}_{\mathcal{D}} = \log \sigma(w \cdot c) + \sum_{i=1}^k \mathbb{E}_{c'_i \sim P(w)} [\log \sigma(-w \cdot c'_i)]$$

where \mathcal{D} is the corpus from which we learn the word embeddings. w refers to the current word, c is the context word, and the function $\sigma(\cdot)$ is the sigmoid function. The word c'_i is a “negative sample” sampled from the distribution $P(w)$, which is typically chosen as the unigram distribution $U(w)$ raised to the 3/4rd power [11].

Instead of training words embeddings directly on the target domain corpus, we first get pre-trained embeddings \mathbf{w}_s for word w , and add a term on the loss function that optimizes the embeddings to be similar to pre-trained embeddings:

$$\mathcal{L}'_{\mathcal{D}_t} = \mathcal{L}_{\mathcal{D}} + \sum_{w \in \mathcal{D}_t \cap \mathcal{D}_s} r_w \cdot (1 - \cos(\mathbf{w}_t, \mathbf{w}_s))$$

where \mathcal{D}_t refers to the target domain text corpus, \mathcal{D}_s refers to the corpus (vocabulary) of pre-trained embeddings, and \mathbf{w}_t is the target domain representation for w . The new loss function will maximize the cosine similarity between \mathbf{w}_t and \mathbf{w}_s , and r_w is used to control the degree of “knowledge transfer” from pre-trained embeddings to target domain embeddings.

It is worth noting that this is done only for words appear in both target domain corpus and pre-trained embeddings, ignoring words that only appear in either the domains.

4 Network Training

4.1 Parameter Initialization

For word embeddings training, we consider the most frequent 100,000 words in the corpus, and all word embeddings size is set to 100. Gradient descent algorithm is used in training with learning rate 0.1. The window size is set to 4, which means four words before and after the center word are considered as context words. And we randomly select four words from eight context words in training. The word embeddings are trained for 10 epochs on the whole corpus.

For the sequence tagging model, the word- and char- level LSTM layer size was set to 100 and 50 respectively in both forward and backward directions. Model is optimized using Adam algorithm with initial learning rate as 0.003, and we set the learn rate decay as 0.95. The batch size is set to 20 unless otherwise specified.

We applied dropout training which is proved to be useful for improving the performance of the model [9]. The dropout rate was set to 0.5. Performance was measured on the development set at every epoch and training was stopped if performance had not improved for 5 epochs.

4.2 Training

After initialization, we shall begin training the model. The training procedure can be described as follows. At each iteration, we sample a task (i.e., either s or t) from s, t in turn. Once given the sampled task, we select a batch of training set from the given task, and then perform an Adam update according to the loss function of the given task. The shared parameters and the task specific parameters are trained simultaneously. The above iterations are repeated until stopping. We adopt dropout and learning rate decay to avoid overfitting. Since the source and target tasks might have different convergence rates, we do early stopping if the target task performance does not improve after 5 epochs.

5 Experiment

5.1 Datasets

For training word embeddings, we use Enwik9¹ as the source-domain dataset, which contains the first 10^9 bytes of the English Wikipedia dump on Mar.3 2006. And we download publication abstracts from PubMed² as the dataset to train our target domain word embeddings.

For NER and POS tagging task, following benchmark datasets are used in our experiment:

¹ <https://cs.fit.edu/~mmahoney/compression/textdata.html>.

² <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>.

- **CoNLL00.** CoNLL-2000 dataset consists of the same partitions of the Wall Street Journal corpus (WSJ) as the widely used data for noun phrase chunking: sections 15–18 as training data (211727 tokens) and section 20 as test data (47377 tokens).
- **CoNLL03.** The CoNLL-03 corpus was created for the shared task on language independent NER. We use the English section of the dataset, containing news stories from the Reuters Corpus.
- **BC2GM.** The BioCreative II Gene Mention corpus [18] consists of 20,000 sentences from biomedical publication abstracts and is annotated for mentions of the names of genes, proteins and related entities using a single NE class.
- **CHEMDNER.** The BioCreative IV Chemical and Drug [8] NER corpus consists of 10,000 abstracts annotated for mentions of chemical and drug names using a single class. We make use of the official splits provided by the shared task organizers.
- **JNLPBA.** The JNLPBA corpus [7] consists of 2404 biomedical abstracts and is annotated for mentions of five entity types: CELL LINE, CELL TYPE, DNA, RNA, and PROTEIN. The corpus was derived from GENIA corpus entity annotations for use in the shared task organized in conjunction with the BioNLP 2004 workshop.

5.2 Performance

Cross-Domain Word Embeddings Evaluation We first evaluate our cross-domain embeddings on 6 datasets for NER task, the char-level LSTM-CRF model mentioned in Sect. 2 is used. Enwik9 and PubMed are used as source and target domain corpus respectively for cross-domain word embeddings training. We consider the following baseline methods when assessing the effectiveness of our approach:

- **No pretrain:** no pretrained word embeddings are used, randomly initialize the look up table in NER model.
- **Enwik9 corpus:** the word embeddings are trained for 10 epochs on Enwik9 corpus using skip-gram algorithm.
- **PubMed corpus:** the word embeddings are trained on PubMed corpus using skip-gram algorithm, same epochs as above.
- **Concat:** we simply concatenate the learned embeddings from both Enwik9 and PubMed domains.

Table 1 presents the results of F1 scores on different embeddings. We use the same hyperparameters on all datasets. As can be seen, randomly initializing look up table has the worst performance on every benchmark, which on the other hand indicates that the quality of the word embeddings has a great impact on the model performance. On CoNLL00 and CoNLL03 benchmark, as we expected, word embeddings trained from Enwik9 corpus has the best performance. Since CoNLL00 and CoNLL03 contains news stories from the Reuters Corpus, word embeddings trained from Enwik9 can express the more accurate word meaning compared with other word embeddings.

Table 1. Comparison F1 scores of NER using different pretrained word embeddings

	CoNLL00	CoNLL03	BC2GM	CHEMDNER	JNLPBA
No pretrain	93.30	87.48	71.28	78.65	73.49
Enwik9 corpus	94.20	89.19	71.93	83.13	74.44
PubMed corpus	94.05	88.74	73.38	84.13	74.91
Cross domain ($r = 0.1$)	93.84	88.74	75.88	83.22	75.19
Cross domain ($r = 1.0$)	93.22	89.14	75.48	83.82	75.05

On BC2GM and JNLPBA benchmark, we can observe word embeddings learned using our algorithm can lead to improved performance. We note that both BC2GM and JNLPBA consist of sentences in biomedical field, which contain many domain-specific term, and it is important for the model to learning good representation of these words. Notice that cross-domain word embeddings also performs better than word embeddings trained on target domain corpus only, we believe that is because the meaning of general words can be more accurately extracted using our algorithm.

Transfer Learning Performance Results for transfer learning on different benchmarks are shown in Tables 2, 3 and 4. In all three benchmarks, we can see that transfer learning can clearly improve the performance of the model. Take BC2GM benchmark for example, using embeddings trained on Enwik9 corpus, the F1 score of NER task is 71.93% using basic LSTM-CRF model. When the regularization (CoNLL03 as source) is used, this value is raised to 75.97%. Same results can be observed on JNLPBA and CHEMDNER dataset. The best performance is obtained when applying both layer-level transfer and word embedding transfer. On BC2GM dataset, the best F1 scores is 77.90 with CoNLL03 as source task.

Table 2. NER on BC2GM dataset using different model configurations

Source task	Word embeddings	Transfer	F1
–	Enwik9	No	71.93
–	PubMed	No	73.38
CoNLL03	Enwik9	Yes	75.97
CoNLL03	PubMed	Yes	76.95
CoNLL03	Cross domain ($r = d0.1$)	Yes	77.90
CoNLL00	Enwik9	Yes	75.86
CoNLL00	PubMed	Yes	77.18
CoNLL00	Cross domain ($r = 0.1$)	Yes	77.56

Table 3. NER on JNLPBA dataset using different model configurations

Source task	Word embeddings	Transfer	F1
–	Enwik9	No	74.44
–	PubMed	No	74.71
CoNLL03	Enwik9	Yes	74.75
CoNLL03	PubMed	Yes	75.27
CoNLL03	Cross domain ($r = 0.01$)	Yes	75.90
CoNLL00	Enwik9	Yes	74.25
CoNLL00	PubMed	Yes	74.88
CoNLL00	Cross domain ($r = 0.01$)	Yes	76.11

Table 4. NER on CHEMDNER dataset using different model configurations

Source task	Word embeddings	Transfer	F1
–	Enwik9	No	83.13
–	PubMed	No	84.13
CoNLL03	Enwik9	Yes	85.26
CoNLL03	PubMed	Yes	85.40
CoNLL03	Cross domain	Yes	86.22
CoNLL00	Enwik9	Yes	83.98
CoNLL00	PubMed	Yes	83.73
CoNLL00	Cross domain ($r = 0.1$)	Yes	84.73

Other than use CoNLL03 dataset, we also conduct transfer learning using CoNLL00 as source task. Note that an important difference is that CoNLL00 is a benchmark used for the task of chunking. The experimental results show that CoNLL00 can still improve the model performance on target task, which proves that our model has the ability to share general feature representations cross different applications.

6 Related Work

Researches on sequence tagging starts in the 1990s, Rau [13] recognizes names of groups in the text in 1991, the study is considered the first sequence tagging project. Studies on sequence tagging has attract concern of many researchers since the MUC-6 in 1999, show rapidly development. In recent years, there are still a lot of researches for it across the world.

However, tagging on single domain cannot satisfy modern needs. There are a lot of domains lacking of sufficient corpora, such as biological, medical domain. Thus, how to increase the performance on those domain with scarce data tend

to be a new trend of sequence tagging. The most common method is transfer learning.

Normally, there are two kinds, transfer between resource and transfer based on models, of transfer learning for NER tasks. Transfer between resource such as cross-lingual dictionaries [20] and word alignments [19] utilizes extra linguistic data as weak supervision. Such methods show great success in related tasks, but they are sensitive to the quality of the additional resources. However, model-based transfer needs less data, it utilizes the similarity and correlation between the tasks in source domain and target domain. Such method trains the model on one domain and then modifies its parameters or architectures to fit for target task. For instance, Ando and Zhang [1] proposed a transfer learning framework that shares structural parameters across multiple tasks, and improve the performance on various tasks including NER; Collobert et al. [3] presented a task-independent convolutional neural network and employed joint training to transfer knowledge from NER and POS tagging to chunking. Cross-domain transfer, or domain adaptation, is also a well-studied branch of model-based transfer in NLP. Techniques in cross-domain transfer include the design of robust feature representations [16], hierarchical Bayesian prior [4], and so on.

For the fusion of word embeddings, the main step is combine the word embeddings of two different domains together in order to increase the adaptability of the new embedding. Glorot et al. [5] used domain adaptation for sentiment classification and McClosky et al. [10] built a domain adaptation model for parsing. However, the contribution in cross-domain word embedding learning is so rare that there is little knowledge to refer to, which may be a major reason to prevent people from using text corpora from different domains for word embedding learning. Also, that causes a confusion on which kind of information is worth learning from the source domain(s) for the target domain. To overcome the challenge, some pioneering work has stressed on this problem. Bollegala et al. [2] considered the frequently-shown and common words in both source domain and the target domain as the “pivots”. Then it tried to use the pivots to optimize the “non-pivots” around them, meanwhile ensuring the pivots to have the similar embedding across two domains (reduce the distance). Embeddings learned from such an approach were shown to be able to improve the performance on a multi-domain sentiment classification task.

7 Conclusion

In this paper, we explore the transfer learning approach for sequence tagging task from two aspects. First, we presented a simple yet effective model for word embeddings transfer. Given a target domain and word embeddings pretrained source domain, our model can learn embeddings incorporate knowledge from both source and target domains. Second, we conduct transfer learning between different tasks by sharing the hidden layers of neural network. Experiment results show that both transfer learning approaches can improve the performance of sequence tagging tasks on various datasets.

Acknowledgements. We would like to thank Prof. Li from Northeast University, China and Dr. Zheng from IBM Innovation Lab, without whose help, our work could not be finished so smoothly. We also thank all the reviewers for their useful feedback to the earlier draft of this paper and the anonymous reviewers for their constructive comments to revise the paper.

References

1. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* **6**(Nov), 1817–1853 (2005)
2. Bollegala, D., Maehara, T., Kawarabayashi, K.i.: Unsupervised cross-domain word representation learning. arXiv preprint [arXiv:1505.07184](https://arxiv.org/abs/1505.07184) (2015)
3. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**(Aug), 2493–2537 (2011)
4. Finkel, J.R., Manning, C.D.: Hierarchical bayesian domain adaptation. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 602–610. Association for Computational Linguistics (2009)
5. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 513–520 (2011)
6. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J.: Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics* **19**(suppl_1), i180–i182 (2003)
7. Kim, J.D., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the bio-entity recognition task at jnlpba. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. pp. 70–75. Association for Computational Linguistics (2004)
8. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: Chemdner: the drugs and chemical names extraction challenge. *J. Cheminformatics* **7**(1), S1 (2015)
9. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint [arXiv:1603.01360](https://arxiv.org/abs/1603.01360) (2016)
10. McClosky, D., Charniak, E., Johnson, M.: Automatic domain adaptation for parsing. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 28–36. Association for Computational Linguistics (2010)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
13. Rau, L.F.: Extracting company names from text. In: Artificial Intelligence Applications, 1991. In: Seventh IEEE Conference on Proceedings, vol. 1, pp. 29–32. IEEE (1991)
14. Rei, M., Crichton, G.K., Pyysalo, S.: Attending to characters in neural sequence labeling models. [arXiv:1611.04361](https://arxiv.org/abs/1611.04361) (2016)
15. Ritter, A., Clark, S., Etzioni, O., et al.: Named entity recognition in tweets: an experimental study. In: Proceedings of the Conference on Empirical Methods in

- natural Language Processing. pp. 1524–1534. Association for Computational Linguistics (2011)
- 16. Schnabel, T., Schütze, H.: Flors: Fast and simple domain adaptation for part-of-speech tagging. *Trans. Assoc. Comput. Linguist.* **2**, 15–26 (2014)
 - 17. Sienčnik, S.K.: Adapting word2vec to named entity recognition. In: Proceedings of the 20th nordic conference of computational linguistics, nodalida 2015, may 11–13, 2015, vilnius, lithuania. pp. 239–243. No. 109, Linköping University Electronic Press (2015)
 - 18. Smith, L., Tanabe, L.K., nee Ando, R.J., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., et al.: Overview of biocreative ii gene mention recognition. *Genome Biol.* **9**(2), S2 (2008)
 - 19. Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: Proceedings of the first international conference on Human language technology research. pp. 1–8. Association for Computational Linguistics (2001)
 - 20. Zirikly, A., Hagiwara, M.: Cross-lingual transfer of named entity recognizers without parallel corpora. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). vol. 2, pp. 390–396 (2015)



Ensemble Models for Enhancement of an Arabic Speech Emotion Recognition System

Rached Zantout¹, Samira Klaylat^{2(✉)}, Lama Hamandi³, and Ziad Osman²

¹ Rafik Hariri University, Almechref, Lebanon
zantoutrn@rhu.edu.lb

² Beirut Arab University, Beirut, Lebanon
samiraklaylat@gmail.com, zosman@bau.edu.lb
³ American University of Beirut, Beirut, Lebanon
lhl3@aub.edu.lb

Abstract. Ensemble classification model has been widely used in the area of machine learning to enhance the performance of single classifiers. In this paper, we study the effect of employing five ensemble models, namely Bagging, Adaboost, Logitboost, Random Subspace and Random Committee, on a vocal emotion recognition system. The system recognizes happy, angry, and surprise emotion from Arabic natural speech where the highest accuracy among single classifiers is obtained by SMO 95.52%. After applying the ensemble models on 19 single classifiers, the best enhanced accuracy is 95.95% achieved by SMO as well. The highest improvement in accuracy was 19.09%. It was achieved by the Boosting technique having the Naïve Bayes Multinomial as base classifier.

Keywords: Machine learning · Ensemble classifiers · Emotion recognition · Arabic speech · Natural corpus

1 Introduction

Emotion recognition from speech is a rapidly expanding field of research. This process can be summarized as a mapping between the emotion class and the audio feature class. Hence, it is simply a machine learning/classification problem and can be solved using machine learning methodologies. The aim of building a robust vocal emotion recognition system is to achieve high accuracy performance when classifying the audio data into emotion classes. Single classifiers are often employed by researchers to recognize discrete and continuous emotions from acted, spontaneous and induced speech corpora of different languages [1]. The performance of these systems varies and is limited by the type/amount of data, features extracted and classifiers applied.

To enhance the performance of classification models, a relatively new machine learning methodology has been adopted by researchers in different machine learning fields. This methodology is using ensemble algorithms that combine several classifiers called base learners in an attempt to improve the overall accuracy of the system [2]. The

base learners can be of same type (ensemble models) or different type (hybrid model). In this paper, we try to benefit from such models in order to increase the performance of a previously proposed system that recognizes happy, angry, and surprise emotions from an Arabic natural speech [3].

In [4], an Arabic natural corpus, referred to as ANAD is collected from live online Arabic TV shows, where each video is labeled as happy, angry, or surprised based on the emotions perceived by human listeners. A set of acoustic and spectral features, known as Low Level Descriptors (LLD) are then, extracted for every 1s speech unit.

Several classification methods were applied on the ANAD corpus in [3] to recognize happy, angry, and surprised emotions. The maximum accuracy performance was 95.52% achieved by the SMO classifier and the worst result was 53.58% obtained by the ZeroR classifier. Table 1 shows the performance of 19 classifiers used in [3], where 6 classifiers had an accuracy performance more than 91%, 8 classifiers had accuracy between 80 and 89.9%, 2 classifiers accomplished an accuracy performance between 79.9 and 73%, 2 classifiers had performance between 68 and 69% and the zeroR had the lowest performance of 53.58%.

Table 1. The original performance of the classifiers.

Classification model	Original performance (%)
SMO	95.52
Simple logistic	95.44
LMT	95.44
MultiClass classifier updateable	92.70
Classification via regression	92.62
KNN	91.11
PART	89.66
RepTree	89.37
Jrip	88.72
J48	88.65
Multi class classifier	84.53
Decision table	84.09
Random tree	82.36
Logistic	80.69
Decision stump	79.10
OneR	78.02
Hoeffding tree	69.99
Naïve bayes multinomial	68.18
zeroR	53.58

Since the accuracy performance of the classification models applied in [3] to recognize vocal emotions is still not close to 100%, we investigate boosting the classifiers by applying popular ensemble models. An ensemble model is a meta-algorithm that tries to eliminate uncorrelated errors and improve predictive performance by

constructing several identical classifiers, and averaging their results [5]. Popular ensemble methods, i.e. Bagging [6], Random Subspace [7], Random-Committee [8], as well as Logit-Boost and Adaptive Boosting are applied in this work. These ensembles are also called homogeneous ensembles because they use a single base learner to produce multiple models. Hence, every classifier listed in Table 1. is been used as a base learner for each of the mentioned ensemble models in an attempt to boost performance accuracy.

A review of related work in ensemble-classifiers speech emotion recognition systems is given in Sect. 2. In Sect. 3, dataset used in this work is described briefly. Section 4 defines the five ensemble models proposed and shows the corresponding results. Finally, summary and comparison of results is given in Sect. 5.

2 Related Work

In the area of emotion recognition, several studies have investigated the effect of applying ensemble techniques on classification models. In [9], Neural Networks, SVM, Random Forest and Naïve Bayes classifiers were combined together using the ROVER framework [10] to recognize neutral, emphatic, angry, and motherese emotions from the induced German corpus AIBO [11]. An enhancement of 5.8% with respect to the highest accuracy obtained among single classifiers. In [12], the Adaboost ensemble was performed on J48, Decision Table, and PART classifiers to recognize neutral, joy, aggressive, sensual and sad emotions from Spanish acted speech dataset.

SVM, multi-layer perceptron, K*, KNN, and Random Forest classifiers were combined in [13] using simple voting to improve the performance of recognizing angry and neutral emotions from the NATURAL [14] dataset. The highest accuracy performance was 76.93% obtained by the SVM model but increase up to 78.04% when performing simple voting on the 5 base learners.

Bagging and Boosting were executed on the C4.5 classifier in [15] to recognize neutral, anger, fear, disgust, sadness, joy, and surprised emotions. The performance of C4.5 was 77.38% before applying the ensemble models, and increased to 83.64 and 84.63% when performing Bagging and Boosting techniques respectively. Also, in [16] the C4.5 classifier was improved from 61.1% by the Adaboosting, Multiboosting and Bagging [6] ensembles to 72.3, 72.5, and 70.7% respectively. This study aimed to recognize anger, sadness, disgust, joy, fear, boredom and neutral emotions from the Berlin speech dataset Emo-DB [17].

The Random Subspace [7] and Random Committee [8] ensemble models have been widely used in many machine learning studies [18–23]. However, no study was found using these 2 ensembles in vocal emotion recognition systems. In this study, we investigate the application of these two ensemble models, as well as the bagging and boosting techniques on different classifiers proposed in [3] to recognize happy, angry and surprised emotions from spontaneous Arabic speech.

3 The ANAD Dataset

The Arabic Natural Audio Dataset (ANAD) used in this paper was constructed by Klaylat et al. [3] and is publically available at [4]. To build this dataset, eight videos were collected from online Arabic talk shows, where live calls were done between an anchor and a human from outside the studio. The videos were then labeled by 18 humans as happy, angry or surprised, depending on the emotions perceived.

After removing silence, laughs and noisy chunks were removed; every video was segmented into 1 s speech units, and hence a corpus composed of 1383 records with 505 happy, 137 surprised and 741 angry units was formed.

Next 25 acoustic low-level descriptors (LLDs) were extracted for each speech unit. These LLD features include: intensity, Zero crossing rate, MFCC 1–12, F0 and F0 envelope, Probability of voicing, and the LSP frequency 0–7. For every feature, 19 statistical functionals were applied, these functionals include: maximum, minimum, range, absolute position of maximum, absolute position of minimum, arithmetic of mean, Linear regression 1 and 2, Linear regressionA, Linear regressionQ, standard deviation, Kurtosis, Skewness Quartiles 1, 2, and 3, and finally, Inter-quartile ranges 1–2, 2–3, and 1–3. Finally, the delta coefficient for every LLD was computed and thus leading to a total of 950 features.

The Kruskal Wallis non-parametric test [24] was then applied to reduce the dimension of the data space. The test determines whether the medians of two or more groups are statistically different by comparing the p-value to a predefined significance level α . If the p-value is less than or equal to α , then the differences between the medians are statistically significant, otherwise they are not statistically significant. For the ANAD, the significance level α was 0.05 and hence the features with p-values less than 0.05 were removed and the final dataset of 1383 records and 845 features was formed.

4 Ensemble Models

Using multiple classifiers and averaging their results reduce the misclassification error and hence boost weak classifiers [5]. In this work, five popular homogeneous ensemble models are investigated to boost the accuracy performance of the Arabic vocal emotion recognition system proposed in [3]. Below, the definition and the corresponding results of applying the listed classifiers as base learners on every ensemble model are presented.

4.1 Bagging

Bagging or *bootstrap aggregation* [6] is an ensemble model used to decrease the variance of classifiers. This is achieved by generating more individual classifiers by resampling-with-replacement of the training set and hence, each training data subset is used to train a different base learner of the same type. Majority vote strategy is applied to combine base learners i.e. the class y that has been predicted the most by base learners is selected for an instance x in the training set. It is useful when there is a

limited amount of data and when the classifier is unstable. The classifier is considered unstable when a slight change in the training set would lead to a big change in the prediction accuracy [25]. Naïve Bayes and KNN are examples of stable classifiers while decision trees are considered one of the most unstable learning models. Figure 1 represents the Bagging ensemble algorithm when applied to single classifiers.

Input: S: learning set T: number of bootstrap samples BL: Base Learner
Output: C^* : multiple classifiers for i=1 to T begin S_i =bootstrap sample from S; C_i =BL(S_i); end; Classification: $C^*(x) = \max_y \sum_{i=1}^T C_i(x) = y$

Fig. 1. The Bagging algorithm.

When executing the Bagging model on the base learners presented in Table 1, nine out of nineteen classifiers showed improvement (see Table 2) where the highest enhancement was 8.89% achieved by Random Tree model.

Table 2. The performance of enhanced classifiers when applying bagging model.

Base learners	Original performance (%)	Bagging performance (%)	Enhancement (%)
Random tree	82.36	91.25	8.89
Decision table	84.09	88.94	4.84
J48	88.65	92.99	4.34
Jrip	88.72	92.99	4.27
PART	89.66	93.78	4.12
Rep tree	89.37	92.99	3.62
OneR	78.02	80.91	2.89
Classification via regression	92.62	93.78	1.16
SMO	95.52	95.81	0.29

Acknowledging what has been stated before, the stable learners KNN and HoeffdingTree are not improved by the Bagging ensemble. The KNN, HoeffdingTree, Naïve Bayes Multinomial classifiers showed a decline of 0.14, 0.07, and 0.07% respectively. The Simple Logistic, LMT, Decision Stump and ZeroR classifiers didn't improve and the Logistic, Multi-Class Classifier, and Multi-Class Classifier Updatable were very slow to show results when used with the Bagging model as base classifiers.

4.2 Adaptive Boosting

The Adaboost [26] model is another popular ensemble machine learning model where a series of identical classification models are iteratively constructed in an attempt to fix the prediction errors made by previous models. It uses the majority vote technique adjusted by the use of weights. Unlike the Bagging ensemble, Boosting sequentially creates different base learners by reweighting the instances in the training dataset. The first model is constructed by running the base classifier normally where all instances have equal weights. At every repetition, all instances in the training dataset are weighted based on the overall accuracy of the prior model where misclassified instances will get a larger weight in the next repetition. Adjusting the weights forces the classifier to focus on different instances and hence different classifiers are built and added until no further improvements are possible. Figure 2 clearly explains the steps of the Adaboost algorithm.

Adaptive Boosting works better on stable and simple classifiers for which the error doesn't increase quickly. When performing Adaboost using the classifiers in Table 1 as base models, eleven classifiers, shown in Table 3, showed improvement. The highest enhancement was 19.09% achieved by Naïve Bayes Multinomial model since the Naïve Bayes Multinomial model is a very stable model, unlike the KNN model which didn't show any improvement due to its instability and complexity.

The SMO and the Logistic classifiers, which are both instable, showed a decline of 0.51, 0.43, and 0.14% respectively. The MultiClass Classifier, MultiClass Classifier Updateable and Classification via Regression models were very slow.

4.3 Random Sub Space

The Random Subspace model [7] (RSM) constructs base models that learn from randomly selected feature subspaces. In Random Subspace, the training dataset is modified as in Bagging; however, this modification is performed in the feature space rather than in the instance space. In this ensemble, each classifier is constructed from the base learner on a randomly selected equal-sized subset of the feature/attribute set without replacement. The final result is obtained by simple majority voting among the constructed classifiers. The Random Subspace ensemble has been proved to work better when there exists a redundancy of values distributed among the feature set [7, 27]. Also, when the training set size is relatively small compared to the dimensionality of the feature set, the RSM is proved to improve the accuracy performance of base classifiers [28]. Figure 3 shows the algorithm followed by the Random Subspace model.

```

Input:
S: learning set
T: number of base learners
N: number of instances/classifier
BL: Base Learner
w: set of instances weight

Initialization:
for i=1 to N
begin
    w1 = [w1 ... wN], wi1 ∈ [0,1], Σi=1N wi1 = 1; wi1 = 1/N
end;

for j=1 to T
begin
    Sj = sample from S using distribution wj;
    Cj=BL(Sj);
    εj = , ∑i=1N wij lji;
    lji = 1 if Cj missclassified instance i, and 0 otherwise;
    If εj = 0 or εj ≥ 0.5
        Ignore Cj;
        for i=1 to N
        begin
            wij = 1/N;
        end
    else
        βj = εj / (1 - εj) where εj ∈ (0,0.5)
    Update individual weights:
    wij+1 = wij βj(1-lji) / ∑k=1N wkj βj(1-ljk) ; i = 1 ... N

Output:
C* and β1 .... βT ;

Classification:
For every class ct the support class is:
μt(x) = ∑Cj(x)=ct ln(1/βj)
C*(x) = max μt(x), t=1...number of classes

```

Fig. 2. The Adaboost algorithm.

When applying random subspace using the classifiers in Table 1 as base models, fourteen classifiers, shown in Table 4, showed improvement. The highest enhancement was 11.86% achieved by using the Logistic model.

Table 3. The performance of enhanced classifiers when applying adaptive boosting model

Base learners	Original performance (%)	Adaboost performance (%)	Enhancement (%)
Naïve bayes multinomial	68.18	87.27	19.09
HoeffdingTree	69.99	78.16	8.17
J48	88.65	94.94	6.29
Decision table	84.09	90.17	6.07
PART	89.66	95.23	5.57
Rep tree	89.37	94.72	5.35
Jrip	88.72	93.78	5.06
OneR	78.02	81.49	3.47
Decision stump	79.10	79.61	0.51
Random tree	82.36	82.50	0.14
Simple logistic	95.44	95.52	0.07

```

Input:
S: feature set
T: number of samples
r:dimension of feature subset
BL: Base Learner
Output:
C*: multiple classifiers

for i=1 to T
begin
    Si = randomly selected r features from the feature set S;
    Ci = BL(Si);
end;

Classification:
C*(x) = max_y Σ_{i=1}^T (C_i(x) = y)

```

Fig. 3. The Random Subspace algorithm.

The Hoeffding Tree and Decision Stump showed a drawback of 1.74 and 0.22%, respectively. The Simple Logistic and LMT models were very slow and the ZeroR didn't show any improvement.

4.4 Random Committee

The Random Committee [8] is a model used to build an ensemble of Randomizable based classifiers where each classifier is constructed by setting different seed using the training data and the final prediction is the average of predictions of the individual classifiers. This ensemble uses tree-base learners and hence only applicable on

Table 4. The performance of enhanced classifiers when applying random sub space model.

Base learners	Original performance (%)	Random Subspace performance (%)	Enhancement (%)
Logistics	80.69	92.55	11.86
Random tree	82.36	93.78	11.42
Multi class classifier	84.53	94.07	9.54
Decision table	84.09	89.95	5.86
J48	88.65	94.29	5.64
PART	89.66	95.23	5.57
Jrip	88.72	93.78	5.06
Rep tree	89.37	94.22	4.84
MultiClass classifier updateable	92.70	95.30	2.60
Classification via regression	92.62	95.08	2.46
KNN	91.11	92.84	1.74
Naïve bayes multinomial	68.1851	69.27	1.08
OneR	78.02	78.67	0.65
SMO	95.52	95.95	0.43

MultiClass Classifier Updateable, REPTree, and Random tree classifiers. Only Multi-Class Classifier Updateable showed a decline of 0.36% while an enhancement was achieved by all the others as depicted in Table 5.

Table 5. The performance of enhanced classifiers when applying random committee model.

Base learners	Original performance (%)	Random committee performance (%)	Enhancement (%)
Random tree	82.36	93.06	10.70
Rep tree	89.37	92.48	3.11

4.5 Logitboost

The Logit Boost [29] model is a boosting model for classifiers that handle weighted instances. It is based on the Adaboost [26] algorithm and performs classification via regression scheme as the base learner and hence is only applicable on base learners that can handle numeric classes. Therefore, it can be applied only on Decision Stump, the REPTree, and Random tree classifiers where the remaining models perform classification scheme (nominal classes) and not regression scheme (numeric classes). Table 6 shows the performance of these three classifiers has increased after applying the Logitboost ensemble.

Table 6. The performance of enhanced classifiers when applying Logitboost model.

Base learners	Original performance (%)	Logitboost performance (%)	Enhancement (%)
Decision stump	79.10	92.91	13.81
Random tree	82.36	92.70	10.34
Rep tree	89.37	93.42	4.05

The main difference between Adaboost and Logitboost is in computing the weight and the classification function. Figure 4 explains the algorithm for j classes applied by Logitboost on base learners.

```

Input:
T: number of classifier instances
w: set of instances weight
N: number of instances/classifier

Initialization:
wij = 1/N, i = 1, . . . , N, j= 1, . . . , T;
Cj(x) = 0;
pj(x) = 1/T ∀ j;

for m = 1 to M
begin

    for j=1 to T
    begin

        zij = (yij* - pj (xi)/(pj (xi)(1 - pj (xi)));
        wij=pj (xi)(1 - pj (xi));

        Fit the function fmj(x) by a weighted least-squares regression of zij to xi with weights
        wij

        end
        fmj(x) ← (T - 1)/T (fmj(x) - 1/J ∑k=1J fmk(x));
        Cj(x) ← Cj(x) + fmj(x) ;

        Update pj(x) = eFj(x)/∑k=1T eFk(x);
    end

    Classification:
    C*(x) =max Cj(x), j=1...T;

```

Fig. 4. The Logitboost algorithm.

5 Conclusion

The main purpose of this work is to evaluate the use of five ensembles machine learning models to increase the performance of the vocal emotion recognition system given in [3]. Below, results are discussed in two different perspectives, the strongest base learner that achieved the highest accuracy performance, and the best ensemble model that achieved highest improvement percentage. From the first perspective, the maximum general accuracy performance is 95.95% obtained by applying the Random Subspace ensemble on the SMO base learner as shown in Fig. 5. The SMO is an unstable classifier, that's why drawback of 0.51% is obtained when applying the Adaboost ensemble since the Boosting algorithm works better for stable base learners. The Logitboost is not applicable on SMO since it is not a numerical class learner, as well as the Random Committee since SMO is not a randomizable classifier and doesn't use seeds.

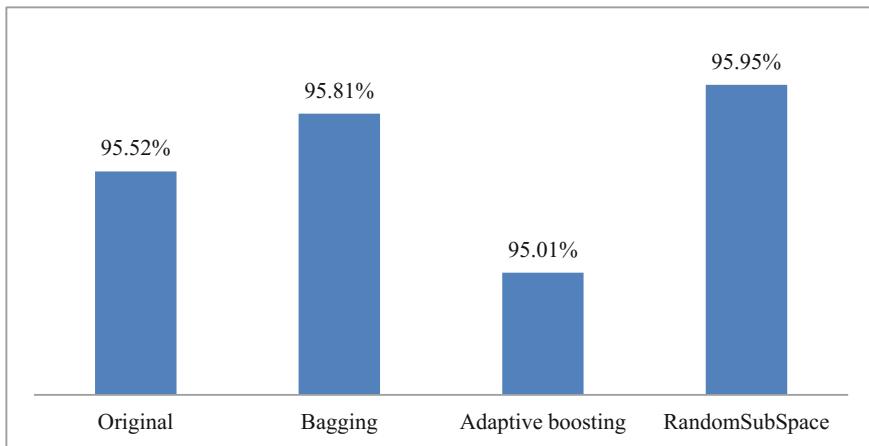


Fig. 5. Performance of applying ensemble models on SMO.

The Bagging ensemble model works well on strong learners [30] and hence improved the accuracy performance of the SMO classifier by 0.29%. However, the performance of Random Subspace outperformed the effect of Bagging and Boosting when applied to SMO. This is due to the fact that the training set (1383 instances) is relatively small with respect to the attribute set (845 features) [4]. Several studies have investigated and compared Bagging, Boosting and RSM ensemble techniques and showed that RSM usually outperforms Bagging and Boosting [27].

When discussing the enhancement percentages, Boosting technique achieved the highest improvement. Having the Naïve Bayes Multinomial as base classifier, the Adaboost model increased the performance from 68.11 to 87.27%, an improvement of 19.09%. The Naïve Bayes Multinomial classifier is a simple stable weak learner, and thus its performance is expected to increase by boosting and decrease by bagging

which is shown in Fig. 6. The Naïve Bayes Multinomial improved by 1.08% when used as base learner for the RandomSubspace.

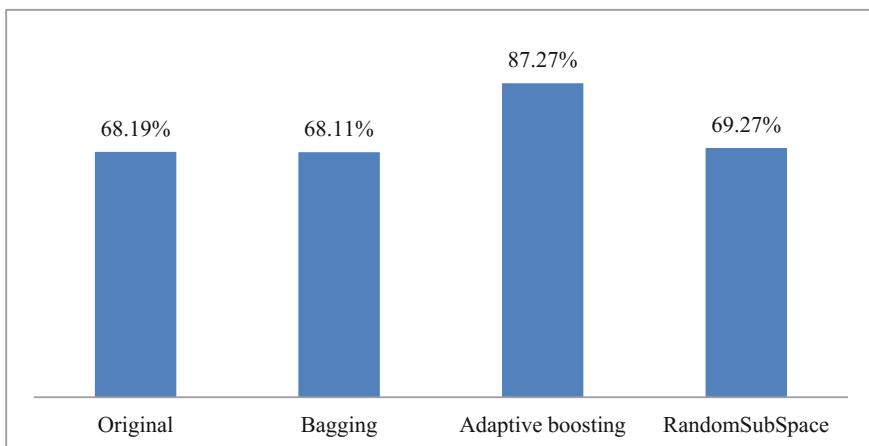


Fig. 6. Performance of applying ensemble models on Naïve Bayes Multinomial.

The results of this paper indicate that applying ensemble machine learning models can improve the accuracy performance of classifiers employed to recognize emotions from speech. However, the size of the dataset, the number of features and the emotions recognized play a significant role on the accuracy performance and improvement percentage. For future work, hybrid ensemble models that use base learners of different types can be investigated such as stackingC and voting. A larger corpus can also be used and results can be compared with the current system.

References

1. Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., Vogt, T., Aharonson, V., Amir, N.: The automatic recognition of emotions in speech. In: Petta, P., Pelachaud, C., Cowie, R. (eds.) Emotion-Oriented Systems, pp. 71–99. Springer, Berlin (2011)
2. Valentini, G., Masulli, F.: Ensembles of learning machines. In: Marinaro, M., Tagliaferri, R. (eds.) WIRN 2002. LNCS, vol. 2486, pp. 3–20. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45808-5_1
3. Klaylat, S., Osman, Z., Zantout, R., Hamandi, L.: Emotion Recognition in Arabic Speech. Analog. Integr. Circuits Signal Process., Springer, **96**(2), 337–351 (2018)
4. Klaylat, S., Hamandi, L., Zantout, R., Osman, Z.: Arabic natural audio dataset. MendeleyData, v1,<http://dx.doi.org/10.17632/xm232yxf7t.1>, Mendeley Data Website
5. Melville, P., Shah, N., Mihalkova, L., Mooney, R.J.: Experiments on ensembles with missing and noisy data. In: International Workshop on Multiple Classifier Systems, pp. 293–302. Springer, Heidelberg (2004)
6. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)

7. Kam Ho, T.: The Random Subspace Method for constructing Decision Forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(8), 832–844 (1998)
8. Frank, E., Hall, M.A., Witten, I.H.: The WEKA workbench. Online appendix for “data mining: practical machine learning tools and techniques”. Morgan Kaufmann, 4th edn. (2016)
9. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: Combining efforts for improving automatic classification of emotional user states. In: Proceedings of 5th Slovenian and 1st International Language Technologies Conference, IS LTC, pp. 240–245. Ljubljana, Slovenia (2006)
10. Fiscus, J.: A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In: proceedings of Automatic Speech Recognition and Understanding, ASRU, pp. 347–354. Santa Barbara, USA (1997)
11. Batliner, A., Hacker, C., Steidl, S., Nöth, E., D’Arcy, S., Russell, M.J., Wong, M.: You stupid tin box-children interacting with the AIBO robot: a cross-linguistic emotional speech corpus. In: Proceedings of 4th Language Resources and Evaluation Conference, LREC, pp. 171–174. Lisbon, Portugal (2004)
12. Iriondo, I., Planet, S., Socoró, J.C., Alías, F.: Objective and subjective evaluation of an expressive speech corpus. In: Proceedings of International Conference on Nonlinear Speech Processing, pp. 86–94. Springer, Berlin, Heidelberg (2007)
13. Morrison, D., De Silva, L.C.: Voting ensembles for spoken affect classification. *J. Netw. Comput. Appl.* **30**(4), 1356–1365 (2007)
14. Morrison, D., Wang, R., De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centres. *Speech Commun.* **49**(2), 98–112 (2007)
15. Schuller, B., Lang, M., Rigoll, G.: Robust acoustic speech emotion recognition by ensembles of classifiers. In: Tagungsband Fortschritte der Akustik-DAGA# 05. München (2005)
16. Schuller, B., Rigoll, G.: Timing levels in segment-based speech emotion recognition. In: Proceedings of International Conference on Spoken Language Processing ICSLP, Pittsburgh, USA (2006)
17. Burkhardt, F., Paeschke, A., Rolfs, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Proceedings of 9th European Conference on Speech Communication and Technology, ISCA, pp. 1517–1520. Lisbon, Portugal (2005)
18. Dastgheib, A., Ranjbar Pouya, O., Lithgow, B., Moussavi, Z.: Comparison of a new ad-hoc classification method with Support Vector Machine and Ensemble classifiers for the diagnosis of Meniere’s disease using EVestG signals. In: Proceedings of Electrical and Computer Engineering (CCECE), IEEE, pp. 1–4. Canada (2016)
19. Dacheng, T., Xiaouo, T., Xuelong, L., Xindong, W.: Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**(7), 1088–1099 (2006)
20. Nanni, L., Lumini, A.: Random subspace for an improved BioHashing for face authentication. *Pattern Recogn. Lett.* **29**(3), 295–300 (2008)
21. Wang, X., Tang, X.: Random sampling LDA for face recognition. In: proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2004)
22. Ali, R., Siddiqi, M.H., Idris, M., Kang, B.H., Lee, S.: Prediction of diabetes mellitus based on boosting ensemble modeling. In: proceedings of International conference on Ubiquitous Computing and Ambient Intelligence, pp. 25–28. Springer, Cham (2014)
23. Thongkam, J., Xu, G., Zhang, Y., Huang, F.: Support Vector Machine for Outlier Detection in Breast Cancer Survivability Prediction. In: Ishikawa, Yoshiharu, He, J., Xu, G., Shi, Y., Huang, G., Pang, C., Zhang, Q., Wang, G. (eds.) APWeb 2008. LNCS, vol. 4977, pp. 99–109. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89376-9_10

24. Leard Statistics website: Kruskal-Wallis H Test using SPSS Statistics. <https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php>
25. Kuncheva, L.: Combining Pattern Classifiers: Methods and Algorithms. Wiley (2004)
26. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: The proceedings of 13th International Conference on Machine Learning, pp. 148–156. San Francisco (1996)
27. Skurichina, M.: Stabilizing weak classifiers. Ph.D. thesis, Delft University of Technology, Delft, The Netherlands (2001)
28. Skurichina, M., Duin, R.P.W.: Bagging, boosting and the random subspace method for linear classifiers. Pattern Anal. Appl. **5**, 121–135 (2002)
29. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: A Statistical View of Boosting. Stanford University (1998)
30. Niculescu-Mizil, A., Caruana, R.: An empirical comparison of supervised learning algorithms using different performance metrics. Technical report, Cornell University (2005)



Automating Vehicles by Deep Reinforcement Learning Using Task Separation with Hill Climbing

Mogens Graf Plessen()

IMT, Lucca, Italy
mogens.plessen@imtlucca.it

Abstract. Within the context of autonomous driving a model-based reinforcement learning algorithm is proposed for the design of neural network-parameterized controllers. Classical model-based control methods, which include sampling- and lattice-based algorithms and model predictive control, suffer from the trade-off between model complexity and computational burden required for the online solution of expensive optimization or search problems at every short sampling time. To circumvent this trade-off, a 2-step procedure is motivated: first learning of a controller during offline training based on an arbitrarily complicated mathematical system model, before online fast feedforward evaluation of the trained controller. The contribution of this paper is the proposition of a simple gradient-free and model-based algorithm for deep reinforcement learning using task separation with hill climbing (TSHC). In particular, (i) simultaneous training on separate deterministic tasks with the purpose of encoding many motion primitives in a neural network, and (ii) the employment of maximally sparse rewards in combination with virtual velocity constraints (VVCs) in setpoint proximity are advocated.

Keywords: Motion planning · Encoding motion primitives in neural networks · Sparse rewards · Hill climbing · Virtual velocity constraints

1 Introduction

There exists a plethora of motion planning methods for self-driving vehicles [31]. The diversity is caused by a core difficulty: the trade-off between model complexity and permitted online computation at short sampling times. Three popular control classes and vision-based end-to-end solutions are briefly summarized.

1.1 Model-Based Control Methods

In [22] a *sampling-based* anytime algorithm RRT* is discussed. Key notion is to refine an initial suboptimal path while it is followed. As demonstrated, this is feasible when driving towards a static goal in a static environment. However, it may

This work was mostly done while MGP was with IMT School for Advanced Studies Lucca, Piazza S. Francesco 19, 55100 Lucca, Italy.

be problematic in dynamic environments requiring to constantly replan paths, and where an online sampled suitable trajectory may not be returned in time. Other problems of online sampling-based methods are a limited model complexity and their tendency to produce jagged controls that require a smoothing step, e.g., via conjugate gradient [9]. In [27], a *lattice-based* method is discussed. Such methods, and similarly also based on *motion primitives* [11, 17, 25, 41], are always limited by the size of the look-up table that can be searched in real-time. In [27], a GPU is used for search. In [10], *linear time-varying model predictive control* (LTV-MPC) is discussed for autonomous vehicles. While appealing for its ability to incorporate constraints, MPC must trade-off model-complexity vs. computational burden from solving optimization problems online. Furthermore, MPC is dependent on state and input reference trajectories, typically for linearization of dynamics, but almost always also for providing a tracking reference. Therefore, a two-layered approach is often applied, with motion planning and tracking as the 2 layers [31]. See [35] for a method using geometric corridor planning in the first layer for reference generation and for the combinatorial decision taking on which side to overtake obstacles. As indicated in [10, Sect. V-A] and further emphasized in [33], the selection of reference velocities can become problematic for time-based MPC and motivated to use spatial-based system modeling. Vehicle dynamics can be incorporated by inflating obstacles [17]. For tight maneuvering, a linearization approach [36] is more accurate, however, computationally more expensive. To summarize, 2 core observations are made. First, all methods (from sampling-based to MPC) are derived from vehicle *models*. Second, all of above methods suffer from the real-time requirement of short sampling times. As a consequence, all methods make simplifications on the employed model. These include, e.g., omitting of dynamical effects, tire dynamics, vehicle dimensions, using inflated obstacles, pruning search graphs, solving optimization problems iteratively, or offline precomputing trajectories.

1.2 Vision-Based Methods

In [37] a pioneering end-to-end trained neural network labeled ALVINN was designed for steering control of an autonomous vehicle. Video and range measurements are fed to a fully connected (FC)-network with a single hidden layer with 29 hidden units, and an output layer with 45 direction output units, i.e., discretized steering angles, plus one road intensity feedback unit. ALVINN does not control velocity and is trained using supervised learning based on road “snapshots”. Similarly, recent DAVE-2 [5] also only controls steering and is trained supervisedly. However, it outputs continuous steering action and is composed of a network including convolutional neural networks (CNN) as well as FC-layers with a total of 250000 parameters. During testing, steering commands are generated from only a front-facing camera. Another end-to-end system based on only camera vision is presented in [7]. First, a driving intention (change to left or right lane, stay in lane and break) is determined, before steering angle is output from a recurrent neural network (RNN). Instead of mapping images to steering control, in [6] and [50], affordance indicators (such as distance to cars in current

and adjacent lanes etc.) and feasible driving actions (such as straight, stop, left-turn, right-turn) are output from neural networks, respectively. See also [32] and their treatment of “option policies”. To summarize, it is distinguished between (i) vision-based end-to-end control, and (ii) perception-driven approaches that attempt to extract useful features from images. Note that such features (e.g., obstacle positions) are implicitly required for all methods from Sect. 1.1.

1.3 Motivation and Contribution

This work is motivated by the following additional considerations. As noted in [48], localization relative to lane boundaries is more important than with respect to GPS-coordinates, which underlines the importance of lasers, lidars and cameras for automated driving. Second, vehicles are man- and woman-made products for which there exist decade-long experience in vehicle dynamics modeling [15, 38]. There is no reason to a priori entirely discard this knowledge (for manufacturers it is present even in form of construction plans). This motivates to leverage available vehicle models for control design. Consider also the position paper [16] for general limitations of end-to-end learning. Third, a general purpose control setup is sought avoiding to switch between different vehicle models and algorithms for, e.g., highway driving and parking. There also exists only one real-world vehicle. In that perspective, a complex vehicle model encompassing all driving scenarios is in general preferable for control design. Also, a model mismatch on the planning and tracking layer can incur paths infeasible to track [17]. Fourth, the most accident causes involving other mobile vehicles are rear-end collisions [29], which most frequently are caused by inattentiveness or too close following distances. Control methods that enable minimal sampling times, such as feedforward control, can deterministically increase safety through minimal reaction times. In contrast, environment motion prediction (which can also increase safety) always remains stochastic. Fifth and to summarize, small sampling times may contradict to use complex vehicle models for control when employed within expensive online optimization or search problems. Therefore, a 2-step procedure is motivated instead: first learning of a controller during offline training based on an arbitrarily complicated mathematical system model, before online fast evaluation of the trained controller. In an automated vehicles settings, it implies that once trained, low-cost embedded hardware can be used online for evaluation of only few matrix vector multiplications.

The contribution of this paper is a simple gradient-free algorithm for model-based deep reinforcement learning using task separation with hill climbing (TSHC). Therefore, it is specifically proposed to (i) simultaneously train on separate deterministic tasks with the purpose of encoding motion primitives in a neural network, and (ii) during training to employ maximally sparse rewards in combinations with virtual velocity constraints (VVCs) in setpoint proximity. The problem formulation, the proposed training algorithm and numerical simulation experiments are discussed in Sects. 2–4, before concluding.

2 Problem Formulation and Preliminaries

2.1 General Setup and Objective

The problem is visualized in Fig. 1. Exteroceptive measurements are assumed to include inter-vehicular communication (car-2-car) sensings as well as the communication with a centralized or decentralized coordination service such that, in general, multi-automated vehicle coordination is also enabled [34]. For learning of controller C it is distinguished between 5 core aspects: the system model used for training, the neural network architecture used for function approximation, the training algorithm, the training tasks selection and the hardware/software implementation. Fundamental objective is to encode many desired motion primitives (training tasks) in a neural network. The main focus of this paper is on the training algorithm aspect, motivated within the context of motion planning for autonomous vehicles that are characterized by nonholonomic system models.

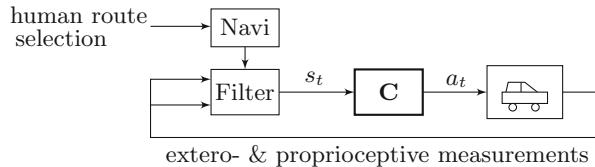


Fig. 1. Closed-loop control architecture. “Navi” and “Filter” (not focus of this paper) map human route selections as well as extero- and proprioceptive measurements to feature vector s_t . This paper proposes a simple gradient-free algorithm (TSHC) for learning of controller C, which maps s_t to control action a_t applied to the vehicle.

2.2 Illustrative System Model for Simulation Experiments

For simplicity for simulation experiments of Sect. 4 a simple Euler-discretized nonlinear kinematic bicycle model [38] is assumed. Equations of motion are $x_{t+1} = x_t + T_s v_t \cos(\psi_t)$, $y_{t+1} = y_t + T_s v_t \sin(\psi_t)$, $\psi_{t+1} = \psi_t + T_s \frac{v_t}{l_f} \tan(\delta_t)$, with 3 states (position-coordinates and heading), 2 controls (steering angle δ and velocity v), 1 system parameter (wheelbase $l_f = 3.5\text{m}$), and t indexing sampling time T_s . Coordinates x_t and y_t describe the center of gravity (CoG) in the inertial frame and ψ_t denotes the yaw angle relative to the inertial frame. Physical actuator absolute and rate constraints are treated as part of the vehicle model on which the network training is based on. Thus, the continuous control vector is defined as $a_t = [v_t, \delta_t]$, with $\max(v_{t-1} + \dot{v}_{\min,t} T_s, v_{\min,t}) \leq v_t \leq \min(v_{t-1} + \dot{v}_{\max,t} T_s, v_{\max,t})$, and $\max(\delta_{t-1} + \dot{\delta}_{\min} T_s, \delta_{\min}) \leq \delta_t \leq \min(\delta_{t-1} + \dot{\delta}_{\max} T_s, \delta_{\max})$. The minimum velocity is negative to allow for reverse driving.

2.3 Comments on Feature Vector Selection s_t

While the system model used for training prescribes a_t , this is not the case for feature vector s_t . In general, s_t may be an arbitrary function of filtered extero- and proprioceptive measurements according to Fig. 1. Thus, a plethora of many different sensors may be compressed through the filtering step to a low-dimensional s_t . Due to curse of dimensionality low-dimensional s_t are favorable, since the easiest way to generate training tasks is to grid over the elements of s_t . For our purpose, s_t must always relate the current vehicle state with reference to a goal state (e.g., via a difference operator). Ultimately, instead of only a single time-instant, s_t may, in general, also represent a collection of multiple past time measurements (time-series) leading up to time t .

2.4 Comments on Computation

For perspective, deep learning using neural networks as function approximators is in general computationally very demanding. To underline remarkable dimensions and computational efforts in practice, note that, for example, in [40] training is distributed on 80 machines and 1440 CPU cores. In [1], even more profoundly, 1024 Tesla P100 GPUs are used in parallel. For perspective, one Tesla P100 permits a double-precision performance of 4.7 TeraFLOPs [30].

3 Training Algorithm

3.1 Neural Network Controller Parametrization

The controller in Fig. 1 may be parameterized by any of, e.g., FCs, LSTM cells including peephole connections [14], GRUs [8] and variants.

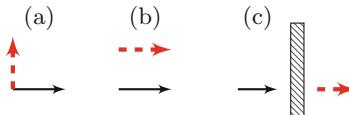


Fig. 2. The problematic of *rich rewards*. Three scenarios **a**, **b** and **c** indicate different start (black) and goal (red dashed) states (position and heading). For **c**, an obstacle is added. See Sect. 3.2 for discussion.

All neural network parameter weights to be learnt are initialized by Gaussian-distributed variables with zero mean and a small standard deviation (e.g., 0.001). Exceptions are adding a 1 to the LSTM's forget gate biases for LSTM cells, as recommended in [21], which are thus initialized with mean 1. In proposed setting, the affine part of all FC-layers is followed by nonlinear $\tanh(\cdot)$ activation functions acting elementwise. Because of their bounded outputs, saturating nonlinearities are preferred over ReLUs, which are used for the hidden layers in

other RL settings [24], but can result in large unbounded layer output changes. Before entering the neural network s_t is normalized elementwise (accounting for the typical range of feature vector elements). The final FC-layer comprises a $\tanh(\cdot)$ activation. It accordingly outputs bounded continuous values, which are then affinely scaled to a_t via physical actuator absolute and rate constraints valid at time t .

So far, continuous a_t was assumed. A remark with respect to gear selection is made. Electric vehicles, which appear suitable to curb urban pollution, do not require gearboxes. Nevertheless, in general a_t can be extended to include discrete gear as an additional decision variable. Suppose N_{gears} gears are available. Then, the output layer can be extended by N_{gears} channels, with each channel output representing a normalized probability of gear selection as a function of s_t , that can be trained by means of a softmax classifier.

3.2 Reward Shaping

Reward shaping is instrumental for learning by reinforcement signals [45]. However, general reward shaping was found to be a far from trivial matter in practical problems. Therefore, our preferred choice is motivated in detail. In most practical control problems, a state z_0 is given at current time $t = 0$, and a desired goal state z^{goal} is known. Not known, however, is the shape of the best trajectory (w.r.t. a given criterion) and the control signals that realize that trajectory. Thus, by nature these problems offer a *sparse* reward signal, $r_{\tilde{T}}(z^{\text{goal}})$, received only upon reaching the desired goal state at some time $\tilde{T} > 0$. In the following, alternative *rich reward* signals and *curriculum learning* [4] are discussed.

The Problematic of Designing Rich Reward Signals A reward signal $r_t(z_t, a_t, s_t)$, abbreviated by r_t , is labeled as *rich* when it is time-varying as a function of states, controls and feature vector. Note that the design of any such signal is heuristic and motivated by the hope for accelerated learning through maximally frequent feedback. In the following, the problematic of rich rewards is exposed. First, for our application let $e_{d,t} = \sqrt{(x_t - x^{\text{goal}})^2 + (y_t - y^{\text{goal}})^2}$, $e_{\psi,t} = |\psi_t - \psi^{\text{goal}}|$, and $e_{v,t} = |v_t - v^{\text{goal}}|$ relate states with desired goal states, and let a binary flag indicate whether the desired goal pose is reached,

$$f_t^{\text{goal}} = \begin{cases} 1, & \text{if } (e_{d,t} < \epsilon_d) \wedge (e_{\psi,t} < \epsilon_\psi) \wedge (e_{v,t} < \epsilon_v), \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $(\epsilon_d, \epsilon_\psi, \epsilon_v)$ are small tolerance hyperparameters. Then, suppose a rich reward signal of the form $r_t = -(\alpha e_{d,t} + e_{\psi,t})$ is designed, which characterizes a weighted linear combination of different measures. This class of reward signals, trading-off various terms and providing feedback at *every* t , occurs frequently in the literature [18, 23, 24, 39]. However, as will be shown, for trajectory planning in an automotive setting (especially for nonholonomic vehicle models), it may easily lead to undesirable behavior. Suppose case (a) in Fig. 2 and a maximum simulation time T^{max} . Then, and omitting a discount factor for brevity,

$-T^{\max} \frac{\pi}{2} > -\sum_{t=0}^{\bar{T}} \alpha e_{d,t}^* + e_{\psi,t}^*$ may be obtained for accumulated rewards. Thus, the no-movement solution may incur more accumulated reward, $-T^{\max} \frac{\pi}{2}$, in comparison to the true solution on the right-hand side of the inequality sign.

Similarly, for specific (α, T^{\max}) , the second scenario (b) in Fig. 2 can also return a no-movement solution, since the initial angle is already coinciding with the target angle. Hence, for a specific (α, T^{\max}) -combination, the accumulated reward for not moving may exceed the value for the actual desired solution.

The third scenario (c) in Fig. 2 shows that even if reducing rich rewards to a single measure, e.g., $r_t = -e_{d,t}$, an undesired standstill may result. This occurs especially in the presence of obstacles (and maze-like situations in general).

To summarize, for finite T^{\max} , the design of rich reward signals is not straightforward and can easily result in solution trajectories that may even be globally optimal w.r.t. accumulated reward, however, prohibit to solve the original problem of determining a trajectory from initial to desired target state.

The Problematic of Curriculum Learning In [4], curriculum learning (CL) is discussed as a method to speed up learning by providing the learning agent first with simpler examples before gradually increasing complexity. Analogies to humans and animals are drawn. The same paper also acknowledges the difficulty of determining “interesting” examples [4, Sect. 7] that optimize learning progress. Indeed, CL entails the following issues. First, “simpler” tasks need to be designed. Second, these tasks must first be solved before their result can serve as initialization to more complex tasks. In contrast, without CL, the entire solution time can be devoted to the complex tasks rather than being partitioned into easier and difficult tasks. Third, the solution of an easier task does not necessarily represent a better initialization to a harder problem in comparison to an alternative random initialization. For example, consider the scenario in Fig. 3. The solution of the simpler task does not serve as a better initialization than a purely random initialization of weights. This is since the final solution requires outreaching steering and possibly reversing of the vehicle. The simpler task just requires forward driving and stopping. This basic example illustrates the need for careful manual selection of suitable easier tasks for CL.

The Benefits of Maximal Sparse Rewards in Combination with Virtual Velocity Constraints In the course of this work, many reward shaping methods were tested. These include, first, solving “simpler” tasks by first dismissing target angles limited to 30° -deviation from the initial heading. Second, ϵ -tolerances were initially relaxed before gradually decreasing them.

Third, it was also tested to initially solve a task for only the ϵ_d -criterion, then for both $(\epsilon_d, \epsilon_\psi)$, and only finally for $(\epsilon_d, \epsilon_\psi, \epsilon_v)$. Here, also varying sequences (e.g., ϵ_ψ before ϵ_d) were tested. No consistent improvement could be observed. On the contrary, solving allegedly simpler task reduced available solver time for the original “hard” problems. Without CL the entire solution time can be devoted to the complex tasks.



Fig. 3. The problematic of *curriculum learning*. The difficulty of selecting “simple” examples is illustrated. The original problem with start (black) and goal (red dashed) state is denoted in (a). A “simpler” problem is given in (b).

Based on these findings, our preferred reward design method is maximally sparse and defined by

$$r_t = \begin{cases} -\infty, & \text{if } f_t^{\text{crash}} == 1, \\ -1, & \text{otherwise,} \end{cases} \quad (2)$$

$$F_i = \begin{cases} 1, & \text{if } \sum_{\tau=t-T^{\text{goal}}+1}^t f_{\tau}^{\text{goal}} == T^{\text{goal}}, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

where f_t^{goal} from (1), and f_t^{crash} denoting an indicator flag for a vehicle crash. Thus, upon $F_i = 1$ the RL problem is considered as solved. In addition, the pathlength incurred for a transition from sampling time t to $t + 1$ is defined as

$$\Delta p_t = -\sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2}. \quad (4)$$

As elaborated further below, accumulated total pathlength is introduced to rank solution candidates that solve all desired training tasks.

The integral for F_i is defined for generality, in particular for problems such as the inverted pendulum [2], which are considered to be solved only after stabilization is demonstrated for sufficiently many consecutive T^{goal} time steps. Note, however, that this is not required for an automotive setting. Here, it must be $T^{\text{goal}} = 1$. Only then learning with $v^{\text{goal}} \neq 0$ is possible. Other criteria and trade-offs for Δp_t are possible (e.g., accumulated curvature of resulting paths and a minmax objective therefore). The negation is introduced for maximization (“hill climbing”-convention). Note that the preferred reward signal is maximal sparse, returning -1 , for all times up until reaching the target. It represents a *tabula rasa* solution critizised in [39] for its maximal sparsity. Indeed, standalone it was not sufficient for efficient learning when also accounting for a velocity target v^{goal} . Therefore, *virtual velocity constraints* (VVCs) in target proximity are introduced. Two variants are discussed. First, VVCs spatially dependent on $e_{d,t}$,

$$v_t^m = \begin{cases} v^m, & \text{if } e_{d,t} \geq R_v^{\text{thresh}}, \\ v^{\text{goal}} + \frac{(v^m - v^{\text{goal}})}{R_v^{\text{thresh}}} e_{d,t}, & \text{otherwise,} \end{cases} \quad (5)$$

where $m \in \{\max, \min\}$, and R_v^{thresh} is a hyperparameter (e.g., the range-view ahead or a heuristic). Second and alternatively, VVCs may be defined as spatially invariant with a constant margin (e.g., 5 km/h) around the target velocity. For

both variants, the neural network output that regulates velocity is scaled with updated v_t^{\min} and v_t^{\max} constraints (i.e., using (5) for spatially dependent VVCs).

Let us further legitimize VVCs. Since speed is a decision variable it can always be constrained artificially. This justifies the introduction of VVCs. In (5), bounds are set to affinely converge towards v^{goal} in spatial goal proximity. This is a heuristic choice. Note that the affine choice does not necessarily imply constant accelerations. This is since (5) is spatially parameterized. Note further that physical actuator rate constraints still hold when a_t is applied to the vehicle.

It was also tested to constrain δ_t . The final heading pose implies circles prohibited from trespassing because of the nonholonomic vehicle dynamics. It was tested to add these as virtual obstacles. However, this did not accelerate learning.

Finally, note that VVCs artificially introduce hard constraints and thus shape the learning result w.r.t velocity, at least towards the end of the trajectory. Two comments are made. First, in receding online operation, with additional frequent resetting of targets, this shaping effect is reduced since only the first control of a planned trajectory is applied. Second, in case of spatially dependent VVCs the influence of hyperparameter R_v^{thres} only becomes apparent during parking when following the trajectory up until standstill. Here, however, no significant velocity changes are desired, such that the R_v^{thres} -choice is not decisive. Ultimately, note that sparse rewards naturally avoid the need to introduce trade-off hyperparameters for the weighting of states in different units. This permits solution trajectories between start and goal poses to naturally evolve without biasing them by provision of rich references to track.

To summarize this section. It was illustrated that the design of *rich reward* signals as well as *curriculum learning* can be problematic. Therefore, *maximal sparse rewards* in combination with *virtual velocity constraints* are proposed.

3.3 The Role of Tolerances ϵ

Tolerances ϵ in (1) hold an important role for 2 reasons. On one hand, nonzero ϵ result in deviations between actually learnt \hat{z}^{goal} and originally desired goal z^{goal} . On the other hand, very small ϵ prolong learning time. Two scenarios apply.

First, for a network trained on a large-scale and dense grid of training tasks and for small ϵ , during online operation, suitable control commands are naturally interpolated even for setpoints not seen during training. The concept of natural interpolation through motion primitives encoded in neural networks is the core advantage over methods relying on look-up tables with stored trajectories, which require to solve time-critical search problems. For example, in [27] exhaustive search of the entire lattice-graph is conducted online on a GPU. In [25], a total of about 100 motion primitives is considered. Then, online an integer program is solved by enumeration using maximal progress along the centerline as criterion for selection of the best motion primitive. In contrast, for control using neural networks as function approximators this search is not required.

Second, the scenario was considered in which existing training hardware does (i) not permit large-scale encoding, and (ii) only permits to use larger ϵ -tolerances to limit training time. Therefore, the following method is devised.

First, tuples $(\hat{z}^{\text{goal}}, z^{\text{goal}})$ are stored for each training task. Then, during online operation, for any setpoint, z^{setpoint} , the closest (according to a criterion) \hat{z}^{goal} from the set of training tasks is searched, before the corresponding z^{goal} is applied to the network controller. Two comments are made. First, in order to reach \hat{z}^{goal} , z^{goal} must be applied to the network. Therefore, *tuples* need to be stored. Second, even though this method now also includes a search, it still holds an important advantage over lattice-based methods. This is the compression of the look-up table in the network weights. Only tuples need to be stored—not entire trajectories. This is especially relevant in view of limited hardware memory. Thus, through encoding, potentially many more motion primitives can be stored.

In practice, the first scenario is preferable. It is also implementable. First, see Sect. 2.4 for computational opportunities. Second, neural networks have in principle unlimited function approximation capability [43]. Hence, implementation is primarily a question of intelligent task setup, and computational power.

3.4 Main Algorithm – TSHC

Algorithm 1 is proposed for simple gradient-free model-based reinforcement learning. The name is derived from the fact of (i) learning from separate training tasks, and (ii) a hill climbing update of parameters (greedy local search).

Let us elaborate on definitions. Analysis is provided in Sect. 3.5. First, all network parameters are lumped into variable θ . Second, the perturbation step 8 in Algorithm 1 has to be interpreted accordingly. It implies parameter-wise affine perturbations with zero-mean Gaussian noise and spherical variance σ_{pert}^2 . Third, $\mathcal{X}(\cdot)$, $\mathcal{R}(\cdot)$ and $\mathcal{Z}(\cdot)$ in Steps 14–16 denote functional mappings between properties defined in the preceding sections. Fourth, hyperparameters are stated in Step 1. While N_{restarts} , N_{iter}^{\max} , n , N_{tasks} and T^{\max} denote lengths of different iterations, $\beta > 1$ is used for updating of σ_{pert} in Step 35 and 37. Fifth, for every restart iteration, i_{restart} , multiple parameter iterations are conducted, at most N_{iter}^{\max} many. Sixth, in Steps 25 and 29 hill climbing is conducted, when (i) all tasks have been solved for current i_{iter} , or (ii) not all tasks have yet been solved, respectively. Seventh, there are 2 steps in which an early termination of iterations may occur: Step 21 and 41. The former is a must. Only then learning with $v^{\text{goal}} \neq 0$ is possible. The latter termination criterion in Step 41 is optional. If dismissed, a refinement step is implied. Thus, even though all N_{tasks} tasks have been solved, parameter iterations (up until N_{iter}^{\max}) are continued. Eighth, note that a discount hyperparameter γ , common to gradient-based RL methods [42], is not required. This is since it is irrelevant in the maximally sparse reward setting. Ninth, nested parallelization is in principle possible with an inner and outer parallelization of Steps 10–22 and 7–22, respectively. The former refers to N_{tasks} solutions for a given parameter vector θ_i , whereas the latter parallelizes n parameter perturbations. Finally, there are 3 options considered for σ_{pert} -selection. First, holding an initial σ_{pert} -selection constant throughout TSHC. Second, updating σ_{pert} randomly (e.g., uniformly distributed between 10 and 1000), whereby this can be implemented either in Step 4 at every i_{restart} , or in

Algorithm 1: TSHC

1 **Input:** system model, network structure, N_{tasks} training tasks; N_{restarts} , N_{iter}^{\max} , n ,
 T^{\max} , ϵ , and a method to update σ_{pert} : constant, random or adaptive based on $(\beta,$
 $\sigma_{\text{pert}}^{\min}, \sigma_{\text{pert}}^{\max})$.

2 Initialize $\theta^* \leftarrow \emptyset$, $N^* \leftarrow 0$, $P^* \leftarrow -\infty$, $J^* \leftarrow 0$.

3 **for** $i_{\text{restart}} = 1, \dots, N_{\text{restarts}}$ **do**

4 Initialize θ randomly, and $\sigma_{\text{pert}} \leftarrow \sigma_{\text{pert}}^{\max}$, $N_{\text{old}}^{\text{tasks},*} \leftarrow 0$.

5 **for** $i_{\text{iter}} = 1, \dots, N_{\text{iter}}^{\max}$ **do**

6 % RUN ASYNCHRONOUSLY:

7 **for** $i = 1, \dots, n$ **do**

8 Perturb $\theta_i \leftarrow \theta + \sigma_{\text{pert}} \zeta$, with $\zeta \sim \mathcal{N}(0, I)$.

9 Initialize $N_i^{\text{tasks},*} \leftarrow 0$, $P_i \leftarrow 0$, $J_i \leftarrow 0$.

10 **for** $i_{\text{task}} = 1, \dots, N_{\text{tasks}}$ **do**

11 Initialize z_0 (and LSTM and GRU cells).

12 **for** $t = 0, \dots, T^{\max} - 1$ **do**

13 Read s_t from i_{task} -environment.

14 $a_t \leftarrow \mathcal{X}(s_t, \theta_i)$.

15 $(r_t, \Delta p_t, f_t^{\text{goal}}) \leftarrow \mathcal{R}(s_t, a_t)$.

16 $z_{t+1} \leftarrow \mathcal{Z}(z_t, a_t)$.

17 $P_i \leftarrow P_i + \Delta p_t$.

18 $J_i \leftarrow J_i + r_t$.

19 F_i according to (3).

20 **if** $(F_i == 1) \vee (r_t == -\infty)$ **then**

21 | Break t -loop.

22 | $N_i^{\text{tasks},*} \leftarrow N_i^{\text{tasks},*} + F_i$.

23 % DETERMINE i^* :

24 **if** $\max_i \{N_i^{\text{tasks},*}\}_{i=1}^n == N_{\text{tasks}}$ **then**

25 | $i^* = \arg \max_i \{P_i \mid N_i^{\text{tasks},*} == N_{\text{tasks}}\}_{i=1}^n$.

26 | **if** $P_{i^*} > P^*$ **then**

27 | | $(\theta^*, N^*, P^*, J^*) \leftarrow (\theta_{i^*}, N_{i^*}, P_{i^*}, J_{i^*})$.

28 **else**

29 | $i^* = \arg \max_i \{J_i\}_{i=1}^n$.

30 | **if** $(J_{i^*} > J^*) \wedge (P^* == -\infty)$ **then**

31 | | $(\theta^*, N^*, P^*, J^*) \leftarrow (\theta_{i^*}, N_{i^*}^{\text{tasks}}, P_{i^*}, J_{i^*})$.

32 | $N_{\text{tasks},*} \leftarrow N_{i^*}^{\text{tasks},*}$.

33 % UPDATE PARAMETERS:

34 **if** $N_{\text{tasks},*} > N_{\text{old}}^{\text{tasks},*}$ **then**

35 | $\sigma_{\text{pert}} \leftarrow \max(\frac{1}{\beta} \sigma_{\text{pert}}, \sigma_{\text{pert}}^{\min})$.

36 **else if** $N_{\text{tasks},*} < N_{\text{old}}^{\text{tasks},*}$ **then**

37 | $\sigma_{\text{pert}} \leftarrow \min(\beta \sigma_{\text{pert}}, \sigma_{\text{pert}}^{\max})$.

38 | $\theta \leftarrow \theta_{i^*}$ and $N_{\text{old}}^{\text{tasks},*} \leftarrow N_{i^*}^{\text{tasks},*}$.

39 % OPTIONAL:

40 **if** $N_i^{\text{tasks},*} == N_{\text{tasks}}$ **then**

41 | Break i_{iter} -loop. % no further refinement step.

42 **Output:** $(\theta^*, N^*, P^*, J^*)$.

Step 6 at every $(i_{\text{restart}}, i_{\text{iter}})$ -combination. Third, σ_{pert} may be adapted according to progress in $N^{\text{tasks},*}$, as outlined in Algorithm 1. For the first 2 options of selecting σ_{pert} , Steps 34–37 are dismissed and at least β can be dismissed from the list of hyperparameters in Step 1.

3.5 Analysis

According to classifications in [12], TSHC is a gradient-free instance-based simulation optimization method, generating new candidate solutions based on only the current solution and random search in its neighborhood. Because of its hill climbing (greedy) characteristic, it differs from (i) *evolutionary* (population-based) methods that construct solution by combining others typically using weighted averaging [40, 49], and (ii) from model-based methods that use *probability distributions* on the space of solution candidates, see [12] for a survey. In its high-level structure, Algorithm 1 can be related to the COMPASS algorithm [51]. Within a global stage, they identify several possible regions with locally optimal solutions. Then, they find local optimal solutions for each of the identified regions, before they select the best solution among all identified locally optimal solutions. In our setting, these regions are enforced as the separate training tasks and the best solution for all of these is selected.

In combination with sufficiently large n , σ_{pert} must be large enough to permit sufficient exploration such that a network parametrization solving all tasks can be found. In contrast, the effect of decreasing σ_{pert} with an increasing number of solved tasks is that, ideally, a speedup in learning progress results from the assignment of more of solution candidates θ_i closer in variance to a promising θ .

Steps 29–31 are discussed. For the case that for a specific i_{iter} -iteration not all tasks have yet been solved, $i^* = \arg \max_i \{N_i^{\text{tasks},*}\}_{i=1}^n$ has been considered as an alternative criterion for Step 29. Several remarks can be made. First, Step 29 and the alternative are not equal. This is because, in general, different tasks are solved in a different number of time steps. However, the criteria are approximately equivalent for sparse rewards (since J_i accumulates constants according to (2)), and especially for large T^{\max} . The core advantage of employing Step 29 in TSHC is that it can, if desired, also be used in combination with rich rewards to accelerate learning progress (*if* a suitable rich reward signal can be generated). In such a scenario, i^* according to Step 29 is updated towards most promising J_{i^*} , then representing the accumulated rich reward. Thus, in contrast to (2), a rich reward could be represented, for example, by a weighted sum of squared errors between state $z_t \in \mathbb{R}^{n_z}$ and a reference $z_t^{\text{ref}} \in \mathbb{R}^{n_z}$,

$$r_t = \begin{cases} -\infty, & \text{if } f_t^{\text{crash}} == 1, \\ -\sum_{l=0}^{n_z-1} \alpha_l (z_t(l) - z_t^{\text{ref}}(l))^2, & \text{otherwise,} \end{cases} \quad (6)$$

where α_l are trade-off hyperparameters and scalar elements of vectors are indexed by l in brackets. Another advantage of the design in Algorithm 1 according to Step 29–31 is its *anytime solution* character. Even if not all N_{tasks} are

solved, the solution returned for the tasks that are solved, typically is of good quality and optimized according to Steps 29–31.

If for all N_{tasks} tasks there exists a feasible solution for a given system model and a sufficiently expressive network structure parameterized by θ , then Algorithm 1 can find such parametrization for sufficiently large hyperparameters N_{restarts} , N_{iter}^{\max} , n , T^{\max} and $\sigma_{\text{pert}}^{\max}$. The solution parametrization θ^* is the result from the initialization Step 4 and parameter perturbations according to Step 8, both nested within multiple iterations. As noted in [19], for optimization via simulation, a global convergence guarantee provides little practical meaning other than reassuring a solution will be found “eventually” when simulation effort goes to infinity. However, the same reference also states that a convergence property is most meaningful if it can help in designing suitable stopping criteria. In our case, there are 2 such conceptual levels of stopping criteria: first, the solution of all training tasks, and second, the refinement of solutions.

Control design is implemented hierarchically in 2 steps. First, suitable training tasks (desired motion primitives) are defined. Then, these are encoded in the network. This has practical implications. First, it encourages to train on deterministic tasks. Furthermore, at every i_{iter} , it is simultaneously trained on all of these tasks. Thus, the best parametrization, θ^* , is defined via Step 25, maximizing the accumulated P -measure over *all* tasks. Second, it enables to provide certificates on the learnt performance, which can be provided by stating (i) the employed vehicle model, and (ii) the list of encoded motion primitives. Note that such certificates cannot be given for the class of *stochastic* continuous action RL algorithms derived from the Stochastic Policy Gradient Theorem [46], which *draw* their controls, typically from a Gaussian distribution. This class includes all stochastic actor-critic algorithms, including A3C [28] and PPO [42].

3.6 Discussion and Comparison with Related RL Work

Related continuous control methods are discussed, focusing on one stochastic [42], one deterministic policy gradient method [24], and one evolution strategy [40]. The methods are discussed in detail to underline aspects of TSHC.

First, the *stochastic policy gradient* method PPO [42] is discussed. Suppose that a stochastic continuos control vector is sampled from a Gaussian distribution parameterized¹ by θ such that $a_t \sim \pi(a_t|s_t, \theta)$. Then,

$$J(\theta) = \mathbb{E}_{s_t, a_t} [g(s_t, a_t \sim \pi(a_t|s_t, \theta))], \quad (7)$$

is defined as the expected accumulated and time-discounted reward when at s_t drawing a_t , and following the stochastic policy for all subsequet times when acting in the simulation environment. Since function $g(s_t, a_t)$ is a priori not know, it is parameterized by $\theta_{V, \text{old}}$ and estimated. Using RL-terminology, in the PPO-setting, $g(s_t, a_t)$ represents the *advantage function*. Then, using the

¹ In this setting, mean and variance of the Gaussian distribution are the output of a neural network whose parameters are summarized by lumped θ .

“log-likelihood trick”, and subsequently a first-order Taylor approximation of $\log(\pi(a_t|s_t, \theta))$ around some reference $\pi(a_t|s_t, \theta_{\text{old}})$, the following parameterized cost function is obtained as an *approximation* of (7),

$$\tilde{J}(\theta) = \underset{s_t, a_t}{\mathbb{E}} \left[\hat{g}_t(\theta_{V, \text{old}}) \frac{\pi(a_t|s_t, \theta)}{\pi(a_t|s_t, \theta_{\text{old}})} \right]. \quad (8)$$

Finally, (8) is modified to the final PPO-cost function [42].

$$\hat{J}(\theta) = \underset{s_t, a_t}{\mathbb{E}} \left[\min \left(\hat{g}_t(\theta_{V, \text{old}}) \frac{\pi(a_t|s_t, \theta)}{\pi(a_t|s_t, \theta_{\text{old}})}, \text{clip}\left(\frac{\pi(a_t|s_t, \theta)}{\pi(a_t|s_t, \theta_{\text{old}})}, 1 - \epsilon, 1 + \epsilon\right) \hat{g}_t(\theta_{V, \text{old}})\right) \right], \quad (9)$$

whereby the advantage function is estimated by the policy parameterized by $(\theta_{V, \text{old}}, \theta_{\text{old}})$, which is run for T consecutive time steps such that for all t the tuples $(s_t, a_t, r_t, s_{t+1}, \hat{g}_t(\theta_{V, \text{old}}))$ can be added to a *replay buffer*, from which later minibatches are drawn. According to [42], the estimate is $\hat{g}_{T-1}(\theta_{V, \text{old}}) = \kappa_{T-1}$ with $\kappa_{T-1} = r_{T-1} + \gamma V(s_T, \theta_{V, \text{old}}) - V(s_{T-1}, \theta_{V, \text{old}})$, $\hat{g}_{T-2}(\theta_{V, \text{old}}) = \kappa_{T-2} + \gamma \lambda \hat{g}_{T-1}(\theta_{V, \text{old}})$ and so forth until $\hat{g}_0(\theta_{V, \text{old}})$, and where $V(s, \theta_{V, \text{old}})$ represents a *second*, the so-called critic neural network. Then, using uniform randomly drawn minibatches of size M , parameters (θ_V, θ) of both networks are updated according to $\arg \min_{\theta_V} \frac{1}{M} \sum_{i=0}^{M-1} (V(s_i, \theta_V) - (\hat{g}_i(\theta_{V, \text{old}}) - V(s_i, \theta_{V, \text{old}})))^2$ and $\arg \max_{\theta} \frac{1}{M} \sum_{i=0}^{M-1} \mathcal{G}_i(\theta)$, with $\mathcal{G}_i(\theta)$ denoting the argument of the expectation in (9) evaluated at time-index i . This discussion is given to underline following observations. With first the introduction of a parameterized estimator, then a first-order Taylor approximation, and then clipping, (9) is an arguably crude approximation of the original problem (7). Second, the complexity with *two* actor and critic networks is noted. Typically, both are of the same dimensions apart from the output layers. Hence, when not sharing weights, approximately *twice* as many parameters are required. However, *when* sharing any weights between actor and critic network, then optimization function (9) must be extended accordingly, which introduces another approximation step. Third, note that gradients of both networks must be computed for backpropagation. Fourth, the dependence on *rich* reward signals is stressed. As long as the current policy does not find a solution candidate, in a *sparse* reward setting, all r_i are uniform. Hence, there is no information permitting to find a suitable parameter update direction and all of the computational expensive gradient computations are essentially not usable². Thus, network parameters are still updated entirely at random. Moreover, *even if* a solution candidate trajectory was found, it is easily averaged out through the random minibatch update. This underlines the problematic of *sparse* rewards for PPO. Fifth, A3C [28] and PPO [42] are by nature *stochastic* policies, which

² It is mentioned that typically the first, for example, 50000 samples are collected *without* parameter update. However, even then that threshold must be selected, and the fundamental problem still perseveres.

draw their controls from a Gaussian distribution (for which mean and variance are the output of a trained network with current state as its input). Hence, exact repetition of any task (e.g., the navigation between 2 locations) cannot be guaranteed. It can only be guaranteed if dismissing the variance component, and consequently using solely the mean for deterministic control. This can be done in practice, however, introduces another approximation step.

Deterministic policy gradient method DDPG [24] is discussed. Suppose a deterministic continuous control vector parameterized such that $a_t = \mu(s_t, \theta)$. Then, the following cost function is defined,

$$J(\theta) = \underset{s_t, a_t}{\mathbb{E}} [g(s_t, a_t = \mu(s_t, \theta))] = \underset{s_t}{\mathbb{E}} [g(s_t, \mu(s_t, \theta))].$$

Its gradient can now be computed by applying the chain-rule for derivatives [44]. Introducing a parameterized estimate of $g(s_t, a_t)$, which here represents the *Q-function* or *action value function* (in contrast to the advantage function in above stochastic setting), the final DDPG-cost function [24] is

$$J(\theta) = \underset{s_t}{\mathbb{E}} [\hat{g}(s_t, \mu(s_t, \theta), \theta_Q)].$$

Then, using minibatches, critic and actor network parameters (θ_Q, θ) are updated as $\arg \min_{\theta_Q} \frac{1}{M} \sum_{i=0}^{M-1} (\hat{g}(s_i, a_i, \theta_Q) - (r_i + \gamma Q(s_{i+1}, \mu(s_{i+1}, \theta_{\text{old}}), \theta_{Q,\text{old}})))^2$ and $\arg \min_{\theta} \frac{1}{M} \sum_{i=0}^{M-1} Q(s_i, \mu(s_i, \theta), \theta_Q)$, with slowly tracking target network parameters $(\theta_{Q,\text{old}}, \theta_{\text{old}})$. Several remarks can be made. First, the Q-function is updated towards only its *one-step* ahead target. It is obvious that rewards are therefore propagated *very* slowly. For sparse rewards this is even more problematic than for rich rewards, especially because of the additional danger of averaging out important update directions through random minibatch sampling. Furthermore, and analogous to the stochastic setting, for the sparse reward setting, as long as no solution trajectory was found, all of the gradient computations are not usable and all network parameters are still updated entirely at random. DDPG is an *off-policy* algorithm. In [24], exploration of the simulation environment is achieved according to the current policy plus additive noise following an *Ornstein-Uhlenbeck* process. This is a mean-reverting linear stochastic differential equation [13]. A first-order Euler approximation thereof can be expressed as the action exploration rule $a_t = \mu(s_t, \theta)(1 - P_\theta^{\text{OU}}) + P_\sigma^{\text{OU}}\epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, with hyperparameters $(P_\theta^{\text{OU}}, P_\sigma^{\text{OU}}) = (0.15, 0.2)$ in [24]. This detail is provided to stress a key difference between policy gradient methods (both stochastic and deterministic), and methods such as [40] and TSHC. Namely, while the former methods sample controls from the stochastic policy or according to *heuristic* exploration noise before updating parameters using minibatches of *incremental* tuples (s_i, a_i, r_i, s_{i+1}) plus $\hat{g}_i(\theta_{V,\text{old}})$ for PPO, the latter directly work in the *parameter space* via local perturbations, see Step 8 of Algorithm 1. This approach appears particularly suitable when dealing with sparse rewards. As outlined above, in such setting, parameter updates according to policy gradient

methods are also entirely at random, however, with the computationally significant difference of first an approximately four times as large parameter space and, second, the unnecessary costly solution of non-convex optimization problems as long as no solution trajectory has been found. A well-known issue in training neural networks is the problem of vanishing or exploding gradients. It is particularly relevant for networks with saturating nonlinearities and can be addressed by batch [20] and layer normalization [3]. In both normalization approaches, *additional* parameters are introduced to the network which must be learnt (bias and gains). These issues are not relevant for the proposed gradient-free approach.

This paper is inspired by and most closely related to [40]. The main differences are discussed. The latter evolutionary (population-based) strategy updates parameters using a *stochastic gradient estimate*. Thus, it updates $\theta \leftarrow \theta + \alpha \frac{1}{n\sigma} \sum_{i=1}^n R_i \zeta_i$, where hyperparameters α and σ denote the learning rate and noise standard deviation, and where R_i here indicates the stochastic scalar return provided by the simulation environment. This weighted averaging approach for the stochastic gradient estimate is not suitable for learning based on separate deterministic training tasks in combination with maximally sparse rewards. Here, hill climbing is more appropriate. This is since most of the n trajectory candidates do not end up at z^{goal} and are therefore not useful. Note also that only the introduction of virtual velocity constraints permitted us to quickly train with maximally sparse rewards. It is well known that for gradient-based training, especially of RNNs, the learning rate (α in [40]) is a critical hyperparameter choice. For hill climbing this issue does not occur. Likewise, *fitness shaping* [49], also used in [40], is not required. Note that above σ has the same role as σ_{pert} . Except, in our setting, it additionally is adaptive according to Steps 29 and 31 in Algorithm 1. As implemented, this is only possible when training on multiple separate tasks. Other differences include the parallelization method in [40], where random seeds shared among workers permit each worker to only need to send and receive the scalar return of an episode to and from each other worker. All perturbations and parameters are then reconstructed locally by each worker. Thus, for n workers there are n reconstructions at each parameter-iteration step. This requires precise control of each worker and may lead to differing CPU utilizations among workers due to differing episode lengths. Therefore, they use a capping strategy on maximal episode length. In contrast, TSHC is simpler with one synchronized parameter update, which is then sent to all workers.

4 Numerical Simulations

This section highlights different aspects of Algorithm 1. Numerical simulations of Sect. 4.1 and 4.2 were conducted on a laptop with an Intel Core i7 CPU @2.80 GHz×8, 15.6 GB of memory, and with the only libraries employed Python’s `numpy` and `multiprocessing`. Furthermore, in Sect. 4.1 for the implementation of 2 comparative policy gradient methods, Tensorflow (without GPU-support) was used. Using these (for deep learning) very limited resources enabled to evaluate the method’s potential when significant computational power

is not available. For more complex problems the latter is required. Therefore, in Sect. 4.3 TSHC is implemented in `Cuda C++` and 1 GPU is used.

4.1 Experiment 1: Comparison with Policy Gradient Methods

To underline conceptual differences between TSHC and 2 policy gradient methods DDPG [24] and PPO [42], a freeform navigation task with $z_0 = [0, 0, 0, 0]$ and $z^{\text{goal}} = [20, 0, \pi/4, 0]$ was considered, where vector $z = [x, y, \psi, v]$ summarizes four of the vehicle's states. The same network architecture from [42] is used: a fully-connected MLP with 2 hidden layers of 64 units before the output layer. Eventhough this is the basic setup, considerable differences between DDPG, PPO and TSHC are implied. Both DDPG and PPO are each composed of a total of four networks: one actor, one critic, one actor target and one critic target network. For DDPG, further parameters result from batch normalization [20]. The number of parameters N_{var} that need to be identified are indicated in Table 1. To enable a fair comparison, all of DPPG, PPO and TSHC are permitted to train on 1000 full rollouts according to their methods, whereby each rollout lasts at most 100 timesteps. Thus, for TSHC, $N_{\text{restarts}} = 1$ and $n = 1000$ are set. For both PPO and DDPG, this implies 1000 iterations. Results are summarized in Fig. 4. The following observations can be made. First, in comparison to TSHC, for both DDPG and PPO significantly more parameters need to be identified, see Table 1. Second, DDPG and PPO do *not* solve the task based on 1000 training simulations. In contrast, as Fig. 4 demonstrates, TSHC has a much better exploration strategy resulting from noise perturbations in the *parameter space*. It solves the task in just 2.1s. Finally, note that no σ_{pert} -iteration is conducted. It is not applicable since a single task is solved with an initial $\sigma_{\text{pert}}^{\max} = 10$. Because of these findings (other target poses were tested with qualitatively equivalent results) and the discussion in Sect. 3.6 about the handling of *sparse rewards* and the fact that DDPG and PPO have no useful gradient direction for their parameter update or may average these out through random minibatch sampling, the focus in the subsequent sections is on TSHC and its analysis.

Table 1. Experiment 1. Number of Scalar Parameters (Weights) that Need to be Identified for DDPG, PPO and TSHC, Respectively. TSHC Requires to Identify the Least by a Large Margin, Roughly by a Factor 4.

	DDPG	PPO	TSHC
N_{var}	19078	18440	4610

4.2 Experiment 2: Inverted Pendulum

The discussion of tolerance levels in Sect. 3.3 motivated an alternative approach for tasks requiring stabilization. An analogy to optimal control is drawn. In linear

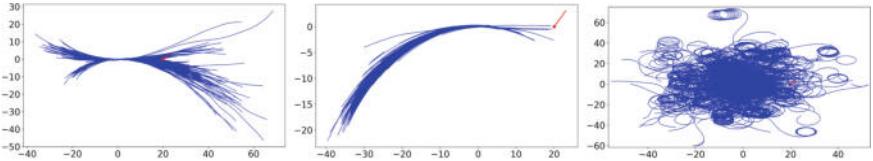


Fig. 4. Experiment 1. 1000 training trajectories resulting from the application DDPG (Left), PPO (Middle) and TSHC (Right), respectively. The effect of virtual constraints on velocity is particularly visible for DDPG. For the given hyperparameter setting [42, Tab. 3], the trajectories for PPO have little spread and are favoring reverse driving. TSHC has a much better exploration strategy resulting from noise perturbations in the parameter space. The task is solved by TSHC in only 2.1s of learning time, when terminating upon the first solution found (no refinement step, no additional restart).

finite horizon MPC, closed-loop stability can be guaranteed through a terminal state constraint set which is invariant for a *terminal controller*, often a linear quadratic regulator (LQR), see [26]. In a RL setting, the following procedure was considered. First, design a LQR for stabilization. Second, compute the region of attraction of the LQR controller [47, Sect. 3.1.1]. Third, use this region of attraction as stopping criterion, *replacing* the heuristic ϵ -tolerance selection.

For evaluation, the inverted pendulum system equations and parameters from [2] were adopted (four states, one input). However, in contrast to [2], which assumes just 2 discrete actions (maximum and minimum actuation force), here a continuous control variable is assumed which is limited by the 2 bounds, respectively. There are 2 basic problems: stabilization in the upright position with initial state in the same position, as well as a swing-up from the hanging position plus consequent stabilization in the upright position. For the application of TSHC, $(N_{\text{restarts}}, N_{\text{iter}}^{\max}, n, T^{\max}, \beta, \sigma_{\text{pert}}^{\max}) = (3, 100, 100, 500, 2, 10)$ are set, and the same MLP-architecture from Sect. 4.1 is used. The following remarks can be made. First, the swing-up plus stabilization task was solved in 43.5s runtime of TSHC (without refinement step) and using sparse rewards (obtained in the upright position $\pm 12^\circ$). For all three restarts a valid solution was generated. Note that $T^{\max} = 500$ in combination with a sampling time [2] of 0.02s corresponds to 10s simulation time. Stabilization in the upright position was achieved from 2.9s on. Rich reward signals were also tested, with the deviation from current to goal angle as measure, however, did not accelerate learning.

In a second experiment, the objective was to *simultaneously* encode the following 2 tasks in the network: stabilization in the upright position with initial state in the same position *and* a swing-up from the hanging position plus consequent stabilization in the upright position. The runtime of TSHC (without refinement step) was 264.4s, with 2 of 3 restarts returning a valid solution and using sparse rewards. Instead of learning both tasks simultaneously, it was also attempted to learn them by selecting one of the 2 tasks at random at every i_{iter} , and consequently conducting Step 6–41. Since the 2 tasks are quite different, this procedure could not encode a solution for *both* tasks. This is mentioned to

exemplify the importance of training *simultaneously* on separate tasks, rather than training on a single tasks with (z_0, z^{goal}) -combinations varying over i_{iter} .

Finally, for the system parameters from [2], it was observed that the continuous control signal was operating mostly at saturated actuation bounds (switching in-between). This is mentioned for 2 reasons. First, aforementioned LQR-strategy could therefore never be applied since LQR assumes absence of state and input constraints. Second, it exemplifies the ease of RL-workflow with TSHC for quick nonlinear control design, even without significant system insights.

4.3 Experiment 3 and 4: GPU-based Training

Experiment 3 is characterized by transitioning from $z_0 = [0, 0, 0, 0]$ to $z^{\text{goal}} = [0, 0, \psi^{\text{goal}}, 0]$ with $\psi^{\text{goal}} \in \{0, 1, \dots, 180\}$ measured in $[\circ]$. This implies $N_{\text{tasks}} = 181$. The feature vector is selected as $s_t = [\frac{x^{\text{goal}} - x_t}{\Delta x_n}, \frac{y^{\text{goal}} - y_t}{\Delta y_n}, \frac{\psi^{\text{goal}} - \psi_t}{\Delta \psi_n}, \frac{v^{\text{goal}} - v_t}{\Delta v_n}, a_t[0]]$ with normalization constants in the denominators and $a_t[0] \in [-1, 1]$ indicating the steering angle-related network output (before scaling to δ_t). A high-resolution tolerance of $\epsilon_\psi = 1^\circ$ is set. In addition, $\epsilon_d = 0.25\text{m}$ and $\epsilon_v = 5\text{km/h}$. Sampling time is 0.01s. As neural network, a MLP-[5,8,2] is used, which implies 1 hidden layer with 8 units. For selections $N_{\text{restarts}} = 10$ and $N_{\text{iter}}^{\max} = 20$, MLP-[5,8,2] was the smallest possible network found to simultaneously encode all 181 training tasks. The second variant of VVCs discussed in Sect. 3.2 is employed. Furthermore, $\sigma_{\text{pert}} \sim \mathcal{U}[10, 1000]$, i.e., uniformly distributed at every $(i_{\text{restart}}, i_{\text{iter}})$ -combination. Several comments can be made. First, by application of *control mirroring* w.r.t. steering all of $\psi^{\text{goal}} \in \{181, \dots, 360\}$ can also be reached. Second, the total learning time (runtime of TSHC) to encode all 181 training tasks was 31.1 min. MLP-[5,8,2] implies a total of 66 parameters to learn. It is interesting that such a small network has enough function approximation capability to encode all 181 tasks for high-resolution $\epsilon_\psi = 1^\circ$. Third, note that the TSHC-trained network controller permits repeatable precision. As mentioned in Sect. 3.6, this is not attainable for stochastic policy gradient-based algorithms, which *draw* their control signals, typically from a Gaussian distribution. Fourth, the learning results are visualized in Fig. 5. These motivated to conduct Experiment 4 with identical basic training setup, however, now encoding only *one* task for the transition from $\psi_0 = 0^\circ$ to $\psi^{\text{goal}} = 180^\circ$. Notice the much reduced number of switches between forward and backward driving, and the different (x, y) -range.

The comparison of Experiment 3 and 4 in Fig. 5 emphasizes the interesting observation that the more motion primitives are encoded in a single network the less performant the single learnt motion trajectories are. This is believed to illustrate potential of partitioning the total number of designated tasks into *subsets* of training tasks for which separate networks are then learnt using TSHC. The promised advantages include faster overall trainig times, higher performance of learnt trajectories, and the ability to employ tiny networks with few parameters for each subset. Further analysis in this perspective is subject of ongoing work.

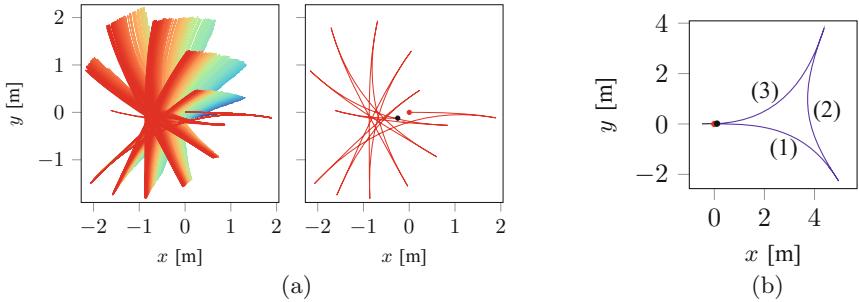


Fig. 5. **a** Experiment 3. In the left subplot, display of *all* 181 trajectories encoded in one MLP-[5,8,2]. Trajectories for each task are visualized in separate colors. In the right subplot, display of learnt result for only the most complex of the 181 training tasks, i.e., for $\psi^{\text{goal}} = 180^\circ$. Recall that $\epsilon_\psi = 1^\circ$ and $\epsilon_d = 0.25\text{m}$. The vehicle's start and end CoG-position is indicated by red and black balls. As indicated, the transition involves frequent forward and backward driving but is constrained locally around the (x, y) -origin. **b** Experiment 4. Display of learnt trajectory when encoding just *one* task in a MLP-[5,8,2] for the transition from $\psi_0 = 0^\circ$ to $\psi^{\text{goal}} = 180^\circ$. Brackets (1) and (3) imply forward, (2) reverse driving and their sequence. A black indicator visualizes the vehicle's final heading. The total learning time (incl. 10 restarts) was 10.6 s.

5 Conclusion

Within the context of automated vehicles, for the design of model-based controllers parameterized by neural networks a simple gradient-free reinforcement learning algorithm TSHC was proposed. The concept of (i) training on separate tasks with the purpose of encoding motion primitives, and (ii) employing sparse rewards in combinations with virtual velocity constraints in setpoint proximity were specifically advocated. Aspects of TSHC were illustrated in 4 numerical experiments. The presented method is not limited to automated driving. Most real-world learning applications are characterized by sparse rewards and the availability of high-fidelity system models that can be leveraged for offline training. Future work will focus on system models of various complexity (kinematic vs. dynamic vehicle models), the partitioning of tasks into separate subsets of tasks for which separate network parametrizations are learnt, analysis of different feature vectors and closed-loop evaluation during recursive waypoint tracking.

References

1. Akiba, T., Suzuki, S., Fukuda, K.: Extremely large minibatch sgd: training resnet-50 on imagenet in 15 minutes. [arXiv:1711.04325](https://arxiv.org/abs/1711.04325) (2017)
2. Anderson, C.W.: Learning to control an inverted pendulum using neural networks. IEEE Control. Syst. Mag. **9**(3), 31–37 (1989)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. [arXiv:1607.06450](https://arxiv.org/abs/1607.06450) (2016)
4. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: International Conference on Machine Learning, pp. 41–48. ACM (2009)

5. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., et al.: End to end learning for self-driving cars. [arXiv:1604.07316](https://arxiv.org/abs/1604.07316) (2016)
6. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: learning affordance for direct perception in autonomous driving. In: IEEE International Conference on Computer Vision, pp. 2722–2730 (2015)
7. Chen, S., Zhang, S., Shang, J., Chen, B., Zheng, N.: Brain inspired cognitive model with attention for self-driving cars. [arXiv:1702.05596](https://arxiv.org/abs/1702.05596) (2017)
8. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation [arXiv:1406.1078](https://arxiv.org/abs/1406.1078) (2014)
9. Dolgov, D., Thrun, S., Montemerlo, M., Diebel, J.: Path planning for autonomous vehicles in unknown semi-structured environments. *Int. J. Robot. Res.* **29**(5), 485–501 (2010)
10. Falcone, P., Borrelli, F., Asgari, J., Tseng, H.E., Hrovat, D.: Predictive active steering control for autonomous vehicle systems. *IEEE Trans. Control. Syst. Technol.* **15**(3), 566–580 (2007)
11. Frazzoli, E., Dahleh, M.A., Feron, E.: A hybrid control architecture for aggressive maneuvering of autonomous helicopters. *IEEE Conf. Decis. Control.* **3**, 2471–2476 (1999)
12. Fu, M.C., Glover, F.W., April, J.: Simulation optimization: a review, new developments, and applications. In: IEEE Winter Simulation Conference, pp. 13–pp. IEEE (2005)
13. Geering, H.P., Dondi, G., Herzog, F., Keel, S.: Stochastic systems. Course script (2011)
14. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* **3**(Aug), 115–143 (2002)
15. Gillespie, T.D.: Vehicle dynamics. Warren Dale (1997)
16. Glasmachers, T.: Limits of end-to-end learning. [arXiv:1704.08305](https://arxiv.org/abs/1704.08305) (2017)
17. Gray, A., Gao, Y., Lin, T., Hedrick, J.K., Tseng, H.E., Borrelli, F.: Predictive control for agile semi-autonomous ground vehicles using motion primitives. In: IEEE American Control Conference, pp. 4239–4244 (2012)
18. Heess, N., Sriram, S., Lemmon, J., Merel, J., Wayne, G., Tassa, Y. et al.: Emergence of locomotion behaviours in rich environments. [arXiv:1707.02286](https://arxiv.org/abs/1707.02286) (2017)
19. Hong, L.J., Nelson, B.L.: A brief introduction to optimization via simulation. In: IEEE Winter Simulation Conference, pp. 75–85 (2009)
20. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 (2015)
21. Jozefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: International Conference on Machine Learning, pp. 2342–2350 (2015)
22. Karaman, S., Walter, M.R., Perez, A., Frazzoli, E., Teller, S.: Anytime motion planning using the RRT. In: IEEE Conference on Robotics and Automation, pp. 1478–1483 (2011)
23. Koutnřík, J., Schmidhuber, J., Gomez, F.: Online evolution of deep convolutional network for vision-based reinforcement learning. In: International Conference on Simulation of Adaptive Behavior, pp. 260–269. Springer (2014)
24. Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., Wierstra, D.: Continuous control with deep reinforcement learning. [arXiv:1509.02971](https://arxiv.org/abs/1509.02971) (2015)

25. Liniger, A., Domahidi, A., Morari, M.: Optimization-based autonomous racing of 1: 43 scale rc cars. *Optim. Control. Appl. Methods* **36**(5), 628–647 (2015)
26. Mayne, D.Q., Rawlings, J.B., Rao, C.V., Scokaert, P.O.: Constrained model predictive control: stability and optimality. *Automatica* **36**(6), 789–814 (2000)
27. McNaughton, M., Urmson, C., Dolan, J.M., Lee, J.-W.: Motion planning for autonomous driving with a conformal spatiotemporal lattice. In: IEEE Conference on Robotics and Automation, pp. 4889–4895 (2011)
28. Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., Kavukcuoglu, K.: Asynchronous methods for deep reinforcement learning. In: International Conference on Machine Learning, pp. 1928–1937 (2016)
29. National Highway Traffic Safety Administration. Traffic safety facts, 2014: a compilation of motor vehicle crash data from the fatality analysis reporting system and the general estimates system. dot hs 812261. Department of Transportation, Washington, DC (2014)
30. Nvidia. Tesla P100. <https://images.nvidia.com/content/tesla/pdf/nvidia-tesla-p100-PCIe-datasheet.pdf> (2016)
31. Paden, B., Čáp, M., Yong, S.Z., Yershov, D., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Trans. Intell. Veh.* **1**(1), 33–55 (2016)
32. Paxton, C., Raman, V., Hager, G.D., Kobilarov, M.: Combining neural networks and tree search for task and motion planning in challenging environments. [arXiv:1703.07887](https://arxiv.org/abs/1703.07887) (2017)
33. Plessen, M.G.: Trajectory planning of automated vehicles in tube-like road segments. In: IEEE Conference on Intelligent Transportation Systems, pp. 83–88 (2017)
34. Plessen, M.G., Bernardini, D., Esen, H., Bemporad, A.: Multi-automated vehicle coordination using decoupled prioritized path planning for multi-lane one-and bi-directional traffic flow control. In: IEEE Conference on Decision and Control, pp. 1582–1588 (2016)
35. Plessen, M.G., Bernardini, D., Esen, H., Bemporad, A.: Spatial-based predictive control and geometric corridor planning for adaptive cruise control coupled with obstacle avoidance. *IEEE Trans. Control. Syst. Technol.* (2017)
36. Plessen, M.G., Lima, P.F., Mårtensson, J., Bemporad, A., Wahlberg, B.: Trajectory planning under vehicle dimension constraints using sequential linear programming. In: IEEE Conference on Intelligent Transportation Systems, pp. 108–113 (2017)
37. Pomerleau, D.A.: ALVINN: an autonomous land vehicle in a neural network. In: Advances in Neural Information Processing Systems, pp. 305–313 (1989)
38. Rajamani, R.: *Vehicle Dynamics and Control*. Springer Science & Business Media (2011)
39. Randlov, J., Alstrom, P.: Learning to drive a bicycle using reinforcement learning and shaping. In: International Conference on Machine Learning, pp. 463–471 (1998)
40. Salimans, T., Ho, J., Chen, X., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. [arXiv:1703.03864](https://arxiv.org/abs/1703.03864) (2017)
41. Schouwenaars, T., Mettler, B., Feron, E., How, J.P.: Robust motion planning using a maneuver automation with built-in uncertainties. *IEEE Am. Control. Conf.* **3**, 2211–2216 (2003)
42. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. [arXiv:1707.06347](https://arxiv.org/abs/1707.06347) (2017)
43. Siegelmann, H.T., Sontag, E.D.: Turing computability with neural nets. *Appl. Math. Lett.* **4**(6), 77–80 (1991)

44. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M.: Deterministic policy gradient algorithms. In: International Conference on Machine Learning, pp. 387–395 (2014)
45. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, vol. 1. MIT press Cambridge (1998)
46. Sutton, R.S., McAllester, D.A., Singh, S.P., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems, pp. 1057–1063 (2000)
47. Tedrake, R., Manchester, I.R., Tobenkin, M., Roberts, J.W.: LQR-trees: feedback motion planning via sums-of-squares verification. *Int. J. Robot. Res.* **29**(8), 1038–1052 (2010)
48. Urmson, C., Anhalt, J., Bagnell, D., Baker, C., Bittner, R., Clark, M.N., Dolan, J., et al.: Autonomous driving in urban environments: boss and the urban challenge. *J. Field Robot.* **25**(8), 425–466 (2008)
49. Wierstra, D., Schaul, T., Glasmachers, T., Sun, Y., Peters, J., Schmidhuber, J.: Natural evolution strategies. *J. Mach. Learn. Res.* **15**(1), 949–980 (2014)
50. Xu, H., Gao, Y., Yu, F., Darrell, T.: End-to-end learning of driving models from large-scale video datasets. [arXiv:1612.01079](https://arxiv.org/abs/1612.01079) (2016)
51. Xu, J., Nelson, B.L., Hong, J.: Industrial strength COMPASS: a comprehensive algorithm and software for optimization via simulation. *ACM Trans. Model. Comput. Simul.* **20**(1), 3 (2010)



Information Augmentation, Reduction and Compression for Interpreting Multi-layered Neural Networks

Ryotaro Kamimura^(✉)

Tokai University, Tokyo, Japan
ryo@keyaki.cc.u-tokai.ac.jp

Abstract. The present paper aims to propose a new type of learning method for interpreting relations between inputs and outputs in multi-layered neural networks. The method is composed of information augmentation, reduction and compression component. In the information augmentation component, information in inputs is forced to increase for the subsequent learning to choose appropriate information among many options. In the information reduction component, information is reduced by selectively choosing strong and active connection weights. Finally, in the information compression component, information contained in multi-layered neural networks is compressed by multiplying all connection weights in all layers for summarizing the main characteristics of connection weights. The method was applied to the improvement of an EC (electric commerce) web site for better profitability. The method could clarify relations between inputs and outputs and its interpretation was more natural than that by the conventional logistic regression analysis. The results suggest that multi-layered neural networks can be used to improve generalization and in addition to interpret final results, which is more important in many applications fields.

Keywords: Information augmentation · Reduction · Compression · Generalization · Interpretation · Multi-layered neural networks

1 Introduction

Because neural networks have the strong ability and complexity to model inputs, it is important to limit their capabilities, depending on data sets. Thus, there have been many different kinds of methods to reduce this complexity to the level of real data sets. For example, variable reduction and weight decay have been well known and frequently used for capability and complexity reduction in neural networks as well as machine learning.

First, the variable reduction method is a popular complexity reduction technique for better generalization by eliminating peripheral information in inputs [1–3]. In addition, variable selection can be used to reduce computational time and to interpret final results through a small number of obtained features.

Recently, pattern augmentation has received considerable attention influenced by recently developed generative models [4–8], which can be considered as a kind of variable reduction [9]. Though variable reduction has played important roles in improving generalization and interpretation in machine learning, one of the major problems lies in difficulty in determining the importance of variables. For example, the importance of input variables cannot be determined by examining individual variables but the combination of those variables should be considered to determine the importance [1]. In addition, there are some cases where it is difficult to obtain and create many different kinds of input variables in actual experiments. This is because the characteristics of inputs cannot be easily determined to represent them in terms of concrete input variables. Naturally, in these cases, the variable selection is of no use, but it is necessary to create as many input variables as possible to explain input patterns.

To cope with this kind of problem concerning the variable selection, we propose an inverse operation, namely, variable augmentation or dimensionality augmentation. In this method, we first increase the number of variables and eventually abundant and redundant information is generated. This can produce many different kinds of similar variables which are slightly different from each other. Thus, the method can increase the number of options, which can be chosen for the subsequent supervised learning. Consequently, supervised learning can more easily choose necessary and useful variables among those redundant options, if appropriate methods to choose them exist.

As mentioned, information augmentation is useful only if strong and efficient information selection methods exist. However, conventional complexity reduction methods such as weight decay are not well suited for choosing important ones among many because of their passive property for the choice. Weight decay and its related techniques have been used to inhibit the complexity [10, 11]. One of the problems is that those types of complexity reduction methods are passive in extracting important features to be used for training. This is because the weight decay is used to inhibit connection weights as much as possible, hoping that important ones emerge spontaneously, but it is not always true to have important ones, because all weights are equally treated. To overcome this passive approach to complexity reduction, we here propose information reduction based on selective information maximization with cost reduction. The cost reduction is similar to the conventional weight decay where all connection weights should be as small as possible. On the other hand, selective information maximization aims to choose important ones selectively based on predetermined criteria. Thus, the method is active in terms of choosing important ones by the predetermined criteria. Combining this active complexity reduction method of selective information and cost reduction with information augmentation, the present method can extract important features with the smallest cost in terms of the strength of weights.

The paper is organized as follows. In Sect. 2, we present how to augment and reduce information by using a simple neural network. Then, how to compress multi-layered neural networks is given, which is realized by multiplying all con-

nexion weights in all layers. In Sect 3, we present experimental results on the improvement of EC (electronic commerce) web site to improve the profitability. We explain how selective information can be increased, while decreasing the corresponding cost. Then, the interpretation of compressed weights on relations between inputs and outputs is more natural than that by the logistic regression analysis.

2 Theory and Computational Methods

2.1 Information Augmentation

In the present model, we have three components: information augmentation, reduction and compression. In the first information augmentation component, information, contained in inputs, is forced to increase. This is because multi-layered neural networks tend to naturally lose its original information by going through many hidden layers. To prevent neural networks from losing its information, it is necessary to increase information in inputs as much as possible. To increase information, we have two options. The first one is to increase input patterns as much as possible. The second one is to increase the inputs or variables. In real experiments, sometimes it is very hard to obtain a sufficient number of input patterns and limited number of patterns are only available. In addition, the number of input variables is also limited in real experiments, because difficulty exists to define input variables to consider all aspects of input patterns. In this context, we try to increase the input variables seemingly by unsupervised learning. By expanding limited number of input variables into many different types of neurons, some useful information concerning subtle change in input variables and the effect of combination of input variables can be extracted.

As shown in Fig. 1, connection weights between the first two hidden layers, are updated by auto-encoders. Usually, the auto-encoders are used to compress information by reducing the number of neurons. However, the auto-encoders are here used to increase neurons and to generate redundant information as much as possible. In this stage, neurons can be increased to extend original information over many neurons. This process is considered to be a kind of dimensionality augmentation. This method is particularly useful, when it is impossible to obtain many different kinds of input variables, meaning that it is hard to obtain necessary information from input patterns.

This information augmentation is simple enough to be implemented. In Fig. 1a, connection weights between inputs to the first hidden layers $w_{j_1 j_0}$ are updated by using auto-encoders and by increasing the number of hidden neurons in the auto-encoders. This processes is repeatedly applied by increasing the number of hidden layers. In Fig. 1, the number of hidden layers was limited to two in the information augmentation component, because it has been impossible to increase generalization performance even if the number of hidden neurons is forced to be increased beyond two hidden layers.

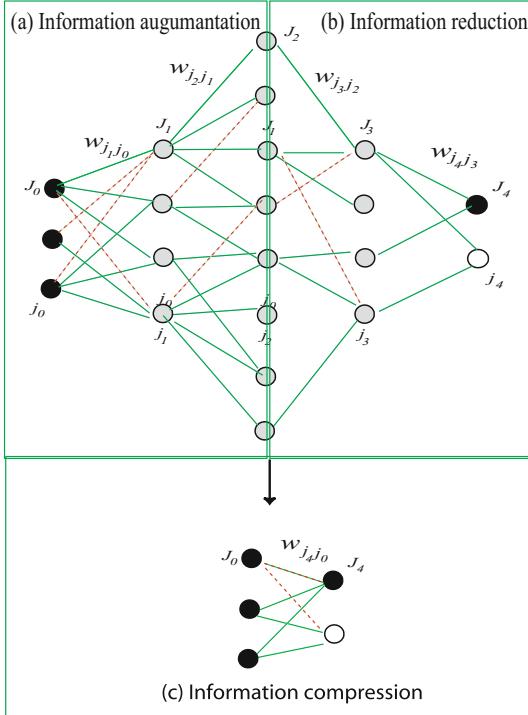


Fig. 1. A network architecture with two-stage information augmentation (a) and information reduction component (b) and information compression (b).

2.2 Information Reduction

The information reduction component uses the ordinary supervised learning with BP and many hidden layers in Fig. 1b. To reduce information or the strength of connection weights, we define selective information. The selective information is defined as difference between maximum number of connection weights and the number of active and strong weights. When the number of active and strong weights becomes smaller, the selective information increases.

For connection weights between the first and second hidden layer $w_{j_2j_1}$ ($j_1 = 1, 2, \dots, J_1$; $j_2 = 1, 2, \dots, J_2$) in Fig. 1a, we define the degree of active and strong weights by using the absolute weights

$$u_{j_2j_1} = |w_{j_2j_1}| \quad (1)$$

Then, the number of normalized active and strong connection weights is defined with respect to maximum connection weight

$$Z_{21} = \sum_{j_2=1}^{J_2} \sum_{j_1=1}^{J_1} \frac{u_{j_2j_1}}{u_{\max}} \quad (2)$$

where the maximum operation is over all connection weights between two layers. When no connection weights between layers exist, it is impossible to transmit information between layers, it is supposed that at least one connection weight between two layers should have a value greater than one

$$Z \geq 1 \quad (3)$$

This Z 's maximum value is computed by $J_2 J_1$, where all connection weights are supposed to be equal. Thus, selective information can be defined by

$$SI_{21} = J_2 J_1 - \sum_{j_2=1}^{J_2} \sum_{j_1=1}^{J_1} \frac{u_{j_2 j_1}}{u_{\max}} \quad (4)$$

Then, the corresponding cost is computed by the sum of all absolute weights

$$C_{21} = \sum_{j_2=1}^{J_2} \sum_{j_1=1}^{J_1} u_{j_2 j_1} \quad (5)$$

Then, this method tries to maximize selective information and at the same time to minimize the corresponding cost. This can be applied by changing the number of weighted active and strong connection weights.

2.3 Information Compression

Finally, we should explain a concept of information compression. The information compression is performed by multiplying all connection weights in all layers. First, connection weights from the third hidden layer to the fourth (output) layer $w_{j_4 j_3}$ and connection weights from the second hidden layer to the third hidden layer $w_{j_3 j_2}$ are compressed into new weights $w_{j_4 j_2}$

$$w_{j_4 j_2} = \sum_{j_3=1}^{J_3} w_{j_4 j_3} w_{j_3 j_2} \quad (6)$$

The same procedures are repeatedly applied and the final compressed weights

$$w_{j_4 j_0} = \sum_{j_1=1}^{J_1} w_{j_4 j_1} w_{j_1 j_0} \quad (7)$$

3 Results and Discussion

3.1 Improvement of EC Web Site by Analyzing Customers' Access Logs

Experimental The experiment aimed to analyze an EC web site of mail-order company and to determine which items in the EC site should be modified to

increase the profitability. The number of customers was 5000, which was evenly divided into ones with high and low profitability. The number of input variables was eight in Fig. 2. The number of hidden layers in the information augmentation component was two, and in the information reduction component, the number of hidden layers increased from one to five. In the information compression component, all connection weights were multiplied by the corresponding weights to have a network without hidden layers in Fig. 2c. The seventy percent of the data set was for training and the remainder was divided evenly into validation and testing data sets. Generalization errors were taken with minimum validation errors. In an analysis by the logistic regression analysis [12], the special site, for example, concerning seasonal sales, negative effects to decrease the profitability. The present experiment aimed to see whether this finding on the special site, could be re-found by the present method.

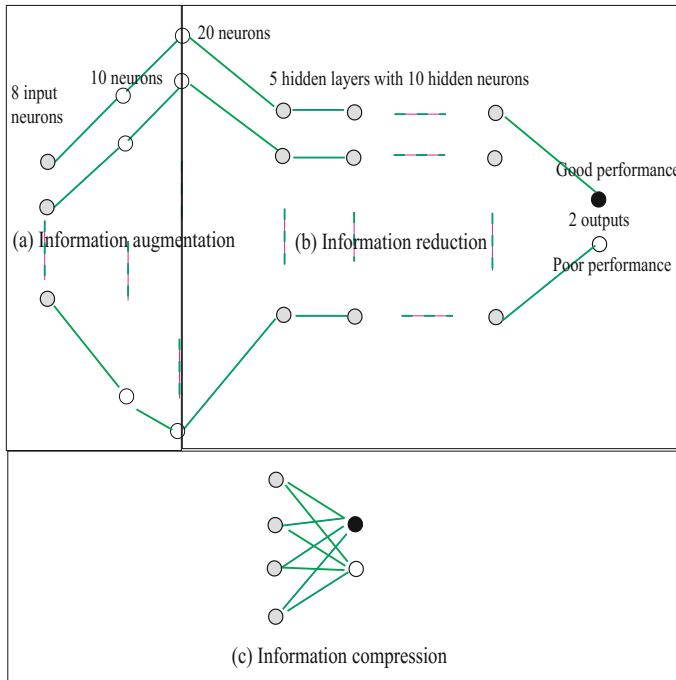


Fig. 2. Information augmentation (a), reduction (b) and compression (c) for the EC web site data set.

Information Augmentation and Generalization Performance First, we should examine to what extent information augmentation contributed to improved generalization. Table 1 shows generalization performance by the present method as well as three other methods. The best errors (0.3524, 0.3667)

were obtained by the present method in terms of average and maximum values where the number of neurons in the first and second hidden layers were 45 and 90, respectively. Thus, to have the best values, the number of neurons should be excessively increased. One interesting fact is that the minimum value (0.3200) was obtained by the conventional multi-layered neural networks with 15 hidden neurons with five hidden layers. Thus, this may imply that the excessive increase in information or neurons tends to decrease the variance of final generalization errors. Actually, the present method could produce generalization errors with the smallest standard deviation (0.0105). The third best average error (0.3750) was by the logistic regression analysis. Finally, the bagging ensemble method [13, 14] produced the worst average error (0.3750). These results show that the information augmentation could decrease generalization errors by increasing neurons and by stabilizing learning processes.

Table 1. Summary of experimental results on generalization performance for the EC site data set. Bold-Face values show the best values

Methods	Neurons	Layers	Avg	Std dev	Min	Max
Cost	45 90	5	0.3524	0.0105	0.3307	0.3667
BP	15	5	0.3619	0.0220	0.3200	0.4000
Bag			0.4059	0.0155	0.3827	0.4267
Logistic			0.3750	0.0170	0.3528	0.4032

Information Reduction In the information reduction component, several important weights were selectively chosen and at the same time, the cost or the strength of connection weights were forced to be smaller. Figure 3 shows selective information and its corresponding cost. As shown in Fig. 3a, selective information increased gradually when the number of learning steps increased and close to 0.65 in the end. Note that the selective information was normalized to take a value from 0 to 1. On the other hand, the cost, computed by the sum of absolute weights, decreased rapidly and close to zero only with 20 steps. The results show that selective information can be increased, reducing the corresponding cost.

Then, we should show how connection weights were changed by the selective information maximization and cost minimization. Figure 4 shows connection weights from the third hidden layer to the fourth hidden layer, though only partially shown. With the first step, many strong positive (green) and negative (red) weights could be seen in Fig. 4a, because the selective information was small and the cost was high. When the number of steps increased from ten in Fig. 4b to 100 steps in Fig. 4e, and the selective information increased and the corresponding cost became smaller, the number of strong connection gradually decreased and only connection weights to the 11th hidden layer were strong

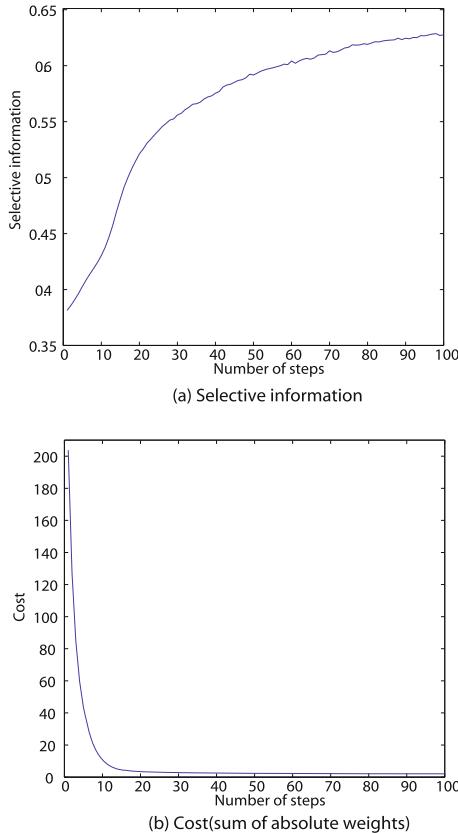


Fig. 3. Selective information (a) and the corresponding cost (sum of absolute weights) (b) as a function of the number of steps.

enough, while all the other connection became quite small. The results confirmed that the present method could reduce the number of strong weights and only a small number of important weights remained in the end.

Information Compression We examine here why and how information augmentation could contributed to improved generalization by compressed weights, namely, relations between inputs and outputs. Figure 5 shows compressed weights, when the number of neurons in the first hidden layer increased from 10 (a) to 45 (d) and correspondingly the number of neurons in the second layer increased from 20 (a) to 90 (d). As above mentioned, 45 and 90 hidden neurons in the first and second hidden layers produced the best generalization performance. When the number of neurons was ten in Fig. 5a, only compressed weights from the fourth input (special site) became negatively stronger. This means that the special site in EC web site decreased the profitability of the company, as already

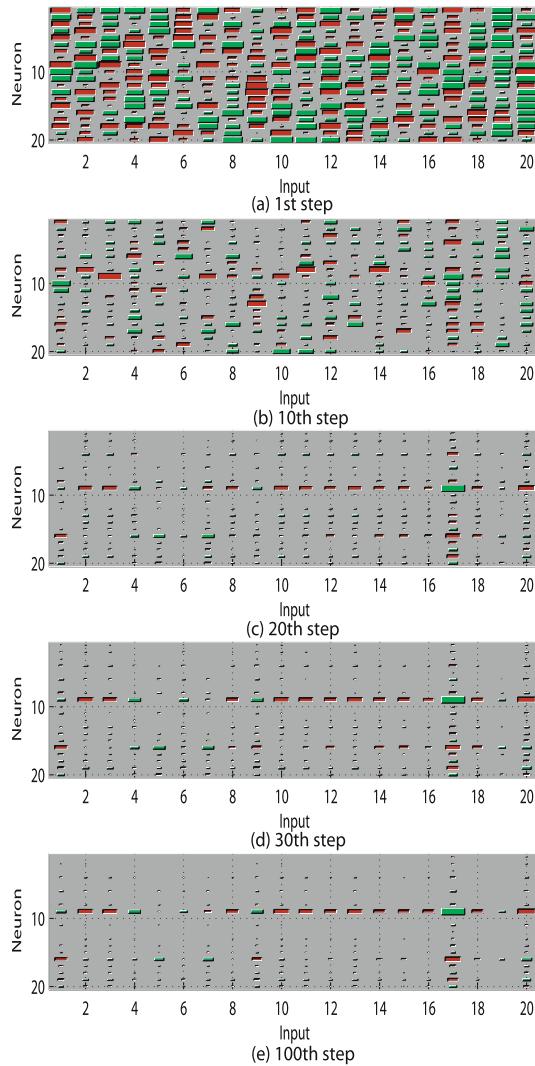


Fig. 4. Connection weights from the first hidden layer to the second hidden layer from the first step to the final 100th step for the EC site data set. Only the first 20 connection weights were shown for simplification.

pointed out in [12]. When the number of neurons increased from 20 in Fig. 5b to 45 with the best generalization performance in Fig. 5d, the other inputs, for example No. 6, 5, 3 and 8 tended to have relatively larger absolute values. This means that when the number of hidden neurons in the information augmentation component was small, only the most important input variable was taken into account to produce the outputs. This is because the information capacity of neural networks was small for accepting all information in inputs. Then, when the

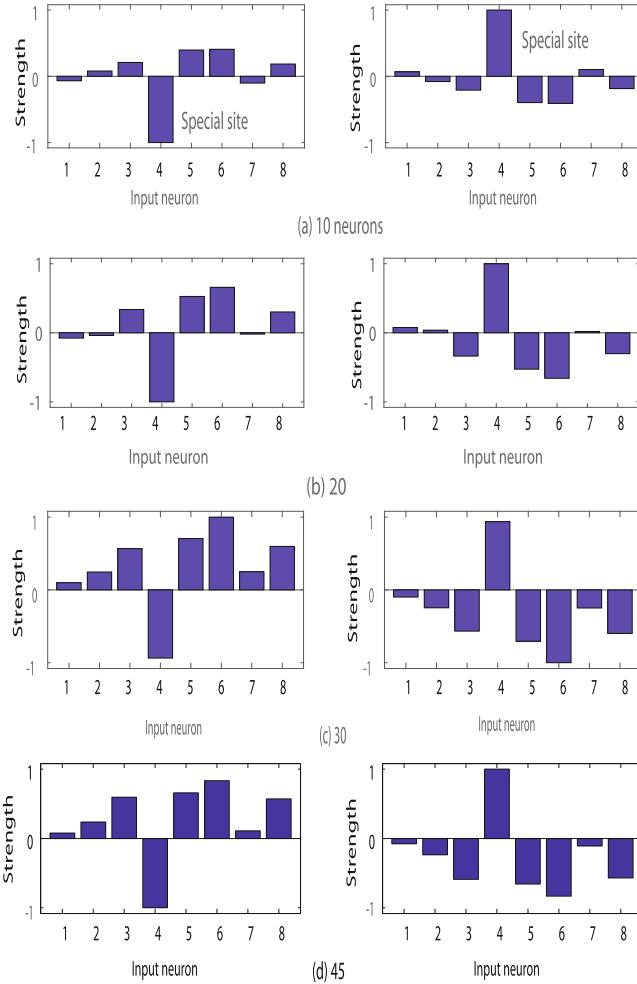


Fig. 5. Compressed weights into the first and the second output neuron when the number of neurons in the first hidden layer increased from 10 (a) to 45 (d) for the EC site data set.

number of neurons gradually increased, redundant information increased, neural networks have more capability to accept redundant information, represented by the variable No. 6, 5, 3 and 8. These relatively important input variables certainly contributed to improved generalization performance.

Then, we examined to what extent compressed weights were similar to or different from other well-known measures obtained by the bagging ensemble method, logistic regression analysis and simple correlation coefficient between inputs and outputs. Figure 6a–c show the predictor importance by the bagging ensemble method, regression coefficients and correlation coefficients between

inputs and outputs for the EC site data set. As shown in the figures, the predictor important values in Fig. 6a were different from the compressed weights by the present method. This is because the predictor importance cannot describe the negative capability of input variables. However, suppose that the variable No. 4 had the negative effect, the predictor importance became similar to compressed weights by the present method, though the variable No. 4 had less importance. On the other hand, the regression coefficients by the logistic regression analysis in Fig. 6b were similar to the compressed weights into the first output neurons in which the variable No. 4 had the largest negative effect for the output. However, the other input variables, for example, variable No. 2 and 8, showed different values. Finally, we computed correlation coefficients between inputs and outputs. As can be seen in the figure in Fig. 6c, the coefficients were more similar to the compressed weights than any other methods, and only one difference is the variable No. 1. Thus, the compressed weights represented the degree of correlation coefficients between inputs and outputs. The neural networks with five hidden layers, could show clearly relations between inputs and outputs. These results with compressed weights shows a possibility that multi-layered neural networks can improve generalization performance, while clarifying relations between inputs and outputs.

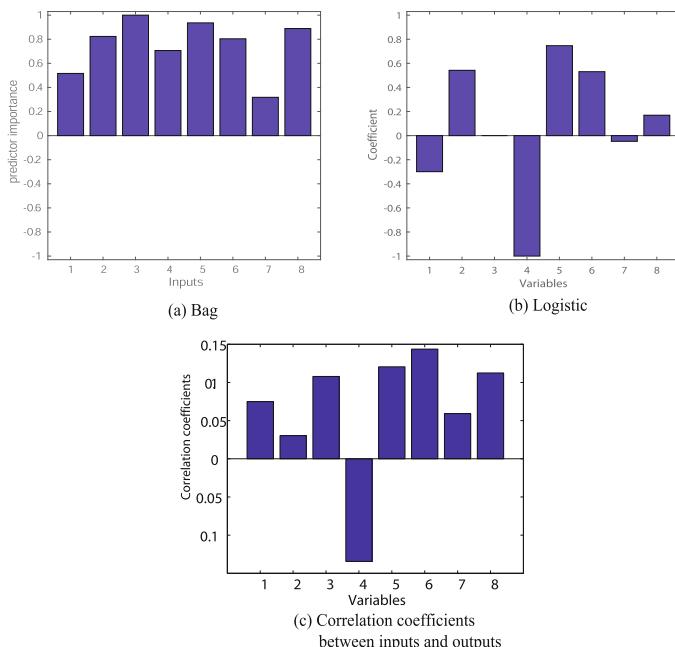


Fig. 6. Predictor importance by the bagging method (a), regression coefficients by the logistic regression analysis (b) and correlation coefficients between inputs and outputs (c) for the EC site data set.

4 Conclusion

The present paper tried to introduce information augmentation, reduction and compression component in training neural networks to improve generalization and to interpret relations between inputs and outputs. In the information augmentation component, the number of inputs or input variables is increased by increasing the number of neurons in hidden layers. The component tries to produce redundant information for the subsequent supervised learning to choose appropriate information easily. Neural networks have so far paid much attention to dimensionality reduction to adjust network complexity, but the present paper stresses that dimensionality should be increased to extract information covering all aspects of input patterns. However, for this information augmentation to be realized, we need a method to choose important information actively and efficiently. In the information reduction component, important information is selectively chosen, based on the strength of absolute weights and at the same time the corresponding cost is minimized. In the information compression component, information in neural networks is compressed by multiplying all connection weights as much as possible. The method was applied to the improvement of EC web site to increase the profitability. The results confirmed that selective information could be increased and the corresponding cost could be decreased. This means that the majority of connection weights are reduced to smaller values. Then, by compressing all connection weights, we could interpret naturally relations between inputs and outputs. The compressed weights showed that the information augmentation could extract more detailed information on relations between inputs and outputs. On the other hand, when the augmentation became weaker or the number of neurons became smaller, only a smaller number of really important variables were taken into account. Thus, some detailed or redundant information may contribute to improved generalization.

The problem is how to determine the importance of weights. In the present paper, the absolute weights were used for the degree of importance. However, more appropriate criteria are necessary for this evaluation. Though some problems should be solved, the present paper shows that the method here proposed can be used to improve generalization and interpretation in neural networks by simple information augmentation.

References

1. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3**, 1157–1182 (2003)
2. Rakotomamonjy, A.: Variable selection using SVM-based criteria. *J. Mach. Learn. Res.* **3**, 1357–1370 (2003)
3. Perkins, S., Lacker, K., Theiler, J.: Grafting: fast, incremental feature selection by gradient descent in function space. *J. Mach. Learn. Res.* **3**, 1333–1356 (2003)
4. Wang, J., Perez, L.: The Effectiveness of Data Augmentation in Image Classification Using Deep Learning. Technical report (2017)

5. Asperti, A., Mastronardo, C.: The effectiveness of data augmentation for detection of gastrointestinal diseases from endoscopical images. [arXiv:1712.03689](https://arxiv.org/abs/1712.03689) (2017)
6. Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., Jin, Z.: Improved relation classification by deep recurrent neural networks with data augmentation. [arXiv:1601.03651](https://arxiv.org/abs/1601.03651) (2016)
7. Marchesi, M.: Megapixel Size Image Creation Using Generative Adversarial Networks (2017)
8. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes (2013)
9. Allen, D.M.: The relationship between variable selection and data agumentation and a method for prediction. *Technometrics* **16**(1), 125–127 (1974)
10. Moody, J., Hanson, S., Krogh, A., Hertz, J.A.: A simple weight decay can improve generalization. *Adv. Neural Inf. Process. Syst.* **4**, 950–957 (1995)
11. Hinton, G.E.: A practical guide to training restricted boltzmann machines. In: *Neural Networks: Tricks of the Trade*. Springer, Berlin, pp. 599–619 (2012)
12. Nishiuchi, H.: Statistical Analysis for Billion People (in Japanese). Nikkei BP marketing, Tokyo, Japan (2014)
13. Breiman, L.: Bagging predictors. *Mach. Learn.* **24**(2), 123–140 (1996)
14. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)



Enhance Rating Algorithm for Restaurants

Jeshreen Balraj^(✉) and Cassim Farook

Informatics Institute of Technology, Colombo, Sri Lanka
jeshreen18995@gmail.com, cassim.f@iit.ac.lk

Abstract. More and more people make their purchase decisions by referring to reviews and ratings provided in online platforms. Visitors to restaurants use online reviews on a larger scale compared with the users of other industries. However, for these visitors, evaluating numerous reviews is a hassle and time consuming as it involves a process of reading through all the reviews, identifying the date of review posting, understanding the reviewer's credibility and identifying the rating of the reviewer and the restaurant before making the decision. This research proposes an enhanced rating algorithm which will calculate an overall rating. Apart from the standard point rating the solution will include the aspect, sentiment, time factor and user credibility of a review. The enhanced algorithm uses Natural Language Processing and Sentiment analysis used with machine learning to identify the thoughts of the user regarding the restaurants. The algorithm is tested with a web-based solution that gives an overall idea of the current performance of a particular restaurant utilizing reviews of those restaurants. The new algorithm gives a much credible rating than the conventional rating systems.

Keywords: Natural language processing · Sentiment analysis · Algorithm · Machine learning

1 Introduction

Purchase decision of many users majorly rely on online sources. The users of restaurant industry stand on the rise when compared with other industries as per the research conducted by BrightLocal [1]. The online sources contain two components as the reviews, the textual content of a formal assessment of an object or a place and the rating which is the rank given to a place or object through its numerical value based on a comparative assessment of the quality making. These two components are used as the measuring facts of restaurants by the users [2]. Among these two, rating is considered highly by the users to make a purchase decision. The confidence to make this purchase decision is gained when the rating for the restaurant is beyond the global average of 3.5 according to the survey carried out by BrightLocal among customers who read reviews before making a purchase decision.

Various factors are analysed by different users while making their purchase decision. A survey conducted by BrightLocal shows that 58% of the users first analyze the overall rating given for the restaurant which gives an idea within seconds as to how the restaurant would be. On the other hand, 47% of the customers checks the sentiment of the reviews provided by the reviewers. This requires a lot of time and critical evaluation

to identify the positive and negative comments provided by the reviewers regarding the specific restaurant. The third factor analyzed by another 41% of customers is the recency of the review. This factor is analyzed to know the current status of the restaurant which is not reflected by the overall star rating displayed against the restaurant. The fourth factor identified by the user is the content alignment between the review and the rating given by the reviewers. The inconsistency gets visible when the rating given does not match with the review provided by the user [3]. Other factors considered by other users are the quality of the review, length of the review and the response rate of the business on the reviews provided.

The reviews that are displayed on the reviewing web sites are thoroughly analyzed by the customers before making the purchase decision. In the process of analyzing customers look into these factors [4]. There are:

- a. Perceived Informativeness—Identifying whether the information displayed is relevant for the reader to make his/her buying decision.
- b. Perceived persuasiveness—Identifying how convincing the review is for the reader.
- c. Source Credibility—analyzing the credibility of the user who has posted the review [5].
- d. Perceived quantity of reviews—analyzing the number of reviews a certain restaurant has received.

The above-mentioned factors do have an effect on the users purchase decision. A sentiment analysis and aspect analysis on all the reviews will provide an idea of the business as a whole. These key factors must be used to analyze a review as it brings out the best meaning from a review.

The rating calculation displayed on the online restaurant rating and reviews site are calculated in different methods such as, Difference value, Average rating, [6] Simple Average System, Lower bound of Confidence Intervals [7] and Bayesian average [8]. These traditional calculation methods fail to cover few of the major features that the user needs to see through the final rating they are, reflection on the time of the review written, analysis on the reviewer's behavior, aspect coverage and sentiment coverage of the review.

The proposed algorithm will cover the drawbacks of the current rating system and include the factors that the user needs in a rating system keeping the Bayesian algorithm as the base. This will provide users a much credible rating which will give them a complete idea of the restaurant at a glance.

This research paper will first carry details regarding the dataset use. The methodology used, testing approach, prevailing products, conclusion of this research and future enhancements of this proposed system is discussed in the given order.

2 Data Set

For this research reviews and ratings of fifteen restaurants around Colombo has been gathered through Zomato. Around hundred reviews were gathered for each of the restaurants.

3 Methodology

The enhance rating algorithm has been divided into different components to carryout different tasks. As displayed in Fig. 1, the aspects were extracted from the given reviews which was be used in the aspect analysis. The completion of aspect analysis triggered the sentiment analysis module to identify the sentiment of the review and score it. Next the user credibility module was triggered and then the reviews were given a score according to the recency. Scoring all the reviews for the restaurant based on these factors resulted in the per user rating calculation module to be executed where the review was given a score based on the scores given. To calculate the final restaurant rating, the newly calculated per user rating was used instead of the user given rating. finally, the rating for the restaurant was calculated.

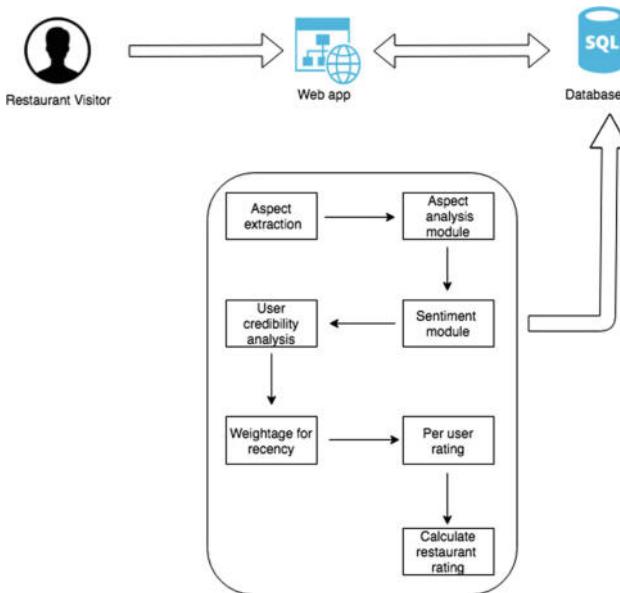


Fig. 1. High level design

3.1 Aspect Extraction Module

The aspects were identified with the highest frequent nouns referred by the reviewers, the reviews of all the restaurants were word tokenized. Once the words were tokenized, they were passed through the POS tagger to identify tokens of the words as Noun, verb, adjective etc. after the tags of the words were identified, the Nouns—singular (NN), Nouns—plural (NNS), proper Noun—singular (NNP) and proper Noun—plural (NNPS) were extracted [9]. The extracted words were lemmatized to identify the base word. After lemmatizing the stop words were removed and frequency of the words

were identified to understand the aspects expected by people and the most frequent ten aspects were chosen as shown in Fig. 2.

```
[('place', 831), ('food', 655), ('service', 354), ('time', 269), ('coffee', 251),
('chicken', 228), ('staff', 200), ('friend', 197), ('price', 194), ('menu', 170)]
```

Fig. 2. Extracted aspects

Out of these ten words the food names were removed using the food names lexicon present in wordnet to identify the important aspects expect by the people and the rest of the aspects were used for the aspect analysis.

Using this method to extract the aspects reduces the errors in identifying the aspects needed by the users.

3.2 Aspect Analysis and Scoring

The aspect analysis identifies whether each individual review has any of the aspects expected by the users contained within it. To carry out the aspect analysis the Stanford CoreNLP Dependency parser [10] was used to extract the dependent word with the relevant aspect word. Out of the many dependencies available in the Stanford Core NLP library, the *amod*—Adjectival modifier and the *nsubj*—Nominal subject were used. As it extracts the Noun Phrase present in the sentence along with the adjectival phrase that brings out the meaning of the Noun Phrase which are highlighted in the red box in the above image. The aspects were passed as a list and the presence of the aspect word was analysed in a sentence and the relevant adjective of the noun was selected as the opinion word. These opinion words were passed through SentiWordNet [11] to get the sentiment score. A score between zero to one will be given to the words as Positive, Negative and objective which brings the total to one [12] this score was multiplied by five for calculation purpose. Figure 3 displays the score given for the opinion words which were extracted from a review.

```
i love this place.. would go anytime when i get the chance... love their bur  
gers.. for me they have the best burgers i have had! pricing is alright. and  
the place is so great and relaxing.  
('alright', 'JJ') nsubj ('pricing', 'NN') 3.125  
('great', 'JJ') nsubj ('place', 'NN') 5.0
```

Fig. 3. Aspect score for a review

3.3 Sentiment Scoring Module

SVM is one of the most efficient text classification technique used by many. The most important principle of SVM is to determine linear separators in the search space which can best separate the different classes [13]. SVM classifiers can performs well with big datasets and will achieve higher accuracy level. This is the main reason why SVM is

widely used in advanced classifiers. Based on the results of different research papers, support vector machine was chosen to identify the sentiment of the review. Table 1 shows an analysis of different research papers among three main binary classifications Naïve Bayes (NB), Support Vector Machine (SVM) and Maximum Entropy (ME). Based on the results of this analysis, SVM was chosen to classify the sentiments of a review in this system.

Table 1. Comparison of different classification techniques

Research paper	Data set used	NB (%)	SVM (%)	ME
[14] Average of all given features	Movie reviews	79.81	80.38	79.68
[15] Average of given results	Sports reviews, scientific electronic devices reviews, computer reviews	95.83	99	94.96%
[16]	News groups	69.3	86.8	–
[17]	Travel reviews	80.71	85.14	–
[18]	Palm readings	85.79	81.01	–

The sentiment scoring module identifies the sentiment of each review and score them accordingly. A score closer to 1 depicts that the review is carrying a negative sentiment and a score of 0 depicts that the review is carrying a positive sentiment. To score the reviews a support vector machine was built. The testing data was manually labelled as positive and negative to train the model to classify reviews accordingly. To build this SVM model scikit-learn [19] was used.

3.4 User Credibility Scoring Module

The credibility of the users is identified through the number of reviews provided by the user. A user who has provided only one review appears to be less credible than a user who has provided fifty reviews. To identify this, a support vector machine was used to classify the users as credible or less credible. The reviews were labelled as positive considering the number of reviews over 20 and negative when the number is below 20. The Support vector machine was able to learn the training data set and provided a score between 0 and 1 for each user. The score provided by this process depicted 1 as negative and 0 as positive where the score was closer to 0 it depicted that the user credible and when the score was closer to the value 1 it depicted that the user is not credible enough. The score such obtained is the User Credibility Score.

3.5 User Score Calculation

The content alignment of each review is identified by the difference between the sentiment score given for the review and rating given by the user for a particular place. This is calculated to identify whether the user has given a consistent rating according to the review provided.

To calculate the score for the user profile rating, the User Credibility Score (UCS) and the Content Alignment Score (CAS) was used. This score will provide an understanding of the reviewer to the user. Below given is the equation used to calculate the User Profile Rating (UPR):

$$UPR = \frac{CAS}{\text{No.of content aligned reviews}} + UCS$$

3.6 Algorithm Implementation

The algorithm implementation of this paper is divided into two categories. As the calculation of per user rating and the calculation of the overall restaurant rating. based on the aspect score, sentiment score, time score and the user credibility score the user given rating for the restaurant will be re-calculated. Once the new user given rating is identified, the Bayesian average algorithm will be used to calculate the overall restaurant rating. The advantage of using Bayesian average calculation is that, it considers the score of a given restaurant as a weighted aggregation from the contribution of the given restaurant ratings and the contribution of ratings from all restaurants. In this the algorithm will give importance to the own restaurant rating and will shift the score to the global average. When the ratings increase the real rating will be displayed moving from the global average as depicted in Figs. 4 and 5. The Bayesian average maintains a fair basis of ratings for the restaurants not considering the number of ratings given [8] The drawback faced in Bayesian average is not providing a weightage to the recency of the review, sentiment, aspect and the user credibility of the review. These drawbacks are addressed in the algorithm while calculating the per user rating.



Fig. 4. Bayesian average calculation [7]



Fig. 5. Growth of rating in Bayesian average [7]

Per User Rating

To calculate the per user rating, the aspect score, sentiment score, user credibility score will be used. Apart for these three scores the recency decay of the review will be identified and scored. The recency of a review is measured comparing the current date with the date that the review was given by the user. Each review will have a time score in comparison with the date of review. Table 2 shows the scores given and different conditions. The score given to a review will depreciate with time.

Table 2. Recency decay conditions and scores

Start date	Time gap	Time score
Current date	1 month	5
Current date	3 months	4
Current date	6 months	3
Current date	12 months	2
Current date	1 year and beyond	1

The aspect scores generated by the system will use the following equation to keep it within the range of 5 to make the overall user rating calculation simple.

$$\text{Aspect score} = \frac{\text{per aspect score} * 5}{\text{No.of aspects}}$$

User credibility score and sentiment score calculated by the system will also be converted to a value within 5 due to the reason of the negative is depicted by 1 and positive by 0. The following equation will be used based on the range—Table 3.

Table 3. Sentiment/user credibility range conditions

Condition	Range
Per sentiment or credibility score > 0.2	4.5
Per sentiment or credibility score > 0.5	3.5
Per sentiment or credibility score > 0.7	3
Per sentiment or credibility score > 1	2.5

S.S *Sentiment Score*

U.C.S *User Credibility Score*

$$S.S \text{ or } U.C.S = Range - Per S.S \text{ or } Per U.C.S$$

With all these scores converted to a range of 5, the final Per user rating was calculated based on the equation given below.

SS *Sentiment Score*

AS *Aspect Score*

UCS *User Credibility Score*

TS *Time Score*

$$Per User Rating = \frac{SS + AS + UCS + TS}{4}$$

Overall Restaurant Rating

To calculate the overall restaurant rating, the newly calculated per user rating will be used instead of the rating given by the user because the newly calculated rating gives a clear understanding regarding the sentiment, aspect, recency decay and user credibility of the review. The equation given below is used to calculate the final restaurant rating.

R *Avg. of user ratings for the restaurant.*

C *Avg. of user rating for all the restaurants.*

V *Number of ratings for the restaurant*

M *Avg. number of reviews for all the restaurants*

$$W = \frac{V}{V + M}$$

$$Bayesian average = W * R + (1 - W) * C$$

4 Testing

To test the performance and accuracy of the individual components, the algorithm was broken down into different sub components as aspect analysis, sentiment analysis, user and credibility analysis to test the accuracy.

4.1 Aspect Analysis

The accuracy of aspect analysis was done manually with a set of 20 different reviews that were randomly selected. The system was able to identify 79.45% of the expected aspects in a sentence. The system also produced only a small variation of +11.07% or—11.07%. The accuracy of the system can be increased with the number of data used for testing.

4.2 Sentiment Analysis

The accuracy of the sentiment analysis module was tested with the generation of the confusion matrix. Accuracy rate of 87.78% was evident for this module. The F1 score for this Support Vector Machine module was 0.64.

4.3 User Credibility Analysis

The accuracy of the Support Vector Machine module to identify the user credibility was tested through the confusion matrix. An accuracy rate of 98.27% was identified for this module and the F1 score was 0.97.

Having the three components with a good accuracy rate, the entire system was tested to identify the accuracy of this enhanced rating algorithm. A manual testing method was carried out for this purpose. 15 different scenarios were selected and at a given moment 5 reviews were passed into the system to notice the fluctuation in the rating provided for the restaurant. The system was able to produce an accuracy rate of 80%. This proved the working of the system to be mostly accurate at a given situation. This algorithm was compared with the other restaurant reviewing sites, and it was more evident that the results from this enhanced algorithm was more reliable rather than the others, due to its uniqueness of including few of the needed components that are being ignored in the other systems as shown in Table 4.

Table 4. Comparison with the prevailing systems

System	Sentiment of review	Aspects of review	User credibility	Recency of review
Zomato	×	×	✓	✓
Yamu	×	×	×	×
Tripadvisor	×	✓	×	✓
Yelp	×	×	✓	✓
Recensione	✓	✓	✓	✓

5 Conclusion

This research identifies an enhanced rating algorithm built taking the following into account: the sentiment score, aspect score, user credibility score and time score of the review. Inclusion of these factors makes the final rating informative and reliable.

These factors are normally considered before making a purchase decision by the user who reads the review.

The key motive of this research is to provide customers with more reliable rating through which they can get an idea regarding a restaurant at a glance, as most of the users identify a good restaurant based on the final rating given for the restaurant. This proposed algorithm is beneficial for the users who frequently visit restaurant reviewing sites to make their buying decision. Further, reliability and accuracy can be increased with the proposed future enhancements.

6 Future Enhancement

Currently emoji are being used more by many people in the written reviews to express their opinion on a certain product or service. The prevailing system gives a scores for the sentiment of the review, aspects covered by the review and for the user credibility of the reviewer. Addition of a score for the sentiment expressed through the emoji will give more value to the restaurant because currently people do not read long reviews but gain an idea through the emojis displayed. Emoticon analysis is part of analysing the sentiment of the user given textual content, therefore this can be included as a part of the sentiment scoring module.

References

1. BrightLocal, “Local Consumer Review Survey 2016 | The Impact Of Online Reviews,” 2017. (Online). Available: <https://www.brightlocal.com/learn/local-consumer-review-survey/>. Accessed 05 Sep 2017
2. Dohse, K.A.: Fabrication feedback: blurring the line between brand management and bogus reviews. *J. Law Technol. Policy* **1**, 363–392 (2013)
3. Lee, I., Sun, Y., Li, Y.S.: An Intelligent Approach to Review Filtering and Review Quality Improvement, pp. 61–66 (2016)
4. J. Gobinath, J., Gupta, D.: Online reviews: determining the perceived quality of information. In: 2016 International Conference Advanced Computing Communication Informatics, pp. 412–416 (2016)
5. Bambauer-Sachse, S., Mangold, S.: Do consumers still believe what is said in online product reviews? A persuasion knowledge approach. *J. Retail. Consum. Serv.* **20**(4), 373–381 (2013)
6. Miller, E.: How Not To Sort By Average Rating—Evan Miller. *Evanmiller.org*, 2009. (Online). Available: <https://www.evanmiller.org/how-not-to-sort-by-average-rating.html>. Accessed 11 Feb 2018
7. EBC.: How to Rank (Restaurants) | ebc,” 2015. (Online). Available: <http://www.ebc.cat/2015/01/05/how-to-rank-restaurants/>. Accessed 05 Jul 2017
8. Rosairo Wenbert Del.: Getting the Bayesian Average for rankings (PHP/MySQL)|Ekini.net by Wenbert Del Rosario. *EKINI*, 2013. (Online). Available: <http://blog.ekini.net/2013/08/18/getting-the-bayesian-average-for-rankings-mysql/>. Accessed 13 Feb 2018
9. University of Pennsylvania.: Penn Treebank P.O.S. Tags. 2003. (Online). Available: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html. Accessed 05 Mar 2018
10. De Marneffe, M.-C., Manning, C.D.: Stanford typed dependencies manual (2008)

11. SentiWordNet.: SentiWordNet. 2010. (Online). Available: <http://sentiwordnet.isti.cnr.it/>. Accessed 05 Apr 2018
12. A. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining (2018)
13. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. *Ain Shams Eng. J.* **5**(4), 1093–1113 (2014)
14. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs Up? Sentiment Classification using Machine Learning Techniques, pp. 79–86
15. Shetty, J.: Sentiment Analysis of Product Reviews no. Icicct, pp. 298–303 (2017)
16. Hassan, S., Rafi, M., Shaikh, M.S.: Comparing SVM and Naïve Bayes classifiers for text categorization with Wikitology as knowledge enrichment. In: *Proceedings of 14th IEEE International Multitopic Conference 2011, INMIC 2011*, pp. 31–34 (2011)
17. Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, (2008)
18. Xue, Y., Chen, H., Jin, C., Sun, Z., Yao, X.: NBA-Palm: prediction of palmitoylation site implemented in Naïve Bayes algorithm. *BMC Bioinform.* **7**(1), 458 (2006)
19. Pedregosa, F., et al.: Scikit-learn: machine learning in Python Gaël Varoquaux. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)



Reverse Engineering Creativity into Interpretable Neural Networks

Marilena Oita^(✉)

The Swiss AI Lab IDSIA, SUPSI, USI, Lugano, Switzerland
marilena@idsia.ch

Abstract. In the field of AI the ultimate goal is to achieve generic intelligence, also called “true AI”, but which depends on the successful enablement of imagination and creativity in artificial agents. To address this problem, this paper presents a novel deep learning framework for creativity, called INNGenuity. Pursuing an interdisciplinary implementation of creativity conditions, INNGenuity aims at the resolution of the various flaws of current AI learning architectures, which stem from the opacity of their models. Inspired by the neuroanatomy of the brain during creative cognition, the proposed framework’s hybrid architecture blends both symbolic and connectionist AI, inline with Minsky’s “society of mind”. At its core, semantic gates are designed to facilitate an input/output flow of semantic structures and enable the usage of aligning mechanisms between neural activation clusters and semantic graphs. Having as goal alignment maximization, such a system would enable interpretability through the creation of labeled patterns of computation, and propose unaligned but relevant computation patterns as novel and useful, therefore creative.

Keywords: Creativity · Neural networks · Imagination · Semantic networks · Knowledge · Interpretability · Neural architecture

1 Introduction

Creative cognition, or *useful imagination*, represents the missing functionality which complements the machine *learning*, towards achieving a sense of flow, and uniqueness necessary to generic AI [1]. The ability to generate novel and useful ideas, i.e., solutions, is a key driver of culture creation and human evolution, and of inestimable value in today’s world.

The broadly adopted definition of creativity refers to the process by which cognitive systems instigate the genesis of novelty, which is goal-appropriate [2]. Humans being the only agents known for their ability to achieve creative outcomes with respect to the above definition, it goes per se that, if we want to instigate the act of creation in machines, we need to understand and mimic the way humans operate. Reverse engineering the process of creativity and putting together latest discoveries and insights from neuroscience, cognitive science, and artificial intelligence, is fundamental in architecting the *creative gene*.

Creativity is often described in terms of novelty, but only through the process of decoding (i.e., interpreting) an outcome can we qualify it as novel or useful. The *new* can therefore be assessed only with respect to a reference, which is our current understanding, i.e., knowledge of the world. Following the intuition that a new solution cannot be found in the same “type of thinking”,¹ the process of creation needs a *continuous exchange* with a greater context than that of the environment in which the problem exist. Nevertheless, an AI architecture that takes into account this necessity has not been yet considered, which results in the lack of means towards *generic AI*.

Various branches of philosophy, psychology and evolutionary biology agree that creative ideas should not be just new and unexpected, but are also successful at providing solutions that are both valuable and *efficient*. Creativity is a computationally efficient process [3]. Driven by optimization, agents use creative cognition to build incrementally on ‘simpler’ (after understanding observations of reality, fewer computational resources are needed for an agent to encode/decode) patterns. When shared and adopted globally, creative solutions result in knowledge. Creativity feeds on what we already know, constantly learn, and occurs in the information integration process. All knowledge is the result of past creativity, and a binding material for future one.

As opposed to learning, which is conceptual acquisition, the core characteristic of creativity is conceptual expansion: the ability to widen one’s conceptual structures to include unusual or *novel associations* [4]. This expansion implies that a new relation is created between the “old” and the “new”, and in order to perform this integration step we need to understand the patterns of the creative output. Following this reasoning, a creative process needs to be interpretable.

Besides interpretability, for which a possible solution is presented in Sect. 4, many interdisciplinary advancements are ready to be used to create a ‘society of mind’ [5]: the neuroanatomy of creativity which has been recently revealed using RMI technology in neuroscience [6], goal-oriented measures which assess the success of generated ideas [7], neural networks which pay attention to semantic cues, successfully reuse computation and share a conceptual space [8], semantic graphs which become larger every day through the use of Linked Data and machine learning at scale.

Semantic technologies and neural networks have been rivals at approaching ‘true AI’, with only few, but memorable influences like Marvin Minsky, aiming at their collaboration [9]. Neuroscience also argues that creative thinking emerges through the dynamic interplay between various, functionally diverse, components. Providing an architecture which blends symbolic and connectionist approaches in a suite of cooperating methods is the goal of the INNGenuity framework.

The structure of this paper continues as follows: Sect. 2 proposes pluridisciplinary perspectives regarding the creation and communication of knowledge, and its link to creativity. Section 3 presents the complementary aspects of symbolic and connectionist approaches to AI when considering a creative purpose.

¹ https://www.brainyquote.com/quotes/albert_einstein_385842.

The INNGenuity framework’s components and envisioned implementation are described in Sect. 4, while Sect. 5 outlines conclusions and future work.

2 Reverse-Engineering Creativity

Forging artificial creativity is dependent on first, our ability as a society to renounce at the belief that humans are particularly special because of their unique ability to exercise creative cognition. Second, we need to overcome the supposed complexity which makes the field of creativity largely unexplored by scientists in comparison with its value.

Following the ability to transfer information from memory to external digital resources, to externalize and perform computation at scale using the cloud, enabling imagination in artificial agents is the ultimate resource to be developed in order to further expand our ability to understand the world.

At a philosophical level, the universe itself is a mass creation, whose components all share, at different intensities, and varying qualities, the property of being creative. Most often built as survival strategies, and explained by evolutionary biology as such [10], the creativity mechanisms ensure not only the persistence of information through precise transmission (e.g., genes), but also its expansion either in sheer quantity, or quality (i.e., sophistication) [11].

Humans in particular have evolved to decompose observations of reality into abstractions, or structured mental representations [12], which can be used as building blocks [13] in the construction of more sophisticated abstractions [14], which aim at re-constructing the “reality”. The world, as we perceive it, is partially creativity understood (or, “decoded”), and partially creativity “encoded” (not yet revealed). Therefore, our goal, as “decoding” machines [15], is to understand it. Indeed, humans seem to be drawn, as soon as the survival conditions have been (even loosely) satisfied [16], towards the understanding of one-self and the world [17]. This understanding is made possible by the use of a communication framework i.e., a collection of conventional codes to be shared between agents, which ensures the endurance and expansion of the global creative outcome. Indeed, recent advances in neuroscience [11, 18–20] agree on the fact that language first evolved as a cognitive tool, and only afterward was externalized for information transfer.

Communicating the creative outcome, whether it is mundane or extraordinary, is essential to its assessment and adoption. For practical reasons, in order for the outcome to be useful to the users of the creation, artificial agents need to perform the resolution of the unknown *in a human interpretable way*.

In this paper, the computation framework typically used in cutting-edge AI, deep neural networks, is complemented by the communication framework represented by semantic networks. In the process of rendering black-box models interpretable using semantics, the neural network’s shortcomings in relation to creativity: opacity, high computational needs, and narrow focus are also reduced.

3 Interpretable Neural Networks

Being able to learn nonlinear latent representations through the activation functions, while being highly effective in capturing local relationships, neural networks have been successfully leveraging powerful computational resources and big data. Choosing not to ‘believe without questioning’, and recalling that model ‘silence’ can create monsters, many scientists denounce the black-box model of the world which is presented to us, and endeavor to incorporate side information (i.e., semantics), but face architectural limits.

Rendering machine learning models interpretable would have huge societal impacts: it would not only enable a wider adoption of AI by domain-critical systems, such as medical, but also for security and ethical reasons.

The symbol-oriented community of AI supports models which are self-describing, but alone too rigid and specialized. Semantic networks are graphs which describe a domain using entities, concepts and relations, and have built-in expansion mechanisms which ensure that the “possibly new” is integrated to the “existing” in a consistent manner. Modelling a series of facts, in the form of triples (Subject Predicate Object), semantic graphs such as ontologies represent a flexible way of reasoning on a domain, either generic (law, chemistry) or specific (e.g., state law regulation, molecule interactions etc.).

Common sense knowledge bases like CyC focus on things that rarely get written down or said, providing a causal understanding of the world we live in, which would come at hand to artificial agents. In addition, generic ontologies built from Wikipedia like Freebase, DBpedia, Yago etc., or built by experts like Wordnet, provide a vast cultural coverage. This knowledge has been successfully leveraged by search and QA systems, e.g., Watson AI system, or the Google Graph.

Semantic technologies have nevertheless the shortcoming of lacking a computation framework that sustains the acquisition of new knowledge, and efficiently updates the existent one. Building the facts of an ontology is usually a semi-automatic or manual, costly process involving expert validation. The expert needs to use her own computation node (ie., the brain) to compensate the lack of an automatic, unsupervised framework.

Since a semantic network is built for reuse, the facts need to be true, i.e., logically provable. In reaction to this excessive care for consistency, the connectionist approaches try to avoid the high-maintenance costs and build architectures endowed with as little knowledge as possible. Nevertheless, the results are integrating data biases² which are not obvious and dangerous in the lack of interpretability. By stripping away knowledge from the computation, relying on pure numerical signals, and building black-box models of the world, connectionist approaches also disregard ways in which models could be improved and reused.

The fundamental flaws which create an algorithmic consumerism in our society, and represent a risk factor in many real-world applications are: (1) model

² <https://web.media.mit.edu/~minsky/papers/SymbolicVs.Connectionist.html>.

opacity, (2) high resource consumption, and (3) narrow focus. The hypothesis that using experience in the form of data, when large, it allows for reaching broader conclusions reaches its limits: a “plateau” of efficacy mitigated only by ever-growing resources. This status-quo favors the passive usage of computing resources by those who can afford them, instead of a compute-encode-reuse strategy which is smarter in terms of optimization. Recently, the fields of transfer and multi-task learning try to mitigate this situation, but their concern is more related to reusing encoded data patterns, but not expliciting them, therefore interpretability continues to be an issue.

In addition, since the generalization provided by an NN solves problems of “the same type”, by definition, this narrow perimeter cannot allow the necessary conceptual leap in forging creative solutions. Although narrow focus could be considered a quality, it maintains the learning in a space in which creation could never be instigated. Currently, once an agent trained with NN has learned to play (e.g., Go), it will beat anyone at that task, but cannot do anything else. Additionally, the model lacks all means to understand what is doing, and in relation to what. This is against the generic AI dream, which states that an agent should be able to perform multiple tasks without being reprogrammed.

Attention added to NN has been successful in a number of applications, as a step towards coping with the noisy data problem by identifying relevant parts of the input for the prediction task. Recently, the Transformer architecture [21] claims better accuracies than LSTMs. Attention can only optimize the learning, and although it does not provide the necessary conditions for creativity, it is a proof that we can alter the architecture of neural networks in various beneficial ways. If attention is used as a mechanism to dynamically inject relevant external knowledge into the computation, then it becomes a bridge for the ideas presented in this paper.

4 The INNGenuity Architecture

The goal of this work is to introduce a variant NN architecture with in/out access to semantics through specialized gates. Its structure contains three modules, inspired by recent advances in neuroscience concerning the anatomy of creative cognition.

Over the past years, an important effort in neuroscience has been pursued to localize the creativity in the brain. INNGenuity framework aims to mimick these reverse-engineered processes concerning the brain circuitries (or hubs) underlying creative thought.

4.1 The anatomy of creative cognition

The human brain is recognized to function in a manner consistent with the notion of “hubs” [22], communicating regions of the brain which have built-in mechanisms optimizing the information transfer [23], even across long distances [24]. These are:

The Imagination Hub is involved in ‘constructing dynamic mental simulations based on past experiences, thinking about the future, and generally imagining alternative perspectives and scenarios to the present’ [25]. It is represented in the INNGenuity framework by typical connectionist approaches, namely NNs, but whose functioning has been semantically biased in an automated way.

The Salience Hub constantly monitors both external events and the internal stream of data, while giving priority to whatever information is most salient to solving the task at hand.³ It is represented in the INNGenuity framework by the IN/OUT control mechanism possible through semantic gates, introduced further. This models the perception and attention modules of the brain in the form of a sensitivity to, or prioritary treatment of relevant observations.

The Executive Hub is active when engaging in reasoning that puts heavy demands on the memory. Represented in the INNGenuity framework by semantic technologies, it solves the following: (i) encoding: finding similarities between the NN patterns’ *structure* and the topical semantic graph which is proper to the domain of computation; (ii) decoding: maps meaning (i.e., labels) to encodings using an interpretation scheme, made of a vocabulary of symbols and the relations between them; (iii) abductive reasoning: combines the explanations obtained through the output semantic gates of the NN, and assesses which parts of the result are novel by comparing them with the *global* knowledge (i.e., alignment with a generic semantic graph), and the goal, respectively.

4.2 Semantic gates motivation and purpose

Knowing that creation comes with the pursue of a ‘different resolution path than the expected one’ [26], INNGenuity introduces semantic biases in order to create meaningful divergence from initial representations.

Adding special gates to NN is not new. The forget gate has introduced to LSTMs, but generally, the gates purpose is to optionally let information through, since they control access to memory cells. In theory, semantic gates can be added to any variant of NNs, for instance Long Short-Term Memory (LSTM) networks.

Current variants of LSTMs have the ability to read, write and erase data from the memory cells, but these data are static to the learning process. An alternative idea promoted in this work is to dynamically bias the input data at every step of the computation with relevant knowledge proper to that current layer level of abstraction. Besides its positive contribution to expanding the solution search space, access to background knowledge opens the door to a larger context that the local one, and has been endorsed as fundamental for creativity by both cognitive science [4], and neuroscience [27, 28]. LSTM networks learn to process data with complex and separated interdependencies, but restricted to the typical input-output settings, they are not capable of imagination, which requires a non-linear relationship to the data.

³ https://med.stanford.edu/content/dam/sm/scsnl/documents/Menon_Salience_Network_15.pdf.

Starting from a typical NN, LSTM [29] for instance, in addition to the typical forget gate, *semantic gates*, denoted by k_{IN} and k_{OUT} , are introduced as in Fig. 1.

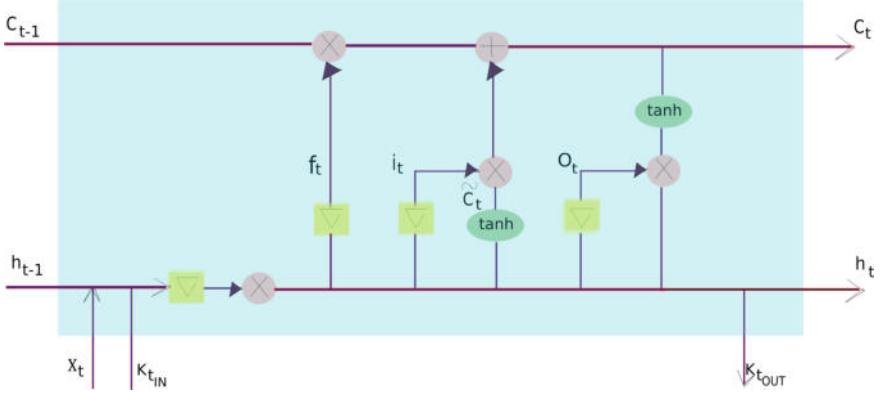


Fig. 1. Semantic gate IN/OUT contribution to memory cell

The purpose of semantic gates to NN is twofold: (1) they operate towards a meaningful bias of the computation by adding contextual external knowledge to the internal data context of the NN's layers, and (2) they allow the output of the patterns identified at different layers of abstraction as clusters, i.e., graphs composed of impactful neural activation dependencies. This latter function is beneficial for interpretability: the clusters will pass through an alignment phase [30] with a topic semantic graph to explicit the computation.

The semantic gate implements the Salience Hub of attention and awareness. Allowing bidirectional communication between NN and semantic graphs, its logic is regulated by the aligning mechanisms of the Executive Hub, dealing with knowledge integration and meta-reasoning. Through the semantic gates, a bidirectional metadata flow provide the necessary conditions for creativity as introduced in [31]: (1) a mechanism for introducing variation (IN semantic bias), and (2) a mechanism for preserving and reproducing the selected variations (OUT persistent expression).

The exploratory and eliminatory aspects of the creativity process can be also mimicked [32]: ‘a sequential back-and-forth begins with informed guesses and progresses to increasingly probable solutions’, can be implemented in semantic networks through the mechanism of alignment towards the increasingly abstract. Semantics feature hierarchies are shown to be effective in boosting the accuracy, besides the interpretability of neural models [33].

A NN outlines a process in which the existence of labeled data (i.e., outcomes) is a condition for learning. In opposition, triggered when we face unknown situations, creativity implements an abductive reasoning process [34], i.e., an adaptive problem solving strategy. The abductive feature of the reasoning supposes that

the outcome instances are not precisely defined, instead the outcome has to be ‘invented’, or *explicated using the most relevant facts*. In order to enable this abductive property of the creative process, but still using the NN’s characteristic of an universal approximator, the ‘focus’ of the NN has to be relaxed. There are two ways of approaching that: (1) if outcomes exist, they will be conceptually expanded to a semantic graph; or (2) if outcomes are not precisely defined, then a semantic graph relevant to the problem to be solved (i.e., a topic) is projected using the access to a generic knowledge base.

4.3 Disentangling neural activations into clustered signals carrying semantics

Rendering an agent interpretable, is equivalent to transforming its black-box model into a self-explaining structure. This cannot be done a posteriori, therefore it needs to happen *while learning*.

Disentangling the underlying explanatory factors hidden in the observed environment is one important goal of Representation Learning [35]. Interpretability needs however an additional step involving attaching meta-information to the explanatory factors. In this paper, the approach to interpretability is employing semantic networks, which have the advantage of providing a self-describing graph structure.

Many successful learning systems benefit from prior knowledge about composition and structure, but the large majority are supervised, while this work describe an unsupervised approach.

The aim of hybrid neural systems is finding best ways to integrate both symbolic and connectionist approaches [36], but for that the connectionist mechanisms need to become more transparent. Only recently differentiable clustering algorithms have been applied on neural signals for the identification of objects and their interactions [37]. Intuitively, the relevant signals identified during computation correspond to object properties encoded. If the encoding (i.e., embedding) scheme is coherent and the transformations consistent, then the clustering would be able to outline the causal dependencies between the primitives of the compositional system, therefore semantically explicit the **isA** relation learned by the model.

For that, input and output data embeddings are used in the selection of more relevant information from a knowledge source, provided that embeddings of the knowledge graph have been computed in advance, e.g. see [38].

As shown in Fig. 2 the typical NN processing is complemented with a phase of clustering, and two phases of alignment with a knowledge graph. The optimization criteria for the NN is then to maximize alignment, in addition to the typical loss minimization. During NN computation the most impactful signals are being conceptually cached and clustered. The resulting clusters represent a graph (set of triples), which is outputted through the semantic gate of that layer. Then, the Executive Hub takes over, and in the process of alignment with a topical semantic graph, the graph of clustered activations is being labelled with relevant concepts based on the structure similarities of the alignment candidates. Linking

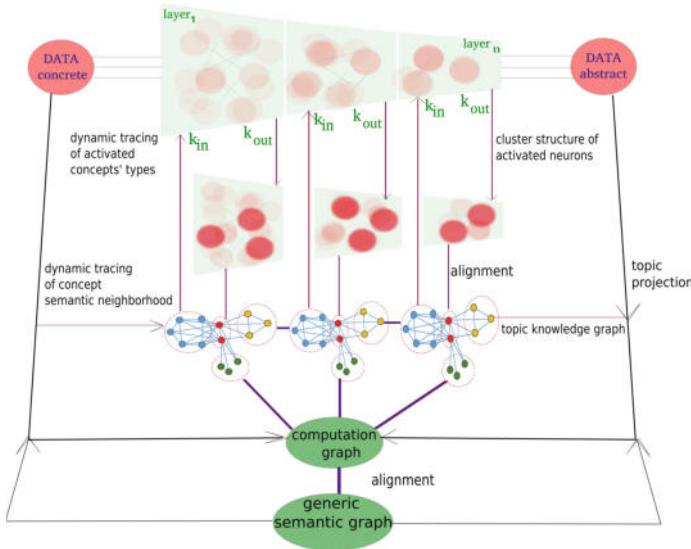


Fig. 2. The clustering and alignment phases

together labelled explanatory factors identified at every layer of abstraction is equivalent to rendering the computation steps explicit.

The use of semantic gates ensures that knowledge is used contextually, therefore the process diverges from a typical learning due to the introduction of a relevant variation. The results become imaginative because of their plausibility, but only novel and useful ones will be considered creative. The assessment of these two properties is the goal of the next phase of alignment.

The *novel* and *useful* attributes of each of the triples (Subject Predicate Object) forming the labelled computation graph are measured. In order to do that, we need a second alignment and to make use of the impact weights of activated neurons. In this process, the generic knowledge base mimicks the ‘culture’ of that environment and serves as a reference for identifying new associations. Partially unaligned structures (triples) which are most impactful would be considered novel (and useful by the fact that they have been activated during computation), therefore by definition creative. Eventually, the culture, or collective memory would be constantly enriched by collaborative agents with new outcomes in the defined domain of operation, contribution which unifies at scale *views* on computations made by different agents.

In contrast with recent causal frameworks which aim at explaining the predictions of a NN model [39], INNGenuity dynamic flow allows a better flexibility and the possibility of designing a fully unsupervised process since both the clustering of patterns, and the semantic alignment are unsupervised.

From a symbolic perspective, INNGenuity designate a semantic network in which relations between concepts represent *computation nodes* (aka NN), and

the relation label is a clarification of the computation node's purpose. From a connectionist perspective, INNGenuity designate a neural network in which the meaningful data patterns and their interactions are being transformed into a semantic graph which explicates the computation.

5 Discussion and Conclusions

Creative AI is the most human invention that we have the chance of pursuing. Using insights from neuroscience, the functions of a creativity framework are defined as the generation and assessment of novelty. This work argues that the generation of novelty needs a dynamic integration of meaningful conceptual structures (aka plausible biases) to memory cells. At the same time, the assessment of novelty needs interpretability. The goal of the INNGenuity architecture is to enable both interpretability and possible creative outputs. Its design instigates towards the conciliation of best AI practices for more impactful progress in this domain.

From a hybrid systems classification, INNGenuity outlines a connectionist symbol processing approach, a tightly coupled system in which knowledge and data are dynamically transferred and shared by the neural and symbolic components, via common internal structures: semantic gates.

The flow of meaningful conceptual structures at each NN layer of abstraction through input semantic gates is thought as enabling means of ‘inspiration’ and context awareness. Purely guided by the recognition of patterns formed during computation, INNGenuity incorporates a differentiable clustering algorithm which outlines the structural dependencies between activated neurons. Activation clusters, which are assumed to carry hidden semantics and compositional logic, are further aligned with the rich and relevant knowledge of our world in the form of (nowadays pervasive) semantic graphs.

Knowledge, seen as the result of past computation and modelled by symbolic systems as semantic graphs, needs to be seemingly integrated into the live computation process for its huge potential in terms of resource optimization, but also for its unique capability to make unknown structural dependencies explicit by means of alignment.

Aligning pattern structures is a *decoding mechanism* which enables the creation of a labelled computation graph. This facilitates model understanding by humans, and communication with other artificial agents. The flow of such structures through output semantic gates is thought as enabling means of ‘expression’. By means of semantic graphs alignment, patterns of computation get labeled, shared, and reused, resulting in knowledge being enriched incrementally and continuously in collaboration.

Eventually, unaligned parts of the semantic computation graph constitute the possible creative outcome. Since they represent structures activated during computation, these unaligned parts are considered useful, besides representing new associations.

One of the main advantages of the INNGenuity architecture is the integration of unsupervised approaches, such as alignment and clustering, thus allowing a high degree of automation.

A current condition towards the successful implementation of the INNGenuity approach is existence of efficient differentiable clustering algorithms operating on neural activations. Another limitation (but which will always exist due to our own limited ability to capture and express knowledge) is that achieving accurate and fine-grained interpretations depends on the quality and depth of the semantic graphs made available to an agent. In this direction, semantic gates are envisioned as a feedback loop between NN and semantic graphs, setting in which knowledge can grow and improve over time due to the controlled interaction between real-world data (sometimes also real-time), and the conceptual and compositional representations we have about these data.

As further work, the author aims at implementing a prototype of INNGenuity producing creative outcomes as a neural network operating in alignment with knowledge graphs. Its potential can be shown in applications in which creativity is more than encouraged: generative systems e.g., language models, dialogue systems and chatbots, or in other processes of assessing and boosting human creation, such as research.

References

1. Weisberg, R.W.: The creative mind versus the creative computer. *Behav. Brain Sci.* **17**(3), 555–557 (1994)
2. Runco, M.A., Jaeger, G.J.: The standard definition of creativity. *Creativity Res. J.* (2012)
3. Schmidhuber, J.: Driven by compression progress: a simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *CoRR* (2008)
4. Terry, D.: Creativity, cognition, and knowledge: an interaction (2002)
5. Minsky, M.: *The Society of Mind*. Simon & Schuster Inc, New York (1986)
6. Jung, R.E., Segall, J.M., Bockholt, H., et al.: Neuroanatomy of creativity. *Hum. Brain Mapp.* (2010)
7. Georgiev, G.V., Georgiev, D.D.: Enhancing user creativity semantic measures for idea generation: Knowl.-Based Syst. **151**, 1–15 (2018)
8. Wong, C., Gesmundo, A.: Transfer learning to learn with multitask neural model search. *CoRR* (2017)
9. Singh, P.: Examining the society of mind. *Comput. Informat.* **22**, 521–543 (2004)
10. Kaufman, S.B., Gabora, L.: Evolutionary approaches to creativity. *Camb. Handb. Creativity*, 279–300 (2011)
11. Baas, M., Nijstad, B.A., De Dreu, C.K.W.: (ed.): The cognitive, emotional and neural correlates of creativity. *Front. Hum. Neurosci.* **9**, 275 (2015)
12. Fodor, J.A., Pylyshyn, Z.W.: *Minds Without Meaning: An Essay on the Content of Concepts*. The MIT Press, Cambridge (2015)
13. Karmiloff-Smith, A.: Is creativity domain specific or domain general? Cases from normal and abnormal phenotypes. *Artif. Intell. Simul. Behav. Q.* **85** (1993) (T.H. Dartnall)

14. Karmiloff-Smith, A.: Digest of beyond modularity. *Behav. Brain Sci.* **17**(4) (1994)
15. Thagard, P.: *Mind: Introduction to Cognitive Science*. The MIT Press, Cambridge (1996)
16. Dawkins, R.: *The Selfish Gene*. 1941-(1989)
17. Singer, I.: Modes of Creativity: Philosophical Perspectives (2013)
18. Reboul, A.C.: Why language really is not a communication system: a cognitive view of language evolution. *Front. Psychol.* pp. 14–34 (2015)
19. Sperber, D., Wilson, D.: *Relevance: Communication and Cognition*. Basil Blackwel, Oxford (1995)
20. Holdgraf, C.R., Rieger, J.W., Michelli, C., Martin, S., Knight, R.T., Theunissen, F.E.: Encoding and decoding models in cognitive electrophysiology. *Front. Syst. Neurosci.* **11**, 61 (2017)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *CoRR*, abs/1706.03762 (2017)
22. Bressler, S.L., Menon, V.: Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn. Sci.* (2010)
23. Sporns, O., Honey C.J., Kotter, R.: Identification and classification of hubs in brain networks. *PLoS ONE* (2007)
24. Bassett, D.S., Bullmore, E.: Small-world brain networks. *Neuroscientist* **12**, 512–523 (2006)
25. Buckner, R.L., Andrews-Hanna, J.R., Schacter, D.L.: The brain's default network: anatomy, function, and relevance to disease. *Ann. N.Y. Acad. Sci.* **1124** (2008)
26. Pascanu, R., Weber, T., Racanière, S., Reichert, D.P., Buesing, L., Guez, A., Rezende, D.J., Badia, A.P., Vinyals, O., Heess, N., Li, Y., Battaglia, P., Silver, D., Wierstra, D.: Imagination-augmented agents for deep reinforcement learning. *CoRR* (2017)
27. Hummel, J., Holyoak, K.: A symbolic-connectionist theory of relational inference and generalization. *110*, 220–264 (2003)
28. Hummel, J.E., Holyoak, K.J.: Distributed representations of structure: a theory of analogical access and mapping. *104*, 427–466 (1997)
29. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
30. Mills, G.J., Healey, P.G.T.: Semantic negotiation in dialogue: the mechanisms of alignment (2008)
31. Campbell, D.T.: Blind variation and selective retention in creative thought as in other knowledge processes. *Psychol. Rev.* **67** (1960)
32. Simonton, D.K.: Creative problem solving as sequential bvsr: exploration (total ignorance) versus elimination (informed guess). *Thinking Skills Creativity* **8** (2013)
33. Peterson, J.C., Soulou, P., Nematzadeh, A., Griffiths, T.L.: Learning hierarchical visual representations in deep neural networks using hierarchical linguistic labels. *CoRR*, abs/1805.07647 (2018)
34. Jung, R., Mead, B., Carrasco, J., Flores, R.: The structure of creative cognition in the human brain. *Front. Hum. Neurosci.* **7**, 330 (2013)
35. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(8), 1798–1828 (2013)
36. Kenneth McGarry, S.W., MacIntyre, J.: Hybrid neural systems: from simple coupling to fully integrated neural networks. *Neural Comput. Surv.*, pp. 62–93 (1999)
37. van Steenkiste, S., Chang, M., Greff, K., Schmidhuber, J.: Relational neural expectation maximization: unsupervised discovery of objects and their interactions. *CoRR* (2018)

38. Wang, X., Ye, Y., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. CoRR (2018)
39. Alvarez-Melis, D., Jaakkola, T.S.: A causal framework for explaining the predictions of black-box sequence-to-sequence models. CoRR, abs/1707.01943 (2017)



Developing a Deep Learning Model to Implement Rosenblatt's Experiential Memory Brain Model

Abu Kamruzzaman^(✉), Yousef Alhwaiti, and Charles C. Tappert

Seidenberg School of Computer Science and Information Systems,
Pace University, Pleasantville, NY, USA
`{ak91252p, yal2919p, ctappert}@pace.edu`

Abstract. This paper describes initial work to develop a Deep Learning model for long-term sequential memory storage to implement Rosenblatt's experiential memory Perceptron architecture brain model. In recent years, deep learning techniques have solved many problems in the area of computer vision, language modeling, speech recognition, and audio/video processing. Further, CNN based models are considered state-of-the-art algorithms to solve perceptron related problems. However, can Deep Learning models store the learned knowledge representation to make better use of classifying and recognizing images and other patterns? The Deep Learning models explored here include CNNs pre-trained models (ResNet50, VGG16, InceptionV3, and MobileNet) on ImageNet datasets and trained model on MNIST datasets.

Keywords: Convolutional neural network · Pre-trained models · Handwritten character recognition · Deep learning · Brain model · Machine learning · Perceptron

1 Introduction

1.1 Long-Term Declarative Episodic Memory

The most popular model for studying memory, the Atkinson-Shiffrin model, consists of a sequence of three stages, from sensory to short-term to long-term memory (Fig. 1). In this study, we are concerned with long-term declarative episodic (experiential) memory, the memory of experiences and events in time in a serial form. This memory allows us to reconstruct the actual events that took place at any given point in our lives, and apparently decays little over time and can store an essentially unlimited amount of information almost indefinitely. In fact, there is currently a debate as to whether we ever forget anything, or whether it merely becomes difficult to retrieve certain items from memory.

1.2 Convolutional Neural Network

Convolutional neural network is a class of deep learning that is used to analyze images. Convolutional neural networks utilize a type of multilayer perceptron, which consists

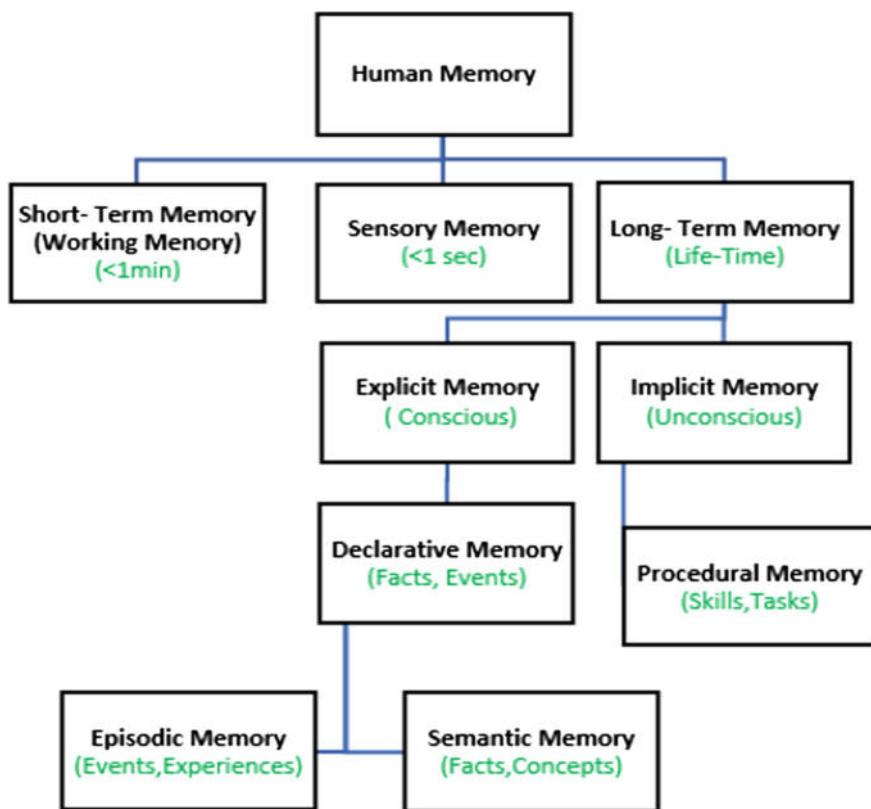


Fig. 1. Types of human memory (adapted) [2].

of an input, output, and at least one hidden layer, that is designed to learn with little to no preloading [1]. The system will translate pixels and features derived from an image to classify the object using provided training. The classification of the output is assigned a probability based on the numeric translation and informed data via training.

A typical convolutional network completes its task over several phases. First, the system will take an input from an image, as a sample for analysis. Convolution is the comparison of a set area of pixels to the rest of an image; these pieces are referred to as ‘features’. The system attempts to match these features using a simple mathematical formula. In order to calculate the match of the source pixel area to the feature, the network multiplies each pixel. This process repeats across the image, attempting to match each pixel. CNN will attempt these comparisons wherever possible, in order to calculate the highest accuracy. Afterwards, this process is repeated on other identified subsamples. CNN also utilizes a tool termed “Pooling”, which is taking large areas in the images and shrinking them down, saving the important calculated information. One more CNN feature is named “Rectified Linear Units”; in its most basic form, the

system swaps out negative calculations from convolution for a zero. This prevents the numbers from becoming too minuscule or too large, effectively simplifying the information to only identify the important characteristic components.

In order to enable a system to achieve highly accurate results, they must be trained and tested on datasets. There are shortcomings of datasets that contain a limited amount of images, whereas new larger datasets increase the amount of variability [3]. This, in turn, reflects reality where larger datasets allow for additional training and accuracy of system functionality.

The MNIST is a database of handwritten characters that contains 60,000 training set examples and 10,000 test set examples [4]. All of the characters within MNIST are numbers derived from a larger set from NIST that have been size-normalized and have been aligned in the center [4]. Currently, the test error rate for the MNIST dataset has been brought down to 0.23% through the usage of a convolution neural network [5].

1.3 Pre-trained CNN Models

A pre-trained model, in essence, is a model that was developed by another entity that is being reused for a similar purpose. In the same rationale theory of the idiom “do not reinvent the wheel”, developers can use an existing model as a starting point for their own directions [3]. The thought process belonging to this practice is to build on the prior successes of other developers that have shared their systems in order to improve the efficiency of the existing system or provide a basis for others, so that they may tailor their experiments to their unique needs without the need to start from the beginning.

The transferability between projects has provided developers effective methods to complete their own ambitions more quickly [3]. At Berkeley in California, a pre-trained system referred to as Caffe is a modular open-source system, which allows outside developers to customize the neural network [3]. Already, several additional research universities, as well as Facebook and Adobe, have collaborated with Caffe to generate results concurrent with their own networks [3].

By modulating the system, developers can modify existing parameters or constraints to support their own system, if they were to fabricate one [6]. In another experiment, two separate pre-trained CNN models were used to categorize items by color and depth, which outputs were then combined into another set of layers to determine an appropriate category and label [6]. By combining two pre-trained CNN networks, that are then analyzed further, the engineers were able to reduce production time and still attain desirable results [6].

From these two examples, we can gather that pre-trained models can optimize a solution for a problem. Existing systems do not need to be set through additional training, if they are close to the intended target. Moreover, it would appear that we save time on development, budget, and personnel time, because tasks are based on modification rather than creation.

In this research project, we are trying to understand if Deep Learning models are suitable to store the learned knowledge representation to make better use of classifying and recognizing images and other patterns. The Deep Learning models explored here include CNNs pre-trained models (ResNet50, VGG16, InceptionV3, and MobileNet) on ImageNet datasets and trained model on MNIST datasets.

2 Literature Review

2.1 VGG16

As a deep learning model the VGG model is a study of convolution neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper “Very Deep Convolutional Network for Large-Scale Image Recognition”. According to the paper, this model achieves 92.7% top-5 test accuracy in ImageNet. ImageNet is a dataset of over 14 million images belonging to 1000 classes. The architecture includes a pre-processing layer which is used to take RGB images with pixel values as input in the range of 0–255, which then subtracts the mean image value (the value is calculated over the entire Imagenet dataset). Figure 2 shows the high level architecture of VGG16.

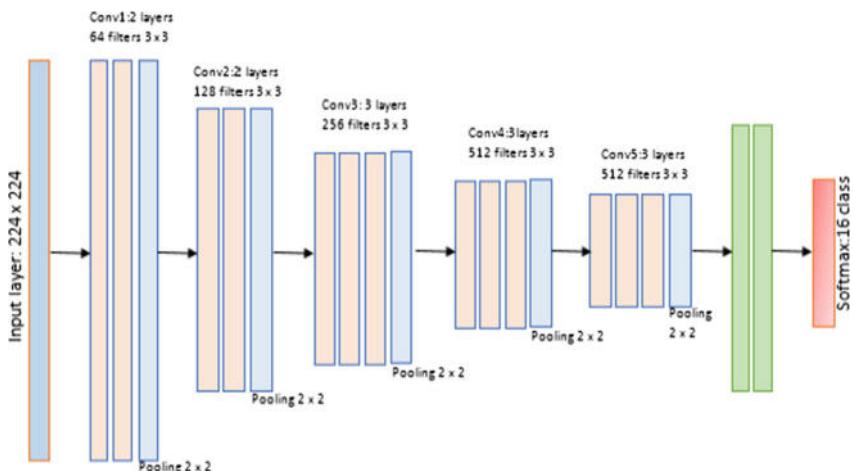


Fig. 2. Macro architecture of VGG16 (adapted) [7].

2.2 ResNet50

ResNet is a short name for Residual Network. As the name suggests, the network introduces residual learning. In general, in a deep convolutional neural network, several layers are stacked and trained to the tasks that need to be processed.

Residual learning can be understood as a subtraction of features learned from input of the layer. Figure 3 displays the building block for the ResNet which uses shortcut connections. It does so by directly connecting input of the nth layer to (n + x)th layer. Training this form of networks is easier than training simple deep convolutional neural networks, and also the problem of degrading accuracy is resolved. ResNet50 is a 50 layer Residual Network. There are other variants like ResNet101 and ResNet152.

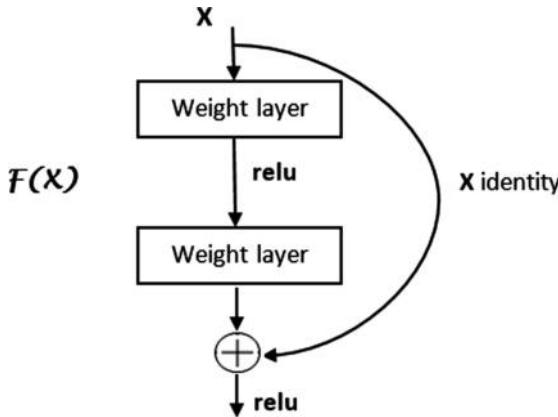


Fig. 3. Residual learning: a building block (adapted) [8].

2.3 MobileNets

MobileNets models are small, low-latency, low-power models. They are parametrized to meet the resource constraints of a variety of use cases. They can be built upon classification, detection, embedding and segmentations similar to how other large scale models, such as Inceptions, are used. The main difference between the MobileNet architecture and “traditional” CNN’s architecture is that instead of a single 3×3 convolution layer followed by a batch norm, MobileNets splits the convolution into 3×3 depth-wise conv. and 1×1 pointwise conv. Figure 4 displays how MobileNets can be applied.

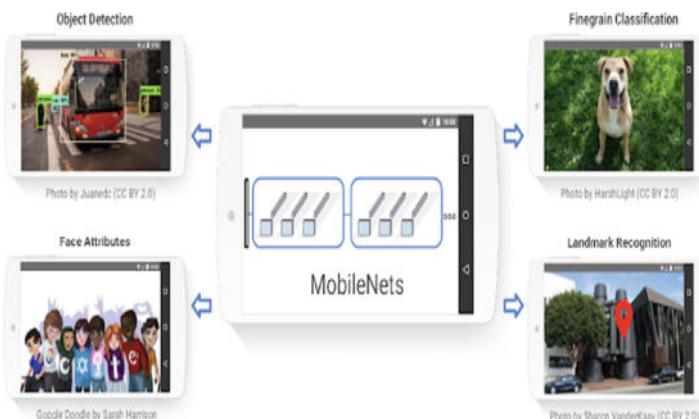


Fig. 4. MobileNets can be applied to various recognition tasks for efficient on device intelligence (adapted) [9].

2.4 InceptionV3

InceptionV3 is a variant of InceptionV2 with an addition of BN-Auxiliary. BN-auxiliary refers to the version in which a fully connected layer of auxiliary classified is also-normalized, not just convolutions. The model InceptionV2 + BN-Auxiliary is referred as InceptionV3. The InceptionV3 architecture is the same as InceptionV2, but with minor changes. One of the benefits of using this model is that it can be done using depth multiplier 8 on the first convolution layer. This reduces computational cost while increasing the memory consumption at training time. Figure 5 displays the architecture of InceptionV3.

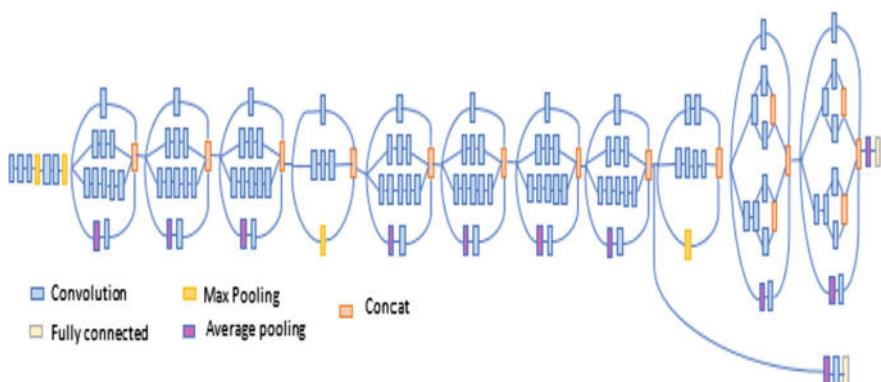


Fig. 5. Architecture of InceptionV3 (adapted) [10].

3 Methodology

The study is to compare, discuss, and experiment the Convolutional Neural Network (CNN) models with the Rosenblatt Perceptron model using a c-system [11] as the memory storage and to discuss the limitation of using CNN. Our experiments and results focus on:

1. How well can CNN classify the spatial data?
2. How does CNN learn hidden layers representation; does it store the hidden layers for future predictions?

We want to identify if CNN alone can be used to build an experiential storage in neural networks, which is known as Rosenblatt's brain model. Rosenblatt defined the neural model with an additional unit (c-system) and he defined it as the Perceptron model. This model consists of a set of units (or neurons) which generates the signal connected together in a network.

After receiving a signal input, the consecutive units (either another unit or the environment) respond by generating an output signal, which may be transmitted to another set of selected units in a network. All the different units in the architecture are called perceptron. Each perceptron has a sensory input and one or more output units, which generates signals. The following are the logical properties of the Rosenblatt memory brain model (Fig. 6):

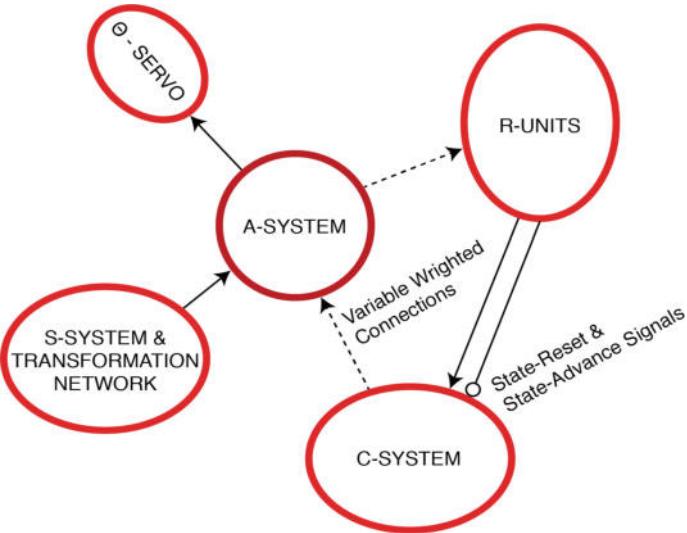


Fig. 6. Rosenblatt C-system (adapted) [11].

1. A set of signal propagation rules which govern the transmission and generation of signals.
2. A set of memory rules for modifying the properties of the network as a result of an activity.

This is a sequential memory model which can be adapted to solve problems related to speech recognition, language modeling, language translation, and audio/video processing. Figure 6 consists of four major units. It starts with a source transmission system (S-system) connected with A-system (Association system), which is capable of recording and detecting the important features and sends the information to R-system (Response unit) as the output. O-servo acts a control mechanism which maintains the constant level of activity, despite changes in the intensity of the input signals. The C-system is connected to both the A-system and R-system, which acts as a memory of the sequences. This system may be synchronous or asynchronous, in which case it goes from one state to another state when only certain events are triggered.

3.1 Convolution Neural Network

In the first layer, an input image (e.g. dog.jpg) is provided to the CNN model as depicted in Fig. 7. The operations of the convolution neural network will learn the features of the input image, and based on the pre-trained weights it will determine the label that identifies the input image showed in the form of output.

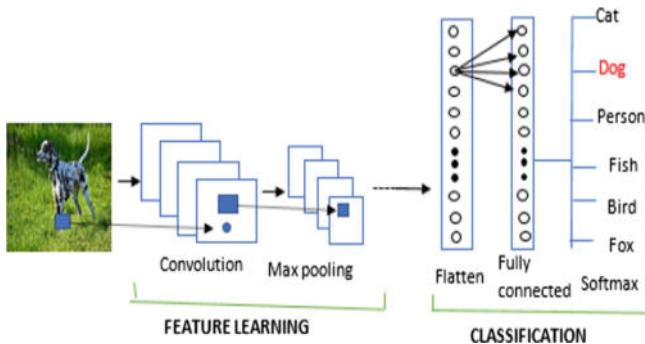


Fig. 7. A general depiction of the convolution process (adapted) [12].

4 Project Requirements

In order to test the performance of the CNN-based models, several key requirements must be met. To begin, anyone who will be using these models will be using Python version 3 and Keras library. Furthermore, a background knowledge of Convolutional Neural Network and an understanding of the Rosenblatt perceptron model is necessary to compare the architectures.

5 Experiments

Several experiments have been conducted to check how CNN performs in the classification and recognition tasks introduced as question #1 (How well can CNN classify the spatial data) in the methodology section. To perform the experiment, pre-trained CNN models (such as ResNet50, VGG16, InceptionV3, and MobileNet trained on ImageNet datasets) have been used to make more generalized classifications with any images (from Experiment 1 to 4). To answer the second question (How does CNN learn hidden layers representation; does it store the hidden layers for future predictions) of the methodology, Experiment-5 has been conducted on the MNIST dataset to see what representations have been learned. Figure 8 (Seashore, Banana, Lion, Orange) are the test case images considered for Experiments 1–4.



Fig. 8. Sample of images (adapted) [13].

5.1 Experiment 1—ResNet50 Model

ResNet50 weights were downloaded in Keras and loaded for the experiment on the images considered. The summary of the ResNet50 model is depicted in Fig. 9.

res5c_branch2c (Conv2D)	(None, 7, 7, 2048)	1050624	activation_97[0][0]
bn5c_branch2c (BatchNormalizati	(None, 7, 7, 2048)	8192	res5c_branch2c[0][0]
add_32 (Add)	(None, 7, 7, 2048)	0	bn5c_branch2c[0][0] activation_95[0][0]
activation_98 (Activation)	(None, 7, 7, 2048)	0	add_32[0][0]
avg_pool (AveragePooling2D)	(None, 1, 1, 2048)	0	activation_98[0][0]
flatten_2 (Flatten)	(None, 2048)	0	avg_pool[0][0]
fc1000 (Dense)	(None, 1000)	2049000	flatten_2[0][0]
<hr/>			
Total params:	25,636,712		
Trainable params:	25,583,592		
Non-trainable params:	53,120		

Fig. 9. Summary of the RestNet50 model.

5.2 Experiment 2—VGG16 Model

VGG16 weights were downloaded in Keras and loaded for the experiment on the images considered. The summary of the VGG16 model is depicted in Fig. 10.

5.3 Experiment 3—InceptionV3 Model

InceptionV3 weights were downloaded in Keras and loaded for the experiment on the images considered. The summary of the InceptionV3 model is depicted in Fig. 11.

5.4 Experiment 4—MobileNet Model

MobileNet weights were downloaded in Keras and loaded for the experiment on the images considered. The summary of the MobileNet model is depicted in Fig. 12.

5.5 Experiment 5—MNIST Experiment

Tests were performed on the MNIST handwritten dataset (Fig. 13) to see what representation CNN is learning.

block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000
<hr/>		
Total params: 138,357,544		
Trainable params: 138,357,544		
Non-trainable params: 0		

Fig. 10. Summary of the VGG16 model.

concatenate_2 (Concatenate)	(None, None, None, 7 0)	activation_190[0][0]
		activation_191[0][0]
activation_192 (Activation)	(None, None, None, 1 0)	batch_normalization_94[0][0]
mixed10 (Concatenate)	(None, None, None, 2 0)	activation_184[0][0]
		mixed9_1[0][0]
		concatenate_2[0][0]
		activation_192[0][0]
avg_pool (GlobalAveragePooling2	(None, 2048)	0
predictions (Dense)	(None, 1000)	2049000
		avg_pool[0][0]
<hr/>		
Total params: 23,851,784		
Trainable params: 23,817,352		
Non-trainable params: 34,432		

Fig. 11. Summary of the InceptionV3 model.

```

conv_pw_13_relu (Activation) (None, 7, 7, 1024)          0
global_average_pooling2d_1 ( (None, 1024)                0
reshape_1 (Reshape)           (None, 1, 1, 1024)         0
dropout (Dropout)            (None, 1, 1, 1024)         0
conv_preds (Conv2D)          (None, 1, 1, 1000)          1025000
act_softmax (Activation)     (None, 1, 1, 1000)          0
reshape_2 (Reshape)           (None, 1000)              0
=====
Total params: 4,253,864
Trainable params: 4,231,976
Non-trainable params: 21,888

```

Fig. 12. Summary of the MobileNet model.**Fig. 13.** MNIST dataset (adapted) [14].

6 Results

After conducting the experiments, it can be noted that the Convolutional Neural Network is performing excellently on the classification and recognition task on the test images considered. The results of all the pre-trained models are compared in Table 1 (from Result 1 to 4). To check how and what representations CNN is learning, the results of Experiment-5 are noted below in Result-5.

Table 1. Comparing the results of the models.

Comparing the pre-trained model accuracy				
Pre-trained models	Image 1: Banana	Image 2: Orange	Image 3: Lion	Image 4: Seashore
ResNet50	0.989	0.998	0.999	0.375
VGG16	0.986	0.982	0.999	0.594
InceptionV3	1.0	0.999	0.992	0.864
MobileNet	0.999	0.988	0.998	0.426

6.1 Result 1—ResNet50 Model

After the experiment of ResNet50, the predictions of the considered images are depicted in Fig. 14.

```
resnet50
('Predicted:', [[(u'n09428293', u'seashore', 0.37578392)]])
('Predicted:', [[(u'n07753592', u'banana', 0.98976445)]])
('Predicted:', [[(u'n02129165', u'lion', 0.99994135)]])
('Predicted:', [[(u'n07747607', u'orange', 0.99839956)]])
```

Fig. 14. ResNet50 model results.

6.2 Result 2—VGG16 Model

After the experiment of VGG16, the predictions of the considered images are depicted in Fig. 15.

```
vgg16
('Predicted:', [[(u'n09428293', u'seashore', 0.5942951)]])
('Predicted:', [[(u'n07753592', u'banana', 0.98645973)]])
('Predicted:', [[(u'n02129165', u'lion', 0.9999863)]])
('Predicted:', [[(u'n07747607', u'orange', 0.9823294)]])
```

Fig. 15. VGG16 model results.

6.3 Result 3—InceptionV3 Model

After the experiment of InceptionV3, the predictions of the considered images are depicted in Fig. 16.

```
inception_v3
('Predicted:', [[(u'n09428293', u'seashore', 0.8640787)]])
('Predicted:', [[(u'n07753592', u'banana', 1.0)]])
('Predicted:', [[(u'n02129165', u'lion', 0.99265844)]])
('Predicted:', [[(u'n07747607', u'orange', 0.99999833)]])
```

Fig. 16. InceptionV3 model results.

6.4 Result 4—MobileNet Model

After the experiment of MobileNet, the predictions of the considered images are depicted in Fig. 17. Seashore predicted as sandbar in the MobileNet model since the human label readable names in ImageNet are 977: ‘sandbar, sand bar’; 978: ‘seashore, coast, seacoast, sea-coast’.

```
mobilenet_1.00_224
('Predicted:', [[(u'n09421951', u'sandbar', 0.42680004)]])
('Predicted:', [[(u'n07753592', u'banana', 0.9993048)]])
('Predicted:', [[(u'n02129165', u'lion', 0.9989668)]])
('Predicted:', [[(u'n07747607', u'orange', 0.9885423)]])
```

Fig. 17. MobileNet model results.

6.5 Result 5—Visualizing CNN Representations on MNIST

From the representation visualization in Fig. 18, it can be seen that the Convolutional Neural Network is only transforming the original input in the form of repeated layers of certain operations (specifically Convolution and Max Pooling operation).

Undertaking all the layers, the final layer is just a high-level layer representation which has learned discrete patterns of features without considering or storing any past sequences of patterns.

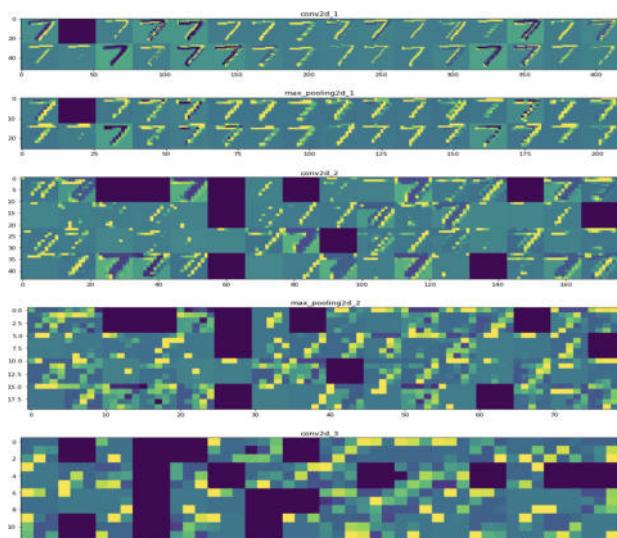


Fig. 18. Transformed MNIST numbers.

7 Conclusions and Future Works

We learned that CNN is mostly used for classification of spatial data. It is good at repetitive transformation of the data into high level discrete representation of knowledge. CNN does not store or remember any past sequences of patterns that can be used for future predictions and is unable to store the learned knowledge representation to make better use of classifying and recognizing images and other patterns. Our future goal is to combine CNN and LSTM to solve the existing storage problem. The memory architecture proposed by Rosenblatt can be for memory storage of other types of data (such as speech recognition, language modeling, language translation, and audio/video processing), particularly where past sequences are important factors for future predictions.

References

1. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. *AAAI* **2**(5), 6 (2016)
2. Mastin, L. : Types of Memory-The Human Memory. *Human-Memory.Net*. (2010)
3. Karayev, S., Jia, Y., Shelhamer, E., Donahue, J., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)
4. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436 (2015)
5. Cires, D., Meier, U., Schmidhuber, J.: Multi-column Deep Neural Networks for Image Classification. *IDSIA/USI-SUPSI*, Manno, Switzerland (2012)
6. Schwarz, M., Schulz H., Behnke, S.: RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In: IEEE International Conference on Robotics and Automation, Seattle (2015)
7. Amarntunga, T.: Build Deeper: Deep Learning Beginners' Guide (2017)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
9. Google AI Blog MobileNet, <https://ai.googleblog.com/2017/06/mobilenebs-open-source-models-for.html>, written 2017/06/14
10. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
11. Rosenblatt, F.: A Model for Experiential Storage in Neural Networks. Spartan Books, Washington, DC (1964)
12. CNN workflow image, <https://flickrccode.files.wordpress.com/2014/10/conv-net2.png>, last accessed 2018/04/20
13. ImageNet Image, <http://www.image-net.org/> (2016)
14. Bottou, L., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Jackel, L.D., LeCun, Y., Muller, U.A., Sackinger, E., Simard, P., Vapnik, V.: Comparison of classifier methods: a case study in handwritten digit recognition. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing, vol. 2, pp. 77–82. IEEE (1994)



The Influence of Media Types on Students' Learning Performance in a Role Playing Game on English Vocabulary Learning

Yuting Li and Jing Leng^(✉)

Department of Educational Information Technology,
East China Normal University, Shanghai, China
jleng@deit.ecnu.edu.cn

Abstract. The application of learning resources to education game not only enriches the content of education game, but also facilitates learners' personalized autonomous learning and satisfies the cognitive preference of learners with different foundation and learning resource preferences. The paper intends to identify the effects of the presentation of learning resources on students' the learning performance in a role playing game on English vocabulary learning. This study first tries to find out which learning resource media type was preferred by the second-year college students in the learning game. Then it examines the influence of different types of learning resources media on learners' learning outcome. Finally, the relationship between students' learning resource preferences and their chosen learning resources and media is analyzed. Results from the experimental study have shown that participants who choose text media perform better on vocabulary test on the whole than those who use other media. However, those who prefer visual learning material are proved to have a good mastery of difficult vocabularies. The study could provide a reference for the collocation of learning resources media in the digital game design on vocabulary learning.

Keywords: Learning resource media · Learning performance · Vocabulary learning

1 Introduction

The study intends to engage college students in English vocabulary learning by designing a Role Playing Game (to be referred as RPG for short). In RPG, the player plays the main character in the game to experience the complete storyline [1]. In order to increase the effectiveness of vocabulary learning, three types of media were provided: text, diagram and video. There are a number of studies addressing the effectiveness of learning with text, diagrams (static animations with simple paraphrase), animations, and other related media [2–6]. However, literature to date fail to indicate that an advantage of animations or videos over static media [7]. Also, limited research has been conducted on exploring which format of these media could effectively facilitate student learning outcome. Besides, Each student's learning style is also linked to learning, there is a lot of literature [8, 9] on learning resource preferences, few

studies even use experimental methods that can test the effectiveness of learning methods applied to education [10]. Therefore, this study designs a RPG for learning English vocabulary and tries to connect student preferences on media with their learning outcomes.

When presenting learning resources, there are many interference factors, which bring more cognitive load to students. Acquisition of learning resources is an active process. When students' autonomy is poor, they cannot complete the whole learning process smoothly. In addition, teachers cannot timely guide them, which will lead to low learning efficiency and even the failure of independent learning. Students' use of learning resources, therefore, is to learn, learning resource is the important foundation of autonomous learning can be smoothly, whether it's a good learning resources use, present system whether the way diversity, systemic, even in the process of learning are convenient communication influence the autonomous learning smoothly. With the continuous development of information technology, the combination of education and information technology is getting closer and closer. RPG is a teaching and learning system of perfect combination of information technology, it not only developed by information technology means real situational funny games, and will learn knowledge contained in it, fun, interactive, both instructive and interesting, can better promote the learners' learning and memory, let learners immerse inside. In RPG games, how to present learning resources is the key that game designers pay attention to. How to present learning resources can be accepted and liked by learners, and how to set up integrated learning resources can improve learning efficiency of users. Learners with different learning resource preferences have different information perception tendencies. In order to meet their different information perception needs, they need to integrate design learning resources by means of text, images, video, animation and other forms of expression. This article embarks from the four types of learning style [10], to explore the learners in learning English word RPG game education learning resources in the media types preference tendency, hope that through data analysis to explore the optimum learning resources media presentation, to the future of learning and game fusion research provides some effective reference, with abundant learning resources application, so as to provide more efficient access to learning resources for learners of the media.

2 Literature Review

2.1 Learning Style

Learning style refers to learners' continuous and consistent learning style with personality characteristics and learning tendency. Reid [11] classifies learning resource preferences into three categories according to the nature and expression of learning resource preferences. Felder (et al.) classify learning resource preferences from five dimensions of information perception, input, organization, processing and understanding, namely perceptive or intuitive, visual or linguistic, inductive or deductive, active or contemplative, sequential or holistic. However, when it comes to measurements, whether instruments are valid and reliable or not should be considered [12].

This experiment to explore the type of learning style is according to the five groups of eight types of learning resource preferences selected four types of learning style, they represent learning style information input and processing of two dimension, more in line with this article explores the learning resources access way of media types to choose, more conducive to the learner learning process according to the analysis of the behavior. The four learning resource preferences selected are: active, reflective, visual, and verbal.

2.2 Learning Resources Media

Various learning resources can promote the independent learning of different basic learners. Can adapt to the self-learning of people with different cognitive preferences; It can also reduce the time cost of autonomous learning; It can provide a basis for teachers to guide students to learn independently. Relevant literature takes six learning resources media as the research subject, including specifically digital documentation, slide show with annotated, animation, film clip, recording clip, interactive multimedia courseware. Based on the perceptive learning style, an analysis of the types of learning resources media from the perspective of sensory media shows that students' learning style influences their preference for learning resources in different media forms.

2.3 Role Play Game

Education game is a kind of computer game software that has education meaning and can cultivate game users' understanding of love. The RPG (Role Play Game) is a role-playing game in which players experience the main characters in the game and complete the story. Relevant literature shows that RPG games have the feature of story, which makes them distinguish from other games. RPG games will set colorful story-lines, in which learning knowledge is integrated in a scientific way, and learners solve learning problems in the game through inquiry. RPG games make players immerse themselves in the game world full of teaching target knowledge with excellent story plots. At the same time, RPG games will combine the relevant knowledge content in the teaching objectives with the story plot in a scientific and careful way, which will be more attractive for learners to have the learning experience in the game scenario. The experiential learning mode of RPG games adopted in this experiment can provide a good situation learning experience platform for English word learners. The learning resource media set in the game is integrated with the plot of the story, which can highlight the multimedia, personalized and interactive features of education games.

3 Method

3.1 Context

The research content is mainly divided into two points. First, based on RPG education game, explore whether learners with different learning resource preferences tend to choose different learning resources and media. Second, in the context of RPG games, is

there any difference in the choice of learning resource media for the learning of two groups of English words with different difficulty levels? If there is any difference, what effect will the selection of different types of learning resource media have on the learning effect? The types of learning resource preferences involved in the study were active and reflective, visual and verbal.

3.2 Participants

The subjects were 30 second-year college students. For the selection of experimental subjects, the following considerations were taken into account: the freshman just triumphed in the battle field of college entrance examination, with a large number of English words stored in his mind and a high enthusiasm for learning; Junior and senior students are too long away from learning English words, which makes learning difficult. Therefore, the second-year students between them are more suitable for real English word learning.

3.3 Design

The second-year learners learned English words in a designed RPG game. In the game, six little girls were set. Each little girl represented an English word that the learners had to learn. The task of the subjects is to control the characters' walking and talking to each little girl through the upper and lower keys, and press enter or the space bar to have a dialogue with the characters (see Fig. 1). The second-year college students who participated in the experiment chose learning resources and media for learning and completed corresponding exercises. When an option is encountered, you can press up and down to browse the option, and press enter or the space bar to select it. At the end of the six words, the learner has to complete 18 questions, each of which corresponds to three questions of increasing difficulty. The experiment was divided into experimental group and control group. In the experimental group, 15 subjects randomly selected learning resources and media to learn English words, and then completed the exercises after class. In the control group, 15 subjects were asked to choose only video media to learn English words, and then complete the corresponding exercises. The behavior data of the participants in the RPG education game will be collected and sorted (see Fig. 2).

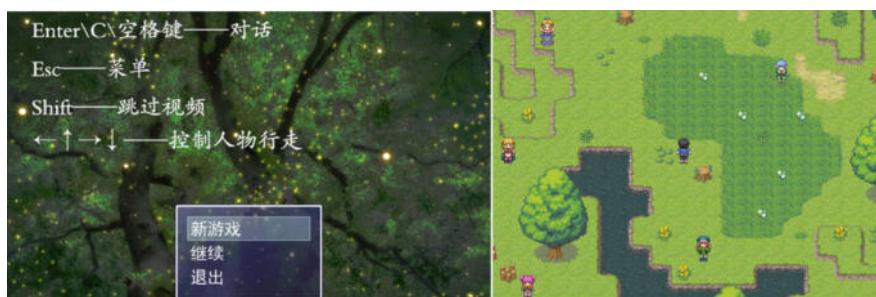


Fig. 1. The screenshot of RPG educational English vocabulary learning game

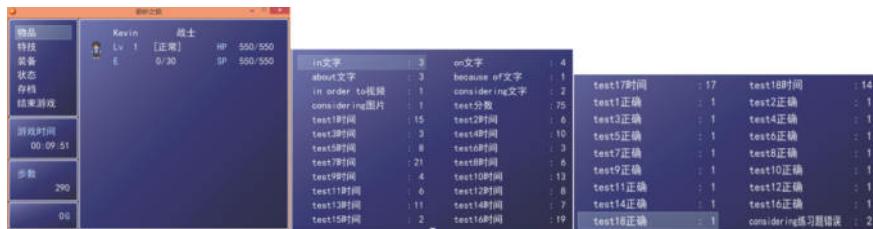


Fig. 2. The test results and test time of participant in RPG (as an example)

As well as the data of the test results after class, it is convenient to compare and deal with the learning style data collected in the questionnaire, which is conducive to further exploration and analysis.

3.4 Method

Empirical analysis. In this study, participants were observed to learn English words in an RPG education game environment. By analyzing the learning behavior in the learning process and combining the methods of questionnaire and interview, real and reliable behavioral data of learners are collected. On this basis, this paper combines with literature analysis method to analyze and discuss relevant contents. Control group was set up to collect, summarize and analyze the experimental data.

4 Results

4.1 Students' Preferences

Figure 3 depicts students' preferences of media as reflected in the questionnaire. There is an inconsistency between learner's learning media tendency in questionnaire and their choice in the RPG learning environment. The results indicate that learners prefer to use video to learning English vocabulary than the other two formats (44.33%). However, in RPG learning context, most participants selected text (60%) to learn vocabulary, only 13.33% of participants chose video. An interview was then conducted to further investigate students' thoughts on the effects of media on English vocabulary learning and their feedback on formats of learning materials. Some participants commented that it was difficult for them to learn English vocabulary with video media, and they feel anxious especially when learning vocabularies with high difficulty level. Meanwhile, a number of other participants suggested that using video format makes learning English fun and effective. As a result, many students went directly to text format when given more than one media format recommendations. It is suggested that it is the text format could keep students' focus on the learning content. But videos and diagrams would shifts the focus of students' attention from the learning content to the additional information and conversation around which that content is situated. So this can partly explain why the students were less likely to select video and diagram format to learn vocabulary.

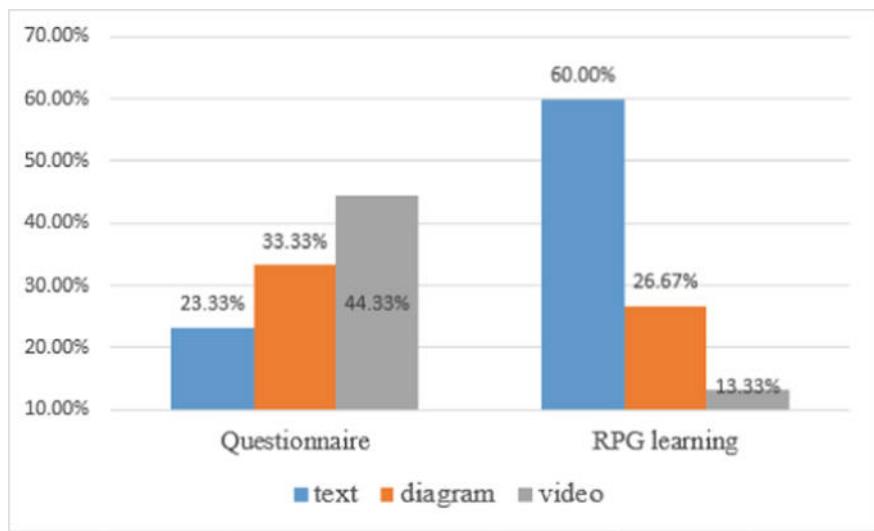


Fig. 3. Learner's learning media tendency in questionnaire and RPG learning context

This phenomenon prompts the author to think deeply, why the percentage of willingness to choose video media for learning in the questionnaire survey is the highest, and the percentage of English word learning in the way of text interpretation is the lowest among the three learning resources media given. However, in RPG, learners' behavioral tendencies are contrary to the results of the questionnaire survey (see Fig. 3). In view of this phenomenon, the author repeatedly observed the video of the recorded participants in the learning process, and found that 46.7% of the participants would choose video media at least once for English word learning in the RPG education game. 10% of the participants will choose two or more learning resources video media for English word learning. For this phenomenon, we conducted follow-up interviews with participants, and now sorted out their explanations of this behavior as follows: among the three media, the text media is the least, and is more consistent with the daily learning method, which is more conducive to the understanding of English vocabulary learning. The content of text interpretation is compact, concise and clear, which makes learners feel clear at a glance. However, the diagram and video are slow and time-consuming, which makes learners more likely to be distracted by the slow progress in the learning process, with low concentration of energy, and affects learning efficiency.

4.2 Vocabulary Test of Two Groups Students with Different Media in the RPG Learning Environment

Besides, based on the students' choice of media, thirty participants were then invited and divided into two groups to complete a vocabulary test in the RPG learning environment. Participants of one group ($N = 15$) can choose the learning media freely during the RPG vocabulary learning process, while students in another group ($N = 15$)

were allowed to select video media only. The total score of the vocabulary test in the RPG learning environment is 18. The average score of the group in which participants could choose learning media freely was 13.2, and the average score in another group was 11.8. The result indicates that compared with learning with video media only, giving learners more choices on media type could effectively foster their learning performance in RPG learning environment.

Figure 4 shows the comparison of the two groups' average score for each question of three difficulty levels. There are some differences among the average score of each question of the two groups under the three difficulty levels'. As a general rule, students score higher on easier questions. Same thing happened with the free media group. However, students in the video group scored unexpectedly well on difficult questions.

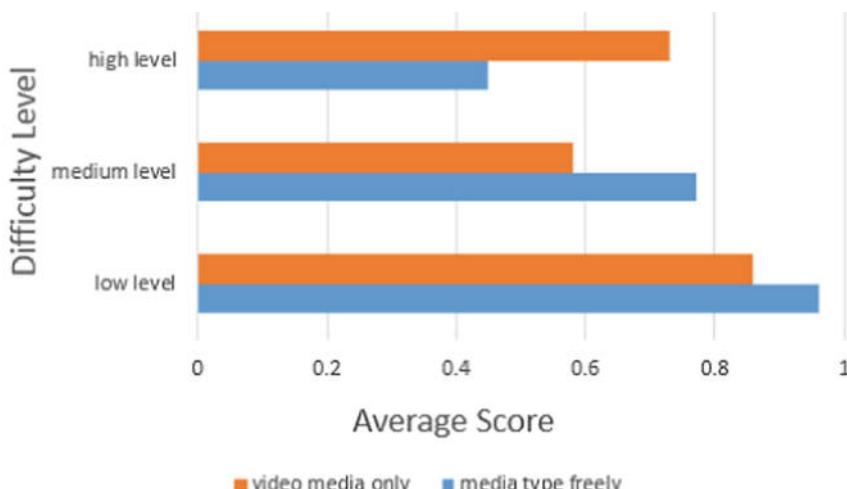


Fig. 4. The average score of two groups under three difficulty levels

4.3 The Media Preference of Different Participants' Learning Style

In another aspect, the survey of learning style shows that visual learners are strongly inclined to choose the learning media with picture interpretation when learning English words (see Fig. 5). They prefer to choose the media with text interpretation to present learning resources, and show curiosity about video media. Learners with visual learning resource preferences tend to be visual. In the process of learning English vocabulary, they also show curiosity about video media, but generally prefer learning resource provided by diagram. Both types of participants belong to visual learning resource preferences, but with slight different degrees of inclination. When learners' learning resource preferences tend to be verbal, participants almost all choose text

media to acquire learning resources. In particular, participants with a balanced learning style in visual and verbal types also have a balanced choice of text and diagram, and their differences in video media's selection are not obvious.

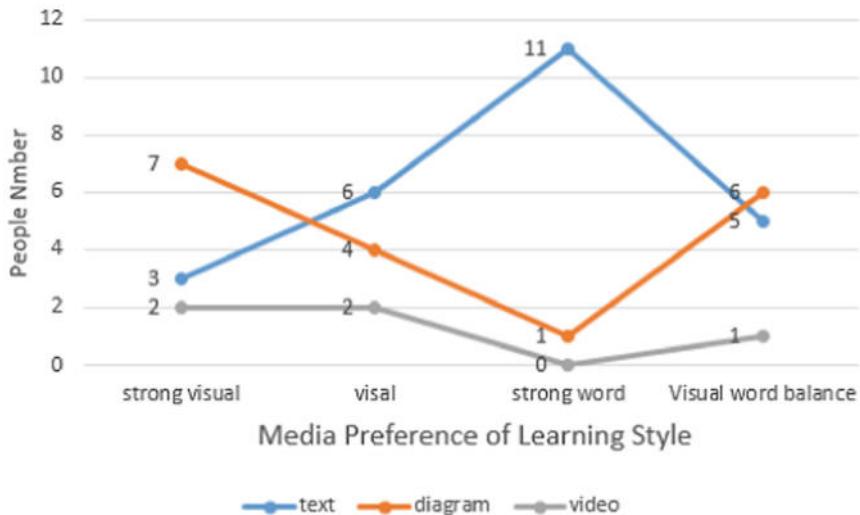


Fig. 5. The media preference of different participants' learning style

5 Discussion

According to the related cognitive theory [13], video media is not easy enough for learners to gain working memory due to a significant extraneous cognitive load [14]. This helps explain why most participants (60%) prefer to choose text media to learn English vocabulary rather than video (13.33%) in this study. The video media in the RPG was designed with some essential prior knowledge structures to optimize learning resources. In order to understand the main information in the video, learners need to spend more time and effort on watching the entire video and then extracting the contents on the vocabulary. This leads to more cognitive energy and reduces learning outcome relatively, especially for students of poor command of English. Therefore, although it is reported that visual nature and audible content of video is beneficial for understanding key concepts, the flexibility and selectivity of the media format is also of great importance. In the teaching, the students' learning style should be evaluated [15], with the experiments and analysis of follow-up interviews, it can be concluded that, with the enhancement of visual orientation, learners will have more preference for learning media such as picture presentation of learning resources in learning, and less preference for learning media that use text presentation of learning resources. However, there was no significant change in video media selection among visual learners of different degrees. This study suggests that just offering one media format such as video

is not a guarantee to good learning results. More media choices should be provided to learners in an online learning environment. For students with different learning resource preferences, the best learning media and methods, for example, which video meets students' learning needs best or which diagram show learning resource best need to be further explored.

References

1. Kolb, D.A.: Experiential learning experience as a source of learning and development. *Coll. Teach.* **1**, 16–17 (1984)
2. Mayer, R.E., Gallini, J.: When is an illustration worth ten thousand words? *J. Educ. Psychol.* **83**, 715–726 (1990)
3. Mayer, R.E., Sims, V.K.: For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *J. Educ. Psychol.* **86**, 389–401 (1994)
4. Mousavi, S.Y., Low, R., Sweller, J.: Reducing cognitive load and visual presentation mode. *J. Educ. Psychol.* **87**, 319–334 (1995)
5. Mayer, R.E., Moreno, R.: A split-attention effect in multimedia learning: evidence for dual processing systems in working memory. *J. Educ. Psychol.* **90**, 312–320 (1998)
6. Tindall-Ford, Chandler, & Sweller (1997)
7. Hegarty, M., Kriz, S., Cate, C.: The roles of mental animations and external animations in understanding mechanical systems. *Cognit. Instr.* **21**(4), 325–360 (2003)
8. Coffield, F., Moseley, D., Hall, E., Ecclestone, K.: Learning resource preferences and pedagogy in post-16 learning: a systematic and critical review. Learning and Skills Research Centre, London (2004)
9. Khacharem, A., Zoudji, B., Ripoll, H.: Effect of presentation format and expertise on attacking-drill memorization in soccer. *J. Appl. Sport Psychol.* **25**(2), 234–248 (2013)
10. Felder, R.M., Silverman, L.K.: Learning and teaching styles in engineering education. *Eng. Educ.* **78**(7), 674–681 (1988)
11. Reid, J.M.: The learning style preferences of ESL students. *Tesol Q.* **21**(1), 87–111 (1987)
12. Caple, J., Martin, P.: Reflections of two pragmatists: a critique of honey and mumford's learning resource preferences. *Ind. Commer. Train.* **26**(1), 16–20 (1994)
13. Sweller, J., Ayres, P., Kalyuga, S.: Cognitive load theory. Springer, New York (2011)
14. Hmoudova, D.E.: The impact of learning style dimensions on computer-based key language competences testing. *Proc Soc Behav Sci* **82**(1), 411–416 (2013)
15. Pashler, H., McDaniel, M., Rohrer, D., Bjork, R.: Learning resource preferences concepts and evidence. *Psychol. Sci. Public Interest* **9**(3), 105–119 (2008)



Edit Distance Kernelization of NP Theorem Proving For Polynomial-Time Machine Learning of Proof Heuristics

David Windridge^(✉) and Florian Kammüller

Middlesex University London, London, UK
`{d.windridge|f.kammueler}@mdx.ac.uk`

Abstract. We outline a general strategy for the application of edit-distance based kernels to NP Theorem Proving in order to allow for polynomial-time machine learning of proof heuristics without the loss of sequential structural information associated with conventional feature-based machine learning. We provide a general short introduction to logic and proof considering a few important complexity results to set the scene and highlight the relevance of our findings.

Keywords: Machine learning · Theorem proving · Kernel methods

1 Logic and Proof—Syntax, Semantics, and Proof Systems

Theorem proving is an attempt to provide machine support for logic and proof. We need logic for problem solving; if we have very small sets of states, we can potentially solve problems just by enumerating all possibilities. Logic, however, provides a convenient way of dealing with larger (potentially infinite) sets of states via the manipulation of compact sentential descriptions instead of large sets of states. These manipulations are made possible by providing *syntax*, *semantics* and a *proof system* for a logic. The syntax defines what constitute legal sentences, the semantics says what they mean, and the proof system allows to syntactically change logic expressions to provide new insights. The proofs allow us to make conclusions about the world in a given state from given percepts and also to conclude properties of the next state given additional state operators (formulas usually contain variables that can be generalized to more complex states). The properties of such a logic description then depends on their interpretation, i.e., the values these variables have. The semantics of a logic formula is usually one of the truth values *true* or *false*, but it may depend on the interpretation of variables contained in the formula. We say a formula is *valid* if it evaluates to *true* for all possible interpretations; we say it is *satisfiable* if there is an interpretation such that it becomes true for inserting these values into the variables, and it is called *unsatisfiable* if there is no interpretation that makes the formula *true*.

Decision Problems Satisfiability (SAT) solver and Satisfiability Modulo Theory (SMT) solvers look at satisfiability of logic formulas; the latter one with respect to an additional axiom system (also called *theory*). Satisfiability problems occur frequently everywhere in the form of constraint systems, for example, scheduling problems. In general, we are often interested in *validity*. That is, we want to know the correspondence of entailment between a set of formulas Γ and a formula ϕ . Entailment is written as

$$\Gamma \models \phi.$$

This statement means that for every interpretation of the variables in Γ the formula ϕ is also *true*. In other words, there is a subset relation between interpretations between the formula ϕ and the set of formulas Γ : $I_\phi \subseteq I_\Gamma$. This, however, is equivalent to the statement that the logical implication between them is *valid*:

$$\models \Gamma \longrightarrow \phi$$

i.e., this implication is true for all interpretations of variables.

Proof Systems Proof is a way of determining validity without examining all interpretations which satisfy a formula. Proofs are commonly written in a notation similar to entailment. The formula

$$\Gamma \vdash \phi$$

means ϕ can be proved from Γ for a given proof system associated with \vdash .

A proof system is a set of so called *inference rules* and a way to apply those to sets of formulas Γ that allows to transform statements, that is, *infer* a statement from a previous one. We say that a proof system is *correct* (or sound) iff

$$\Gamma \vdash \phi \implies \Gamma \models \phi$$

and we say that it is *complete* iff

$$\Gamma \models \phi \implies \Gamma \vdash \phi.$$

Natural Deduction There are different styles of proof systems. The proof system of *Natural Deduction* uses a style of proof that suits human understanding. For example, the rule of *modus ponens* allows to infer Q from the set of premises $\{P, P \longrightarrow Q\}$ written as

$$P \quad P \longrightarrow Q \vdash Q.$$

This could also be written in a two-dimensional style thus creating a tree like structure for proofs that visualises the proof structure, i.e., chaining up of rules. For example, consider the proof of $P \wedge Q \longrightarrow Q \wedge P$.

$$\frac{\frac{[A \wedge B]}{B} [conjE_{\text{left}}] \quad \frac{[A \wedge B]}{A} [conjE_{\text{right}}]}{B \wedge A} [conjI]$$

$$\frac{B \wedge A}{A \wedge B \longrightarrow B \wedge A} [impI]$$

In each step, the items above the lines are the premises specified by the inference rule. The rules name is given in square brackets at the side, for example, $conjE_{\text{left}}$ for the rule $P \wedge Q \vdash Q$. The last step of the proof uses the rule $impI$: $[P] \vdash Q \vdash P \longrightarrow Q$ to eliminate the assumption $A \wedge B$ by adding it as a premise to the current formula. This elimination of premises is called cancellation syntactically indicated by the square brackets around the cancelled assumption. Note that it is possible to cancel various occurrences of the premise at once and that the cancellation may reach across several proof steps as shown in the example.

Resolution Refutation A proof system that is different in style but more suitable for the implementation of automated proving on computers is that of resolution refutation [12]. It uses the resolution rule

$$P \vee Q \quad Q \longrightarrow R \vdash P \vee R$$

which can also be written as

$$P \vee Q \quad \neg Q \vee R \vdash P \vee R.$$

Resolution refutation works as follows. Given a set of assumptions Γ and a conclusion ϕ , we prove $\vdash \Gamma \longrightarrow \phi$ by the following procedure.

1. Transform each premise $\gamma \in \Gamma$ into Conjunctive Normal Form (CNF), (i.e., a conjunction of disjunctions, e.g. $(P \vee Q \vee \neg R) \wedge (\neg P \vee R)$), obtaining Γ' as the set of conjuncts.
2. Conjoin the negated conclusion $\neg\phi$ to Γ' transformed into CNF.
3. Apply repeatedly the resolution rule to Γ' extending Γ' by adding the conclusion.

This procedure either terminates by reaching a contradiction in Γ' by having P and $\neg P$ in the set, or by reaching a Γ' such that the resolution rule cannot be applied any more to extend Γ' . In the former case, we have proved

$$\Gamma \quad \neg\phi \vdash \text{false}$$

which is equivalent to having proved $\Gamma \vdash \phi$ in classical logics. In the latter case, we have shown that we could not prove ϕ using the resolution refutation procedure. For propositional logic, refutation resolution is a complete procedure. Hence, we know that in the latter case if ϕ cannot be proved from Γ using refutation resolution, then it is also not entailed in Γ .

Since transformation of a set of formulas Γ into CNF is a simple algorithm and applying just the one resolution refutation rule defines the whole process of

resolution, it seems intuitively clear how it could be implemented as a decision procedure on a computer.

We illustrate the resolution refutation procedure on the example that we have used for natural deduction. To prove

$$\vdash A \wedge B \longrightarrow B \wedge A$$

we apply Steps (1) and (2) negating this conclusion ϕ and adding it to the empty set of premises.

$$\neg(A \wedge B \longrightarrow B \wedge A).$$

Next, we transform this (set of) premises into CNF by translation the implication and applying de Morgan's laws.

$$\begin{aligned} \neg(\neg(A \wedge B) \vee (B \wedge A)) &\equiv \\ (\neg\neg(A \wedge B)) \wedge \neg(B \wedge A) &\equiv \\ A \wedge B \wedge (\neg B \vee \neg A). \end{aligned}$$

We then apply Step (3), i.e., the resolution rule once and add the conclusion.

$$A \wedge B \wedge (\neg B \vee \neg A) \wedge B \wedge \neg B.$$

Finally, we have the contradiction $B \wedge \neg B$ in the set of conjuncts and have thus proved ϕ .

Decidability of Proof Systems As we have seen above, resolution refutation is a complete and sound procedure for propositional logic. Gödel's Completeness Theorem proves that there exists a complete proof system for First Order Logic (FOL) but it is only later that Robinson showed a more constructive version proving that resolution refutation is a complete proof system for FOL. This resolution refutation for FOL is more complicated than the one for propositional logic as it has to deal with quantifiers. As a consequence, provability, i.e. $\Gamma \vdash \phi$, is only semi-decidable for FOL. That is, if we apply resolution refutation to a formula ϕ and a set of premises Γ such that ϕ is actually entailed by ϕ , then it will arrive at a contradiction for Γ and $\neg\phi$, i.e. find the proof. But if ϕ is not entailed in Γ , the resolution refutation might not terminate. By contrast, resolution refutation terminates for propositional logic in both cases, i.e., provability for propositional logic is decidable.

If we consider more expressive logics than FOL, we have to consider Gödel's Incompleteness Theorem: there is no complete and consistent (sound) proof system for FOL if Arithmetic is added. So for any potential proof system there would be either a true statement it cannot prove (incomplete) or it would be possible to prove a false statement (inconsistent). Intuitively, the proof works by assuming an arbitrary proof system for FOL with Arithmetic, then showing that one of the two is the case. Arithmetic provides the means to construct code names for sentences in the logic (using a procedure called “Gödel-numbering”) and therefore construct sentences that are self-referential. The sentence used in

the proof is sometimes called the Gödel-sentence $G = "G \text{ is not provable}"$. Now, if G would be true then it is not provable by definition and therefore the proof system would be incomplete. If G , however, would be false, then G is provable which means the proof system is not sound (inconsistent).

Having no complete and sound proof system for FOL with Arithmetic implies that there is also no complete and sound procedure to implement automated proofs on computers. Consequently, any logics that can formalise Arithmetic cannot be adequately automated. This includes all Higher Order Logics. Yet, theorem provers for Higher Order Logics exist, e.g., Coq or Isabelle, but they are interactive, that is, need human intervention. And even with human intervention, the proof systems they implement must be incomplete (if not inconsistent as one would hope). Surprisingly, Gödel's Incompleteness Theorem itself could be proved in Isabelle [10]. So, these logics are not so incomplete as one might think although the proof document containing the transcript of the interactions with Isabelle in the Archive of Formal proof has more than 200 pages [9]. Also, incompleteness is still a great loss since it says that the proof system will never reveal all truth. However, the main motivation for interactive proof in HOL is to provide a means to have sound proofs. Although traditionally mathematical proof is a social process performed by humans in a community of scientists there are cases where the human capacity for guaranteeing the consistency of proofs is reached and where formalisation and proof in interactive theorem provers is necessary to provide acceptable proofs. Examples are the proof of the Four Colour Theorem in Coq [6] and the proof of Kepler's conjecture in HOL-light [7].

Computational Complexity of Propositional Logic Proof Systems If we want to investigate the feasibility of theorem proving, we need to describe the computational complexity of decision procedures for (provability of) logical statements precisely. This only makes sense for decidable questions since for undecidable or semi-decidable proof procedures we could only ever compare efficiency relative to some varying portion of the problem. Therefore, complexity theory focuses only on propositional logic because even for FOL we already have no decision procedure: as we have seen above, the complete and sound procedure of resolution for FOL is only semi-decidable.

It appears from the literature that the complexity of propositional proofs is a well studied subject and still many questions remain open. There are some very remarkable relations to general complexity problems.

Let Σ be a finite alphabet and Σ^* denote the set of all finite strings over it. A language L is defined as a subset of Σ^* , that is, a set of strings. The length of a string s is written as $|s|$. Let us first recall the basic definition of computational complexity theory. We adopt the definitions of [4, 13] adapting them slightly. A set of strings L is in the class P (NP) if it is recognized in time polynomial in the length of the input by a deterministic (non-deterministic) Turing machine. A set of strings L is in the class co-NP if L is the complement $\Sigma^* \setminus \hat{L}$ of a language \hat{L} that is in NP. The following set \mathcal{L} defines characteristic functions for the elements of class P and thus makes the notion of “polynomial-time function” more precise.

Definition 1. (*Polynomial function class*) \mathcal{L} A function $f : \Sigma_1^* \rightarrow \Sigma_2^*$ for finite alphabets Σ_1, Σ_2 is in \mathcal{L} , if it can be computed by a deterministic Turing machine in time bounded by a polynomial in the length of the input.

To compare the efficiency of propositional proof systems, we need a general yet abstract definition of proof system to encompass the existing variety of existing systems.

Definition 2. (*Abstract Proof System*) Let $L \subseteq \Sigma^*$ be a language. A proof system for L is a function $\mathcal{P} : \hat{\Sigma}^* \rightarrow L$ for some alphabet $\hat{\Sigma}$ such that \mathcal{P} is in L and onto. A proof system \mathcal{P} is *polynomially bounded* if there is a polynomial p such that for all $A \in L$ there is a $\pi \in \hat{\Sigma}^*$ such that $\mathcal{P}(\pi) = A$ and $|\pi| \leq |p(\pi)|$.

The language L will be the set of all valid propositional formulas. The idea of this definition is that $\mathcal{P}(\pi) = A$ if π is a proof for A in the proof system \mathcal{P} . The special property of a polynomially-bounded proof system as defined above is that there is a feasible function \mathcal{P} that checks whether a potential proof π of a property A is a proof in that proof system and that it is a proof of property A .

To illustrate this on our running example, we encode the resolution proof of $B \wedge A \longrightarrow A \wedge B$ in one string

$$[A \wedge B \wedge (\neg B \vee \neg A), B \wedge \neg B, \text{false}].$$

The function \mathcal{P}_{res} implementing a proof system for resolution refutation checks that this string represents a sequence of resolution refutation steps according to the algorithm sketched above. In addition, it re-transforms $A \wedge B \wedge (\neg B \vee \neg A)$ into $\neg(A \wedge B \longrightarrow B \wedge A)$ and by deleting again the negation to check that it is a proof of the property $A \wedge B \longrightarrow B \wedge A$. These steps can be clearly done in time polynomial to the length of the input.

A problem p is called NP-complete, if it is at least as hard (complex) than any other NP problem. That is, if every other problem in NP can be transformed (or reduced) into p in polynomial time. If a polynomial solution to any NP-complete problem would be found, all other NP problems would also follow to be in P thereby proving $P = NP$. The idea of completeness transfers also to the class coNP. Simply, every coNP-complete problem is the complement of an NP-complete problem.

A celebrated result by Cook and Reckhow [2] shows that *SAT solving for propositional logic is NP complete*.

From this theorem follows a result on propositional logic [4].

Corollary 1. *Validity for propositional logic is coNP-complete.*

This corollary immediately implies another one [4].

Corollary 2. *Validity for propositional logic is in P if and only if P = NP.*

This is relevant for the implications of the work we are going to present next.

2 Machine Learning Theorem Proving

Theorem proving/formal verification is, for the most part (i.e. the majority of axiom systems deployed in the field), NP complete. Any polynomial-time heuristic system for reducing the search space involved in proving a given theorem that performs better than chance can therefore be effectively utilized for increasing the efficiency of theorem-proving/formal-verification.

Such heuristics are generally provided manually. However, an alternative is to learn these via the techniques of *machine learning*. The typical machine learning paradigm is *supervised classification* in which discretely-labeled (i.e. class-delineated) data of arbitrary kinds are represented in a feature space of potentially very large dimensionality, within which an optimal class decision boundary is determined via an appropriate optimization process (such that arbitrary unlabeled data can be attributed to one of the training classes).

Syntactically-valid sentences in a formal language may thus seem superficially well suited a machine learning approach in that they consist of a finite set of labeled data, for which the attribution of a label is a difficult, potentially non-polynomial process (the labels in question being the True/False values of the theorem prover [or appropriate alternative discrete truth values]).

However, the set of syntactically-valid sentences does not naturally lend itself to a feature-based model. In particular, the potentially infinite variation of sentence length would suggest that sentences have an intrinsically arbitrary feature-dimensionality, and are thus not collectively embedded in the *same* feature space. (While it is possible to enforce a consistent dimensionality amongst sentences by, for example, using a normalized symbol histogram ('bag of words') approach [14], this would invariably constitute an information losing process).

In this paper we propose a 'featureless' polynomial-time approach to the learning of proof heuristics that eliminates feature space representation entirely and works directly on the individual sentences as given; the classification process thus takes place in an entirely implicit feature-space. This implicit feature space will turn out to have interesting characteristics; in particular it will provide a continuous Hilbert (or Krein) space in which the theorems exist. As well as providing an intriguing implicit discrete-to-continuous space mapping, continuity provides a range of useful properties for efficient optimization.

Our principle contribution is thus the outlining of strategy for NP-complete → polynomial heuristic mapping for formal-verification problems.

2.1 Adopted Machine Learning Paradigm

Support Vector Machine (SVMs) are archetypal *binary classifiers*, i.e. entities capable of learning an optimal discriminative decision hyperplane from labeled vectors $\{(\mathbf{x}, y) \mid \mathbf{x} \in \tilde{X}, y \in \{-1, +1\}\}$ existing within a feature space \tilde{X} .

SVMs are especially useful in classical machine learning because they have the capacity to be *kernelized*; that is, the gram matrix inherent in the (dual form of the) SVM optimization problem can be replaced by an arbitrary kernel function obeying the Mercer condition (essentially positive semi-definiteness -see later), enormously extending its capability.

Kernel functions thus constitute a form of similarity measure (specifically, a highly generalized inner product) between classification objects. Indeed they can be demonstrated to be the equivalent of inner products within an implicit embedding space (the Mercer space) generated by the kernel as a *feature mapping* of classification objects (which need not be directly computed in itself). This is enormously powerful in machine learning in that it enables classification to apply in areas in which there is not a readily apparent real vector space of feature measures. A relevant motivating example is genomics, for which it is much more straightforward to compute a similarity measure between pairs of DNA strands (using e.g. least common ancestor distance or mutation distance) than it is to embed each strand individually into a vector space of feature measurements. However, computation of kernels can be a complex, potentially NP-hard, exercise; we will thus, in the following section set out the specifics of SVM learning computation in order to show how it applies to the theorem-proving problem.

3 Methodology

In order to specify our strategy for kernel machine learning within a theorem proving context, we shall firstly give a general description of the support vector machine optimization problem:

3.1 The SVM and its Kernelization

The standard SVM [3] seeks to maximize the margin (i.e., the distance of the decision hyperplane to the nearest data point), subject to a constraint on the classification accuracy of the labeling induced by the hyperplane's delineation of a general decision boundary. In its primal form, the soft margin SVM optimization takes the form of a Lagrange optimization problem:

$$\arg \min_{(\mathbf{w}, b)} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^M \xi_i \right\} \quad (1)$$

subject to:

$$\forall i \ y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (2)$$

where (\mathbf{x}_i, y_i) $i = 1 \dots M$ are the training vectors/labels, $y_i \in \{-1, +1\}$, \mathbf{w} is the weight orientation vector of the decision hyperplane, and b is its bias offset. (The margin is inversely proportional to $\|\mathbf{w}\|$.) The ξ_i are slack variables that give rise to the soft margin with sensitivity controlled by hyper-parameter C .

In the dual form [3], the slack parameters disappear such that the problem is solved in terms of the KarushKuhnTucker (KKT) multipliers α_i :

$$\arg \max_{(\alpha_i)} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) \quad (3)$$

subject to:

$$\sum \alpha_i y_i = 0 : \quad \forall i \quad 0 \leq \alpha_i \leq C \quad (4)$$

The problem is one of quadratic programming. As the optimization proceeds, only a sparse set of the α_s s retain non-zero values. These denote the support vectors defining the decision hyperplane. This sparsity (i.e. the low parametric complexity of the decision boundary with respect to the training data) gives the SVM substantial resilience to over-fitting (and thus reduces classifier variance).

Notably, it may be shown that the term $(\mathbf{x}_i^T \mathbf{x}_j)$ in the above (equating to the training vector Gram matrix) may be freely replaced by any kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ that satisfies the Mercer condition (guaranteeing positive semi-definiteness). This is the principle (but by no means the only) use of kernel methods in machine learning, one which vastly extends the utility of the SVM by enabling the mapping of the input decision space into a large variety of alternative Hilbert spaces of potentially infinite dimensionality (thus guaranteeing linear separability). The decision boundary in the input space may thus undergo significant morphology variation while crucially retaining the low parametric support vector characterization of the decision boundary in the Mercer embedding space (the space defined by $\phi(\mathbf{x})$, where $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T(\phi(\mathbf{x}_j))$). Critically, at no stage are we required to compute $\phi(\mathbf{x}_i)$. The Mercer condition guarantees the existence of ϕ , but the kernel itself may be calculated based on any similarity function that gives rise to a legitimate (ie PSD) kernel matrix. (It is also possible to formulate the SVM problem in the explicit absence of similarity functions encompassing a guarantee of positive semi definiteness via the use of *Krein spaces*. However, in this case the SVM optimization problem is no longer strictly convex and requires an alternative (though often straightforward) saddle-point gradient descent process without guarantees of achieving the global minimum).

3.2 Edit Distance Kernels

In the terminology of Neuhaus and Bunke [8], a string t over V consists in an ordered, finite sequence of symbols drawn from V such that:

$$t = t_1 \dots t_n \in V^* = \bigcup_{i=0}^{\inf} V_i \quad \text{with } V_0 = \{\} \& n > 0 \quad (5)$$

where V_i is the set of strings of length i over V and V^* is the set of all finite sequences of symbols drawn from V . V is thus typically a finite set of symbols, but may equally represent vector spaces etc. This format can thus capture a wide range of sequential data within machine learning, from written text to DNA sequences.

The standardized set of string edit operations that can be performed on such a string consists in:

- (1) An insertion operation $\{\} \rightarrow q$ (inserting symbol q into a string)
- (2) A deletion operation $p \rightarrow \{\}$ (removing symbol p from a string)
- (3) A substitution operation $p \rightarrow q$ (exchanging symbol p in a string with symbol q).

Clearly, by recursive operation, any string can be transformed into any other string using just these operations. For example, to utilize our running example, we may obtain the final proof string “[$A \wedge B \wedge (\neg B \vee \neg A)$, $B \wedge \neg B$, *false*]” via a series of transformations of the initial string “ $A \wedge B \wedge (\neg B \vee \neg A)$ ” (which in the proof theoretic context corresponds to the application of the resolution refutation rule). Further type-dependency can be introduced into the above rules while retaining its character as an edit distance e.g. by including rules disallowing the deletion of negation operators.

In order to arrive at a kernel, however, it is necessary to obtain a *unique* measure of the edits required to relate arbitrary pairs of strings. Following [8], if we let $e(t, t')$ denote the set of all edit operation sequences that connect t to t' , then the *string edit distance*, $d(t, t')$, between the two strings is defined as the minimum cost required to edit t into t' :

$$d(t, t') = \arg \min_{(w_1, \dots, w_k) \in e(t, t')} \sum_{i=1}^k c(w_i) \quad (6)$$

where c is the positive real-valued *edit cost function*. The fact that this distance is parametrized means that, we in effect, have a *family* of distance measures (and later kernels).

Note, critically, that this is a true metric obeying the triangle inequality. Recursively evaluating this distance, however, takes exponential time; it is consequently computed within feasible machine learning scenarios via a dynamic programming approach that reduces the cost to approximately quadratic time (though provably not strongly subquadratic time [1]).

By contrast, the *graph edit distance* between discrete attributed graphs can be computed in essentially the same way as the general edit distance, with the exception that the nodes of a graph (whose connectivity may be represented via a matrix) have no intrinsic order to them, meaning that dynamic programming does not readily apply [5]). Thus (to extend the above terminology), the node substitution $u \rightarrow v$, replaces node u in a graph with node v ; an edge insertion $\{\} \rightarrow (p, q)$ on the other hand, inserts the edge connecting node p with node q into the graph.

Graph edit distance is hence similarly NP-complete in its native formulation to the general edit distance, however only non-upper-bounded heuristic processes are available to render it tractable [11].

It is thus the fact that theorems are one-dimensional strings of symbols that we propose to exploit; any reasonable (i.e. monotonic) variant of the string edit distance will retain its polynomial-time characterization in the theorem-proving domain (following the application of dynamic programming); thus any non-polynomial proof-theoretic problem can, in principle, be substituted by a polynomial-time Kernel matrix formulation problem (in conjunction with a polynomial $O(n^3 \log(n))$ SVM optimization problem); the attribution of a truth value to an arbitrary sentence is essentially linear once the learning process has taken place.

Applying these notions within a machine learning context thus involves construction of a kernel function between two string objects x and x' based on their edit distance with respect to a fixed pattern string x_0 :

$$k(x, x') = k_{x_0}(x, x') = \frac{1}{2}(d(x, x_0)^2 + d(x_0, x')^2 - d(x, x')^2) \quad (7)$$

In particular, we can guarantee the positive semidefiniteness of this kernel by virtue of the metricality of d , which thus obeys the Mercer condition.

Neuhaus and Bunke [8] go on to construct *sum* and *product* kernels with respect to the full set of fixed pattern strings I in the training set:

$$k_I^+(x, x') = \sum_{x_0 \in I} k_{x_0}(x, x) \quad (8)$$

$$k_I^*(x, x') = \prod_{x_0 \in I} k_{x_0}(x, x) \quad (9)$$

which both retain the properties of a kernel. (Indeed any arbitrary convex sum or product over kernels is also a kernel and we can thus treat the determination of the optimal kernel coefficients as a polynomial-time optimization problem in its own right, for instance when we are confronted with a family of kernels as in the parametrized edit distances above.)

Having thus obtained a kernel suitable for application to theorem proving, such that *any* logical sentence is guaranteed to exist within the Mercer space of the kernel, it becomes possible to apply machine learning (and, in particular, the SVM algorithm) in order to anticipate (i.e. predict) the outcome of the theorem prover in question on the basis of its previous outcomes (i.e. via an implicit labeling of previously derived sentences using the binary class labels *theorem/non-theorem*). Moreover, it does so in an adaptive, on-line manner such that the mechanism improves over time. It can thus be deployed alongside the theorem prover in a hybridized form in order to determine promising directions of application of the full machinery of theorem proving.

It is thus clear that although computing an edit-distance kernel constitutes an optimization problem in its own right (the problem being NP-hard in its recursive form without optimization, and irretrievably NP-hard in the case of graph edit distance kernels), the specific case applicable to theorem proving is always tractable in polynomial time.

Thus, provided that the domain exhibits learnability (which only requires that the learning domain is not fully topologically discontinuous), machine learning has the capability to provide a polynomial substitution for a non-polynomial theorem-proving problem (with a utility that is proportional to domain learnability).

4 Conclusions

We have provided a concise introduction to the theorem proving problem with respect to the issues of decidability and computational complexity before pro-

ceeding to demonstrate that *kernelized machine learning* is capable of providing a hybridized approach to reducing the theorem proving problem to a polynomial one with probabilistic success dependent on the intrinsic learnability of the domain. Critically, domain learnability is related to the metric proximity of similar data; a domain thus only exhibits non-learnability when its class labels are *entirely* discontinuous. There is no reason, however, to suppose that this extreme case applies in relation to theorem proving with sequential strings.

In terms of the state of the art on the theorem proving problem, more precisely Corollary 2, it may appear that we are thus claiming $P = NP$ in contradiction to the current conviction in the scientific community that $P \neq NP$. Far from trying to do so, we believe that our results presented in this paper are fully valid given that what we have presented is *heuristic* learning. Therefore, findings in relation to machine learning theorem proving (which is the subject of ongoing experimental work) need to be understood as approximate behaviours true in the average case.

Acknowledgements. The first author would like to acknowledge financial support from the Horizon 2020 European Research project DREAMS4CARS (number 731593).

References

1. Backurs, A., Indyk, P.: Edit distance cannot be computed in strongly subquadratic time (unless SETH is false). CoRR abs/1412.0348 (2014), <http://arxiv.org/abs/1412.0348>
2. Cook, S.A., Reckhow, R.A.: The relative efficiency of propositional proof systems. J. Symb. Log. **44**(1), 36–50 (1979). <https://doi.org/10.2307/2273702>
3. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
4. Das, A.: The Complexity of Propositional Proofs in Deep Inference. Ph.D. thesis, University of Bath Department of Computer Science (2014)
5. Fischer, A., Uchida, S., Frinken, V., Riesen, K., Bunke, H.: Improving hausdorff edit distance using structural node context. In: International Workshop on Graph-Based Representations in Pattern Recognition, pp. 148–157. Springer, Berlin (2015)
6. Gonthier, G.: Formal proof—the four color theorem. Not. Am. Math. Soc. **55**(11), 1382–1393 (2008)
7. Hales, T.C., Adams, M., Bauer, G., Dang, D.T., Harrison, J., Hoang, T.L., Kaliszyk, C., Magron, V., McLaughlin, S., Nguyen, T.T., Nguyen, T.Q., Nipkow, T., Obua, S., Pleso, J., Rute, J.M., Solovyev, A., Ta, A.H.T., Tran, T.N., Trieu, D.T., Urban, J., Vu, K.K., Zumkeller, R.: A formal proof of the kepler conjecture. CoRR abs/1501.02155 (2015), <http://arxiv.org/abs/1501.02155>
8. Neuhaus, M., Bunke, H.: Edit distance-based kernel functions for structural pattern classification. Pattern Recogn. **39**(10), 1852–1863 (2006). <http://www.sciencedirect.com/science/article/B6V14-4K48N7S-4/2/1e7743302ffe0f0662da24f14c7d5a8f>
9. Paulson, L.C.: Gödel's incompleteness theorems. Arch. Formal Proofs (2013). <http://isa-afp.org/entries/Incompleteness.html>, Formal proof development

10. Paulson, L.C.: A mechanised proof of gödel's incompleteness theorems using nominal isabelle. *J. Autom. Reasoning* **55**(1), 1–37 (2015). <https://doi.org/10.1007/s10817-015-9322-8>
11. Riesen, K., Fankhauser, S., Bunke, H.: Speeding up graph edit distance computation with a bipartite heuristic
12. Robinson, J.A.: A machine oriented logic based on the resolution principle. *J. ACM* **12**(1), 23–41 (1965)
13. Urquhart, A.: The complexity of propositional proofs. *Bull. Symbolic Logic* **1**, 425–467 (1995)
14. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 977–984. ACM (2006)



Performance Analysis of Artificial Neural Networks Training Algorithms and Activation Functions in Day-Ahead Base, Intermediate, and Peak Load Forecasting

Lemuel Clark P. Velasco^(✉), Noel R. Estoperez,
Renbert Jay R. Jayson, Caezar Johnlery T. Sabijon,
and Verlyn C. Sayles

Premier Research Institute of Science and Mathematics, Mindanao State University-Iligan Institute of Technology, Iligan City, The Philippines
lemuelclark.velasco@g.msuiit.edu.ph

Abstract. Artificial Neural Networks (ANN) has been highly utilized in short term electric load forecasting not just among aggregated consumed load but also in predicting the specified base, intermediate and peak loads. To ensure success in its predictive capability, every ANN model implementation should employ the appropriate training algorithm and activation function that will be suitable to the historical data that it is processing. This study conducted performance analysis of six models having different combination of training algorithms namely Quick Propagation, Resilient Algorithm and Back Propagation and activation functions namely Gaussian and Sigmoid. Electric load data preparation was conducted through data correction, Min-Max data normalization and clustering to identify the base, intermediate and peak loads. After determining the ANN models' input, hidden and output neurons from its respective layers, the ANN model having the combination of Quick Propagation training algorithm and Gaussian activation function yielded the lowest MSE and MAPE values having 0.005700397 and 5.88% respectively. The day-ahead base, intermediate, and peak load forecasting model developed in this study has the potential to be implemented in order to suffice the need of electric power systems in predicting the necessary system loads for their economic decisions, power dispatching, system planning, and reliability evaluation.

Keywords: Artificial neural network · Electric load forecasting · Base intermediate and peak load

1 Introduction

Artificial Neural Network (ANN) is a computational model which derives its machine learning functionality based on the behavior of biological neural systems. It consists of an interconnected group of artificial neurons and information processes which employs pattern recognition and function approximation using a connectionist approach to computation [1–4]. The ANN's model implementation schemes can vary greatly on its

applicability and depends on model architectural choices such as the number of neurons, the training approach and the activation function used for each neuron [1, 4, 5]. Due to the nature of ANNs having input layers, hidden layers and output layers, varied ways of network learning are executed from the input data up until its associated neural networks output [5, 6]. Just like biological neurons found in the human brain which received electrochemical signals from various sources, artificial neurons are fired depending on the network architecture and configuration of its training algorithm and activation function. More than the type of data with its intended application and its associated data preparation technique, performance analysis of neural networks play an important role in implementing ANNs [2, 4]. It is in the ANN model consisting of not just training algorithm and activation function but also the necessary configurations like the epoch, maximum error, learning rate and momentum which determines the success in the applicability of the supervised ANN in a particular problem domain that it tries to solve.

Commonly used among ANN architectures are multilayer perceptron neural networks, a universal approximator feed forward architecture which transports signals and adjusts weights in a one way manner from input, hidden and output neuron [7]. It is a well-known perceptron based technique that can solve problem such as set of instances which cannot be separated linearly that could cause the learning and classification process erroneously done. It is the frequently used neural network structure in load forecasting models that seems to offer the best possible performance [3, 8, 9]. In the field of power management, multilayer perceptron neural networks has the promising ability to predict electric load of a certain locality which can be helpful inputs for decisions involved in the functions of power generation, transmission and distribution. For day-ahead electric load forecasting, researchers usually aggregate the consumed electric load of the geographic locality in the data preparation phase in order to come up with a data set ready to be fed into the multilayer perceptron's input layer [1–3, 9]. Although developing an ANN model which generates short-term, medium term and long term forecast on consumed load have helpful applications in the operations and maintenance of power utilities, a challenge goes beyond the aggregated load input and forecasted load output in order to come up with sound analysis of a locality's load consumption. As stated by a group of researchers, the three major parts of the electric system should be determined because these contributes a major issue in power system planning and as the demand fluctuates over a day, power companies needs to use various types of power sources [10–12]. These three load compositions namely base load, intermediate load and peak load poses a challenge in developing ANN models that have applications in load forecasting since the three loads are not always correlated from each other [10, 11, 13]. Despite the known applicability in power management decisions, very few literatures are devoted to developing ANN models used in determining and forecasting the day-ahead base, intermediate and peak loads.

Power utilities need personalized predictive models like ANN to forecast their base, intermediate, and peak load demand as the absence of such scientific method could lead to over or under purchasing in the electricity market. Accurate base, intermediate, and peak load forecast is essential for the decision making of the electric utility such as determining the need for new and existing resources and type of resources that must be added to meet customers' demand [11, 12]. This study aims to contribute in the

development of appropriate ANN forecasting models for base, intermediate and peak loads by conducting a performance analysis of the applicable ANN training algorithms and activation functions. During the training and testing phases of the ANN models, evaluation of the learning rate and momentum for each formulated model consisting of varied training algorithms and activation functions was then performed to ensure that the ANN model can be appropriate for day-ahead forecasting of base, intermediate and peak loads.

2 Methodology

2.1 Electric Load Data Preparation

There are various factors that can affect the success of ANN model performance, one of which is the quality of the data. Three years of consumed electric load from January 2012 to January 2015 of a certain locality in the Philippines was used in this study. As shown in Table 1 the 96 daily observations are composed of the date, time and kilowatt delivered (KW_DEL) representing the consumed electric load. A total of 101,760 rows of daily fifteen-minute observations were used in this study.

Table 1. Format of the raw electric load data

DATE	TIME	KW_DEL
5/14/2013	00:15	XXXX
5/14/2013	00:30	XXXX
5/14/2013	00:45	XXXX
5/14/2013	01:00	XXXX
5/14/2013	01:15	XXXX

Data corrections were then made for the outlier values of the data set. Clustering of the daily electric load data was then performed using K-Means to determine three daily base, intermediate and peak loads. The daily clustering of the electric load was done through K-Means with $k = 3$ and $i = 4$ all throughout the 96 observation groups of the 101,760 rows of data for the 1060 days resulting to 3180 data sets after the clustering. Data normalization was then performed to the clustered data set to ensure the efficiency of the ANN model [14, 15]. This involves transforming the data to fall within a smaller or common range such as -1 to 1 or 0 to 1 . In this study, the researchers only used Min-max normalization method because this technique produces positive values which can be appropriate for the available data. A study showed that Min-Max normalization technique produces the highest accuracy in all their experimental locations and it is identified as the normalization having the fastest computational time for ANN [16]. Another study compared Min-Max normalization, Z-score, and Decimal Scaling found out that Min-Max normalization datasets as the best performing data normalization method [17]. Furthermore, the Min-Max normalization method was used in this study for the reason that it produces a value ranging from 0 to 1 , the same range for clustered

electric load data. Equation (1) shows how the data set was computed using Min-Max normalization where \bar{x} is the normalized load, x is clustered load, x_{MIN} and x_{MAX} are the minimum and maximum of the dataset respectively.

$$\bar{x} = \frac{x - x_{MIN}}{x_{MAX} - x_{MIN}} \quad (1)$$

After transforming the data, the clustered data set was partitioned into training and testing sets. Partitioning the electric load data is necessary in order to group each data set according to their use for the ANN. As shown in Fig. 1, the electric load dataset was divided into training set and testing set where 2/3 of the data with 2120 rows were partitioned into training dataset which comprises the clustered load data from 2012 to 2013 and the remaining 1/3 of the data with 1060 rows were partitioned for testing dataset which is the year 2014. This partitioning scheme is supported by a study which also used three years' worth of electric load data from January 2003 to December 2006 [18]. Another study also divided the input dataset into three parts: 70% data was used to train the network, 15% was used for testing and another 15% was used for validation [2]. These studies suggest that the training data set should comprise the largest part of the entire dataset.



Fig. 1. Data partitioning of the electric load data

2.2 ANN Model Design Evaluation

A multilayer perceptron neural network with one layer each for the input, hidden and output layers was used as the architecture for this study. The number of neurons used in the input layer was determined and it was then represented through numerical values in order to be fed in the ANN for training. Single hidden layer is sufficient for ANN to approximate any complex nonlinear function with any desired accuracy [4, 7, 9]. Determining the number of hidden neurons in the hidden layer is a vital aspect in designing the overall neural network architecture [4, 19]. Despite this importance, there is no established theoretical approach in determining the number of hidden neurons since most of the popular methods still depends on trial and error basis. Table 2 shows suggested approaches in determining the number of hidden neurons. Using these approaches, the possible number of neurons was calculated and the Mean Squared Error (MSE) is computed for each calculated number of neurons upon implementation. Moreover, the number of hidden neurons with the lowest MSE will be regarded as the most optimal configuration.

Table 2. Approaches in determining the number of hidden neurons

Authors	Process
Param (2015)	The number of hidden neurons is within the range of the input and output layer
Panchal et al. (2014)	Two-third (2/3) of the number of input neurons plus the output neurons
Karsoliya (2012)	Less than twice the number of the input neurons

This study formulated different neural network models with combinations of training algorithms and activation functions for the hidden and output layers. After the models were formulated, they were implemented using Encog Java Library for training and testing thereafter. Factors such as learning rate and momentum were also considered in training the model as these affect the learning process of the ANN1. The epoch value and maximum error were also considered in training the model to determine how much iteration it takes to get the error rate of the network below the maximum error [20]. Both learning rate and momentum were determined and the researchers conducted trial and error in order to find the optimal combination for the models' learning rate and momentum. Setting up the learning rate will help the network determine the weight adjustments each time the weights are changed during the training process [3, 20]. It also helps to specify the degree on which the weight matrixes will be modified through iteration and how quickly the network will learn. The momentum allows the network to speed up its convergence and maintain generalization performance of the network and also helps to specify how much previous learning iteration affects the current iteration [3, 4, 8]. Attempts of training were then conducted using the combination of training algorithms, activation functions, epoch, maximum error, learning rate and momentum.

The researchers then formulated different types of training algorithms and activation functions. These training algorithms work to iteratively modify each weights of the network to minimize overall error between the actual and predicted output for all output nodes over all input nodes. In this study, the researchers first set the epoch value and maximum error to 5000–10,000 and 0.01, respectively. Training the network takes a lot of iterations or epoch until it gets an error rate below the acceptable level [3, 20]. In training the network, Quick Propagation, Back Propagation, and Resilient Propagation algorithm were used as training algorithms. Quick Propagation algorithm performs a multiplicative update and also handles each component of the weight vector independently. It also does not employ learning rates and momentum values that need to be determined which is sometimes advantageous since it can be difficult to determine the exact optimal learning rate thus, it is considered as the fast learning algorithm which sometimes reduces the computational time a hundred times [20]. On the other hand, Back Propagation algorithm uses both learning rate and momentum. Learning rate and momentum were then set to a range of 0.1–0.9 to accommodate Back Propagation. Moreover, Resilient Propagation does not require learning parameters and each weight has an evolving update-value. These training algorithms were then paired up with different activation functions. There are several types of activation functions, each of

which have their unique uses. These activation functions determine the relationship between the output and the input node and a network that introduces a degree of nonlinearity of the data that is valuable in most ANN [8, 19]. Sigmoid and Gaussian were used as activation functions due to its ability to make the learning weight of a neural network simple. Furthermore, these activation functions were chosen since their computations can handle exponential functions which fit on most learning algorithms that involve a lot of differentiations than other arbitrary functions [20]. These activation functions also produce positive values between 0 and 1 which fit to the training and testing datasets that were normalized in a scale of 0 and 1. As shown in Table 3, every combination of training algorithm and activation function denotes one model.

Table 3. Formulated ANN models

Model	Training algorithm	Activation function
Model 1	Quick Propagation Algorithm	Gaussian
Model 2	Quick Propagation Algorithm	Sigmoid
Model 3	Resilient Algorithm	Gaussian
Model 4	Resilient Algorithm	Sigmoid
Model 5	Back Propagation	Gaussian
Model 6	Back Propagation	Sigmoid

Evaluating the models was done by doing test runs on the different models. Error measure was computed to assess the neural network's accuracy since accuracy is the most important criteria in evaluating the forecasting models and in choosing models which produces a better prediction. To test the prediction accuracy of the model, error measure was calculated to determine and compare the predictive capability of the model. In order to evaluate the ANN models, Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE) were calculated. Measure of error was computed in each test run to determine and to compare the predictive accuracy of the models [1, 12]. The model that produces the minimum MSE and MAPE values will be regarded as the most optimal ANN model in predicting the day-ahead base, intermediate and peak loads.

3 Results and Discussion

3.1 Electric Load Data Preparation Results

The electric load data was originally in a .xls format, thus it was converted into a .csv file and then imported into a database for query convenience. Linear and nonlinear properties were also found from the electric load data. As evident from the graph plotted shown in Fig. 2 the plot follows a constant trend representing the linearity of the electric load data while the spikes represents the non-linearity of the electric load.

Due to scheduled and unscheduled power interruptions, the electric load data was found out to have zero values, which causes the historical data to become inefficient



Fig. 2. Sample electric load data plot

and out of range. Data correction was made to outliers and missing values because once disregarded, it will yield to poor results [17]. There is no right or wrong way of data correction, the method would just depend on the researchers [4, 6, 9]. Data correction was made on the raw electric load data by removing zero values and by replacing them with the values of the average of the preceding and succeeding days. Table 4 depicts a sample of electric load data which has zero values where the zero values were replaced by the values of the average of preceding days and the succeeding days.

This data pre-processing procedure is supported by a study conducted which did corrections to out-of-range values observed on the obtained historical load data [9]. The outlier values were replaced with the average of both the preceding and succeeding values in the series. Researchers also used pre-processing of data by filled missing values of measures by linear regression in order to detect and correct bad data [21]. Shown in Table 5 are final clusters for a particular date after applying K-Means to identify the daily base, intermediate and peak loads. The value below 16,820.9705 kW is the base load while the values above 26,202.7391 kW comprise the peak load [13]. On the other hand, the value below and above the intermediate cluster center 20,645.4180 kW represents the intermediate load [11]. After K-Means clustering, there were a total of 3,285 units of data for 1095 rows.

Shown in Table 6 are the electric load data that was being normalized using Min-Max normalization having values ranging from 0 to 1. Min-Max normalization performs a linear transformation on the original data and preserves the relationship among data values [15, 16]. This normalization technique was used in order for the data to be fed into the neural network for training and to test the accuracy of the prediction. A study on the comparison of normalization techniques in prediction concluded that Min-Max normalization showed a second highest accuracy in all experimental locations with the average of 86.84% compared to Z-score and identified to be the fastest computational time compared to Decimal Point Normalization [16]. Another study found out in the experiment results and suggested in choosing Min-Max normalization as the best design for the training dataset [17]. The said researchers compared Min-Max, Z-Score, and Normalization by Decimal Scaling and concluded that Min-Max normalization dataset has the lowest tree growing time, so it is the faster than the other two methods of normalization.

Table 4. Data correction for zero values

DATE	TIME	KW_DEL	DATE	TIME	KW_DEL	TIME	PRECEEDING	SUCCEEDING
AAA	09:00	0	AAA	09:00	5808	09:00	6768	4848
AAA	09:15	0	AAA	09:15	5760	09:15	6624	4896
AAA	09:30	0	AAA	09:30	5592	09:30	6432	4752

Table 5. Sample base, intermediate and peak loads after K-means clustering

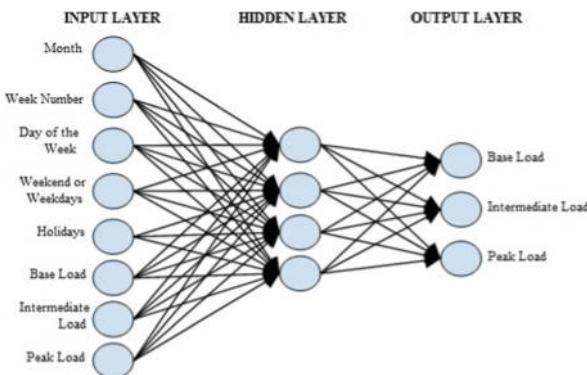
DATE	BASE	INT	PEAK
BBB	16,820.9705	20,645.4180	26,202.7391

Table 6. Normalized electric load data

BASE	INT	PEAK
0.240666598	0.23465889	0.238462481
0.239323519	0.232076194	0.2454447171
0.252431361	0.246454658	0.234943997
0.24463666	0.230605132	0.213712437
0.230328569	0.176836543	0.218362952

3.2 Model Design Evaluation Results

Figure 3 illustrates the ANN architecture showing the neurons found in each of the input, hidden and outputs layers. The input layer has eight (8) inputs neurons consisting of the month, week number, day of the week, weekend or weekday indicator, holiday indicator, and the three (3) system load which consists of base, intermediate, and peak loads.

**Fig. 3.** Proposed ANN architecture

The input neurons of the ANN architecture is supported by a study who also used month, day of the week, working day or weekends and holiday or non-holiday as part of the input neurons [22]. The result showed that exogenous and endogenous inputs to the network are better than just exogenous inputs to the network, as it is difficult for the network to learn when only exogenous inputs are presented into it. Studies also used day of the week and holiday or non-holiday as one of their input neurons in training a neural network as they are very material in the behavior of electric load consumption

[1, 23]. Studies also used input neurons that includes the year, month, week number, and day of the week [2, 3, 9, 22]. The idea behind including year, month, week number, and day of the week in the input layer is to make the network learn in giving the predictions based on them for the future. However, researchers also expressed that there is no suggested systematic way to determine the number of input [4, 6, 13]. Thus, the input neurons included in this study was based on several researchers who suggested that these input neurons can affect the performance of the neural network for training.

Table 7 shows the value representation for the input neurons. The day of the week represented by three binary input neurons, which is for Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday. Weekend/Weekday and Holiday/non-holiday will be represented by 0 and 1 [22]. Weekend for 0 and weekdays for 1, and holiday for 1 and non-holiday for 0. The month will represented by 4 binary input neurons which corresponds to the month of January up until to the month of December. And the week number will represented by seven binary inputs which corresponds by the number of weeks within the whole year. A who also did data representation suggested that binary numbers should be used for the month of the year, day of the week, week number, and holiday as a binary input for optimum performance [3]. The ultimate goal of these methodologies is to ensure that data can be represented in numerical forms that the ANN can process in the input layer.

Table 7. Data representation for input neurons

Day of the week	Weekend/weekday	Holiday indicator	Month	Week number
001	0	0	1000	1000000
010	1	1	1100	1100000
011			1110	1110000
100			1111	1111000
101			1001	1111100
110			1011	1111110
111			1010	1111111

This study utilized only one hidden layer. Determining the number of hidden neurons is a rule-of-thumb approach as there is no theoretically established technique in its standardization. In order to identify the appropriate number of hidden neurons, Table 8 summarizes the hidden neurons results. Each number of hidden neurons was tested using the best performing model and yielded results in MSE. Based on these approaches in determining the number of hidden neurons, the computation of the size of neurons in the hidden layer ranges from 3 to 8 yielded the smallest MSE. Thus, among the three hidden neurons approaches, the first choice presented with 4 hidden neurons yielded the lowest MSE while the 3rd choice with 6 hidden neurons yielded the highest MSE.

In evaluating the performance and the quality of each model, the training dataset was fed into each model during the training phase. Models 1 and 2 were trained using

Table 8. Hidden neurons results

Process	Hidden neurons	Mse
Within the range of the number of input and output neurons	3	0.0097995
	4	0.0067880
	5	0.0075200
	7	0.0089357
Two-third of the number of input neurons plus the number of the output neurons	8	0.0084619
Less than twice the number of the input neurons	6	0.0094607

Quick Propagation algorithm while Models 3 and 4 were trained using Resilient Propagation algorithm and Models 5 and 6 were trained using Back Propagation algorithm. Generally, there are a total of seventy-eight (78) training attempts where each model has thirteen (13) training attempts. In every training attempt, the epoch error was recorded in order to observe the efficiency and accuracy of every iteration. Table 9 shows the MSE and MAPE results for each model.

Table 9. ANN models results

MODEL	MAPE	MSE
Model 1	5.88%	0.005700397
Model 2	8.38%	0.006925646
Model 3	13.04%	0.015086133
Model 4	9.11%	0.008177917
Model 5	N/A	0.2916356349
Model 6	6.03%	0.00980438459

Model 1 which is the combination of Quick Propagation algorithm and Gaussian activation function produced the lowest MSE that yields to 0.005700397. On the other hand, Model 5, a combination of Back Propagation algorithm and Gaussian activation function produced the highest MSE that yields to 0.2916356349886915. This could be the result of the non-convergence of the formulated model. Model 2 and Model 4 produced MSE of 0.006925646 and 0.0081779167 respectively. This could be interpreted that these two models are insufficient in yielding precise forecast. As shown in Fig. 4, Model 3 and Model 6 that yields to 0.01508613256 and 0.009804384592 MSE respectively, turned out to have higher MSE as these two models have MSE nearer to the maximum error of 0.01 of the neural network.

As shown in Fig. 5, the model that yields the lowest MAPE is Model 1, which has a MAPE of 5.88% and the model that exhibits the highest MAPE is Model 3 having a MAPE of 13.04%. Model 2, Model 4, and Model 6 have lower MAPE of 8.38, 9.11, and 6.03% respectively and are the same models which have a lower MSE. This implies that these models have the capability of predicting though not greater accuracy.

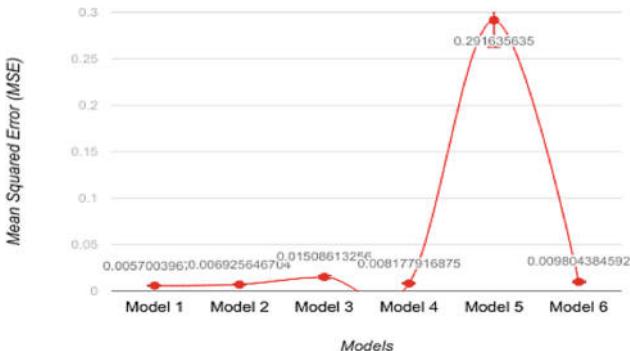


Fig. 4. Proposed ANN architecture

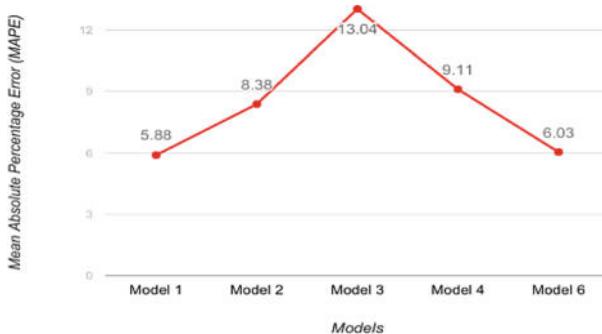


Fig. 5. Performance of the ANN models in terms of MAPE

On the other hand, Model 5, has no MAPE result due to the non-convergence of the formulated model. This phenomenon can lead to the interpretation that this pair of training algorithm and activation function tends to be incapable in achieving precise forecast. Studies also express that Back Propagation algorithm and Gaussian activation function is not a good combination to aim for an accurate forecast [8, 9, 23]. Furthermore, Back Propagation is sensitive to initial conditions and researchers have argued that there is no proof of convergence of the technique. In addition, it has also been stated that convergence time are related to the difficulty of the problem. However, a study contradicted that Back Propagation algorithm is much better compared to other training algorithms for the reason that it exhibits greater accuracy as it allows to learn and improve itself [24]. But in the case of this research, it was found out that Back Propagation algorithm doesn't perform well especially if paired up with Gaussian activation function.

Based on the MAPE and MSE results, the combination of Quick Propagation algorithm with Gaussian activation function turns out to be the model which yielded the lowest error in both MAPE and MSE. Thus, this implies that out of the six formulated models, Model 1 tends to have better prediction accuracy. Quick

Propagation is considerably more efficient than Resilient Propagation and Back Propagation for supervised feedforward neural networks. This made an advantage for the Quick Propagation because it requires no parameter setting before using it [9, 12, 20]. There are no learning rates, momentum values or update constants that need to be determined that is why it is much better because it can be difficult to determine the exact optimal learning rate. Figure 6 shows the comparison of Gaussian and Sigmoid activation functions in terms of MSE and results. This illustrates on which activation function is better to be paired off to the three training algorithms. It was found out that Gaussian activation function doesn't perform well when paired off to Back Propagation and Resilient Propagation algorithms. However, it performs well when combined with Quick Propagation algorithm having the MSE of 0.005700397. On the other hand, Sigmoid activation function produces better MSE results when paired up with the three models and there is no material difference of the results having the MSE of 0.006925646, 0.008177917, and 0.0098043845 for Quick Propagation, Resilient Propagation, and Back Propagation algorithm respectively. As supported by a study, it was shown that Sigmoid activation function performs better than Gaussian with the corresponding error of 0.0079952301 and 0.0003930641 having a final epoch of 50 and 30 respectively [8]. Hence, it can be inferred that Sigmoid activation function performs better than Gaussian activation function.

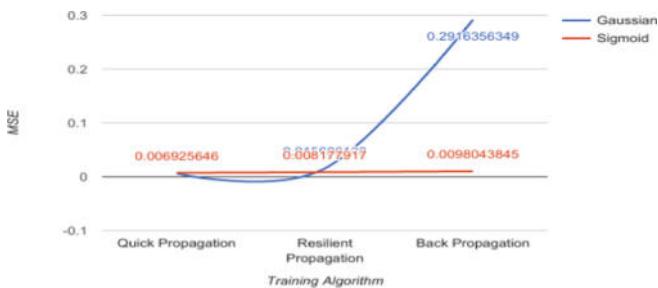


Fig. 6. Performance of the ANN models in terms of MAPE

As shown in Fig. 7 where MAPE comparison was performed on Gaussian and Sigmoid activation functions, it was also found out that Gaussian activation function doesn't perform well on Resilient Propagation with a MAPE of 13.04% and giving out no result when paired to Back Propagation algorithm due to the non-convergence of formulated model, however, Gaussian activation function performs well when paired to Quick Propagation algorithm with a MAPE result of 5.88%. Sigmoid activation function, on the other hand, also shows better MAPE results when paired up with the three training algorithm, results show that there is no material difference among the results having a MAPE values of 8.38%, 9.11%, and 6.03% for Quick Propagation, Resilient Propagation, and Backpropagation, respectively.

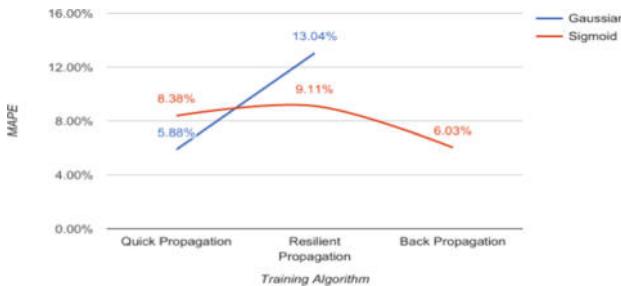


Fig. 7. Comparison of the activation functions in terms of MAPE

References

1. Olagoke, M.D., Ayeni, A.A., Hambali, M.A.: Short term electric load forecasting using neural network and genetic algorithm. *Int. J. Appl. Inf. Syst.* **10**(4), 22–28 (2016)
2. Bala, A., Yadav, N.K., Hooda, N.: Implementation of artificial neural network for short term load forecasting. *Curr. Trends Technol. Sci.* **3**(4), 247–251 (2014)
3. Ortiz-Arroyo, D., Skov, M.K., Huynh, Q.: Accurate electricity load forecasting with artificial neural networks. In: Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce. IEEE (2006)
4. Velasco, L.C.P., Granados, A.R.B., Ortega, J.M.A., Pagtalunan, K.V.D.: Performance analysis of artificial neural networks training algorithms and transfer functions for medium-term water consumption forecasting. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **9**(4) (2018)
5. Nieminen, P.: Multilayer Perceptron Training with Multiobjective Memetic Optimization. Jyväskylä University Printing House (2016)
6. Hush, D.: Classification with neural networks: a performance analysis. In: International Conference on Systems Engineering. IEEE (2002)
7. Cireşan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Deep big multilayer perceptrons for digit recognition. In: Neural Networks: Tricks of the Trade, pp 581–598. Springer, Berlin (2012)
8. Sibi, P., Allwyn Jones, S., Siddarth, P.: Analysis of different activation functions using back propagation neural networks. *J. Theor. Appl. Inf. Technol.* **47**(3), 1264–1268 (2013)
9. Velasco, L.C.P., Villezas, C.R., Palahang, P.N.C., Dagaang, J.A.A.: Next day electric load forecasting using Artificial Neural Networks. In: International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management. IEEE (2016)
10. Abdulaal, A., Buitrago, J., Asfour, S.: Electric load pattern classification for demand-side management planning: a hybrid approach. In: Proceedings of the IASTED International Symposium Advances in Power and Energy Systems (2015)
11. Salimi-beni, A., Fotuhi-Firuzabad, M., Farrokhzad, D., Alemohammad, S.J.: A new approach to determine base, intermediate and peak-demand in an electric power system. In: Proceedings of International Conference on Power System Technology. IEEE (2006)
12. Velasco, L.C.P., Estoperez, N.R., Jayson, R.J.R., Sabijon, C.J.T., Sayles, V.C.: Day-ahead base, intermediate, and peak load forecasting using K-means and artificial neural networks. *J. Adv. Comput. Sci. Appl. (IJACSA)* **9**(2), 62–67 (2018)

13. Grant, J., Eltoukhy, M., Asfour, S.: Short-term electrical peak demand forecasting in a large government building using artificial neural networks. *Energies* **2014**(7), 1935–1953 (2014)
14. Panigrahi, S., Karali, Y., Behera, H.S.: Normalize time series and forecast using evolutionary neural network. *Int. J. Eng. Res. Technol.* **2**(9) (2013)
15. Jayalakshmi, T., Santhakumaran, A.: Statistical normalization and back propagation for classification. *Int. J. Comput. Theory Eng.* **3**(1) (2011)
16. Mustaffa, Z., Yusof, Y.: A comparison of normalization techniques in predicting dengue outbreak. In: Proceedings of the International Conference on Business and Economics Research. IACSIT Press (2011)
17. Shalabi, L., Zyad, S., Al-Kasasbeh, B.: Data mining: a preprocessing engine. *J. Comput. Sci. (Science Publications)* **2**(9) (2006)
18. Ramachandran, P., Senthil, R.: Locational marginal pricing approach to minimize congestion in restructured power market. *J. Electr. Electron. Eng. (Academic Journals)* **2**(6), 14–153 (2010)
19. Panchal, F.S., Panchal, M.: Review on methods of selecting number of hidden nodes in artificial neural network. *Int. J. Comput. Sci. Mob. Comput.* **3**(11), 455–464 (2014)
20. Heaton, J.: Encog: Library of interchangeable machine learning models for Java and C#. *J. Mach. Learn. Res.* **15**, 1243–1247 (2015)
21. Rodrigues, F., Duarte, F.J.F., Silva, V., Cordeiro, M.: Comparative analysis of clustering algorithms applied to load profiling. In: Proceedings of Machine Learning and Data Mining in Pattern Recognition, MLDM 2003 (2003)
22. Rashid, T.: Study of artificial neural networks for daily peak load forecasting. In: Proceedings of 2nd International Conference on Information Technology (2005)
23. Hernández, L., Baladrón, C., Aguiar, J., Carro, B., Sánchez-Esguevillas, A.: Classification and clustering of electricity demand patterns in industrial parks. *Energies* **2012**(5), 5215–5228 (2012)
24. Chaudhary, R., Patel, H.: A survey on backpropagation algorithm for neural networks. *Int. J. Technol. Res. Eng.* **2**(7) (2015)



Quantum Deep Learning Neural Networks

Abu Kamruzzaman^(✉), Yousef Alhwaiti, Avery Leider, and Charles C. Tappert

Seidenberg School of Computer Science and Information Systems,
Pace University, Pleasantville, NY 10570, USA
[{ak91252p,ya12919p,aleider,ctappert}](mailto:{ak91252p,ya12919p,aleider,ctappert}@pace.edu)@pace.edu

Abstract. This study surveys the current status of Quantum Deep Learning Neural Networks. Exciting breakthroughs may soon bring real quantum neural networks, specifically deep learning neural networks, to reality. Three main obstacles have been limiting quantum growth in the deep learning area, and this study has found that new discoveries have changed these obstacles. The first obstacle was the lack of a real quantum computers to experiment with, not simulators. Several companies have significantly increased their inventory of quantum computers in the last year, including IBM. The second obstacle was the impossibility of training quantum networks, but a new algorithm solves this problem. The third obstacle was that neural networks have nonlinear functions, but that has been solved with a new quantum perceptron. This study explains the historical background briefly for context and understanding, then describes these three major accomplishments that will likely lead to real quantum deep learning neural networks.

Keywords: Deep learning · Quantum computing · Neural networks · Perceptron

1 Introduction

Our study first explains the necessary information about how neural networks and the deep learning types of neural networks function. This is to understand the significance of the recent discoveries and place them in context of the larger whole. This is followed by enough detail about quantum computing that one can see how the quantum computing forces will enhance in the quantum version of the neural network. The second part of our study brings attention to the algorithm that enables the quantum neural network to accomplish training. Training is a key component of a neural network. The third part of our study describes the quantum perceptron which is vital to achieving a true quantum neural network and is superior to other ideas that would have most of the perceptron action happening in classical, instead of quantum, computing processes.

Most important in the future of quantum neural networks, especially deep learning ones, is their use to create artificial minds that are human-like in the

Thanks to the IBM Faculty Award that made this research possible.

way they learn and process information [27]. These computers will have great capacity for good. There is already a large field of researchers with algorithms, models, designs and hopes that wait now, researching and planning, who seek to work with these artificial minds [4].

A neural network is an information processing model that mimics the design of mammalian brains in nature, which are comprised of neurons, that are connected to each other with many connections [6]. In nature these neurons form special associations, transformations and mappings as a part of learning in the brain [7]. In computer science, these neurons are artificially created, may be called units, nodes, perceptron [22] or neurons and they also perform learning, through their many connections to each other and their organization in layers. In a standard neural network model, input neurons are activated through sensors that represent in nature the senses that connect to the brain, sensors such as microphones or cameras. The activation signals from the sensors become data inputs that pass along through multiple layers for analysis and adjustments, based on the parameters set by the architecture. When the model outputs its findings, the results are used to enforce or deter as a teaching method to improve quality in the learning system. In the memory enhanced models [25], such as the Long Short-Term Memory (LSTM) [23], which is a neural network that includes memory-based learning, the data is recorded and used during the processing phase to enhance the next analysis stage. Bringing the possibilities of quantum computing into the equation will allow for additional variables to be applied to specific decisions made by the analysis engine.

Quantum computing applies the knowledge of quantum physics and quantum mechanics to manipulate elemental particles such as electrons, photons or ions, to create processing power [14]. Most important are entanglement and superposition, which are used to develop special circuits that offer options beyond the classical computers binary options of one and zero [26]. Binary computers are based on transistors, whereas quantum computing uses quantum bits, qubits, to perform calculations. Qubits are objects that can exist in two exclusive states in the same instance.

Some research findings into quantum artificial neural networks support the idea that a completely quantum neural network will not perform effectively [17]. The term, fully quantum is not meaning totally quantum, as all quantum computers have a classical component. What is meant by fully quantum is that the neural network computations would take place in the quantum computer. This idea is based on a popular model which creates a network that has quantum computing processing components interfacing with classical computing components in a blended network designed specifically for neural network processing [15]. In that system, only small processing steps that are optimal for quantum are delegated to the quantum component, and most of the neural network non-linear processing is done on the classical side.

There is another way to accomplish a quantum neural network that is not yet modeled, but the first step has been designed, which is to create a quantum

perceptron with a circuit [8]. Future research is sure to work on building a new model that is built with the quantum perceptron as the building blocks.

A deep learning neural network is the neural network system design that is most often used to analyze images. Deep learning neural networks utilize a type of multilayer perceptron, which consists of an input, output, and at least one hidden layer, that learns with little to no preloading [9]. The system translates pixels and features derived from an image to classify the object using provided training. The classification of the output is assigned a probability based on the numeric translation that has been compared to the data in the training set. Many research papers on deep learning neural networks mention the Modified National Institutes of Technology (MNIST) database of images of handwritten digits to test ideas [12]. This is because MNIST has become a benchmark by which new deep learning neural network designs are compared. The challenge, with a quantum computer based design, is teaching the system on the training data [21]. Research on a proposed quantum multi-layer neural network with self-organizing design promises to identify the handwritten digits, but it is completely theoretical [10].

A typical deep learning network completes its task over several phases. First, the system will take an input from an image, such as one from the MNIST, as a sample for analysis. Next, the system will compare pieces, known as characteristic features, to determine if the compared images are similar in structure. The convolutional neural network, a type of deep learning network, will attempt these comparisons wherever possible, in order to calculate the highest accuracy. Afterwards, this process is repeated, on other identified subsamples. This process can include pooling, which is a technique to alter large images and downsize them while preserving the most important features.

In this research project, the authors of this study are trying to understand if Deep Learning models are suitable to store the learned knowledge representation to make better use of classifying and recognizing images and other patterns. The Deep Learning models explored here include CNNs pre-trained models (ResNet50, VGG16, InceptionV3, and MobileNet) on ImageNet datasets and trained model on MNIST datasets.

The rest of this research paper proceeds with a literature review to provide historical background briefly for context and understanding of concepts. It explores if Quantum Computing Deep Learning models are suitable to store learned knowledge. The authors describe the quantum concepts that help with understanding these capabilities in Quantum Neural Networks. Then in this paper we offer the advantages and obstacles of Quantum Neural Networks. Finally, this study examines the relationship between Artificial Neural Networks and the Quantum counterpart Neural Networks, the latter evolving as quantum computation quickly increases in usability. The doors to quantum computing enhanced deep learning neural networks, as well as the storage of the learned data, may be unlocked sooner than researchers think.

2 Literature Review

Ricks [21] realized that the gradient descent algorithm that had been suggested for Quantum Neural Networks was too fatiguing for the qubits, and created a unique algorithm, that is executed in two steps. The first step is called the Simple QNN, that is a complex quantum circuit that uses qubits in a pattern of gates to compute the XOR function. The second step, called QNN Training, is a macro view of the Simple QNN design that shows how it can accomplish training in a quantum neural network [21]. In that unique algorithm, an assumption was made, about the existence of quantum perceptron, which at the time of their paper, did not yet exist. The perceptron operate just as they do in a classical neural network system, however these perceptron are doing quantum processing. Figure 1 shows how a quantum neuron receives input training. Figure 2 shows how the trained neuron sends the output.

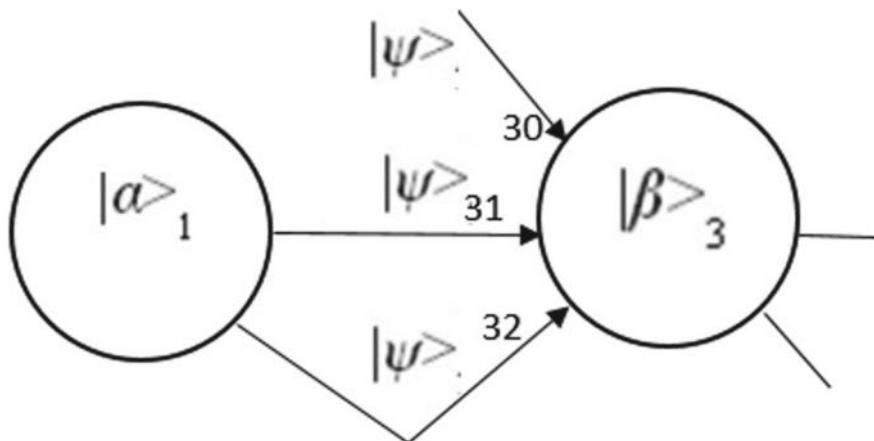


Fig. 1. Quantum neuron receiving input training [16]

From that unique algorithm, a system can be developed using convolutional neural network to demonstrate the effectiveness in quantum computing through the process of analyzing a dataset of characters in images and its resulting output [9]. The proposed methodology and experiment will use quantum mechanics to accurately identify text with an image dataset.

Kim [9], at NYU, set out to determine the effectiveness of convolutional neural network for sentence classification. Kim proposed using word vectors that were pre-trained to effectively process information through a single layer of convolution. Kims experiments assigned text to its correct category based on the sentence structure. In his experiment, several model variants were implemented, however the variants were not well differentiated as for comparison each model had the same data source and same training algorithm was applied to identify the data category.

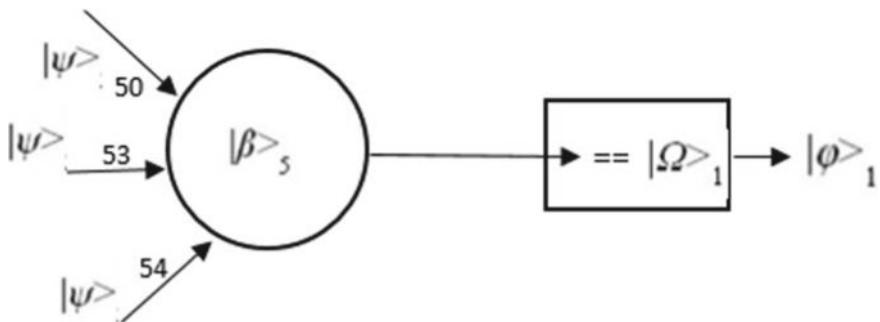


Fig. 2. Trained neuron sending output [16]

Kim [9] discovered that at its rudimentary level, with a single layer of convolution, the system performed as expected with the given hyper parameters. From this, it follows that supervised pre-training of word vectors is beneficial for neuro-linguistic programming. This may transfer over into the design of researchers future systems, if it can appropriately identify first the text in the image and then follow that with determining the tone of the sentence and its meaning.

In the convolutional neural network , which is a type of neural network [2], there are multiple models that are in two basic groups: (1) scene labelling and (2) graphical interpretation approaches. In scene labelling, each individual pixel of an image is identified with a calculated value, whereas in graphic interpretation a field (class or feature area) may be comprised of a pixel array. The scene labelling method uses a feed-forward system, using a mechanism that recalls labels over processing layers. This method of supervised learning has proven to be both feasible and effective.

The acceptance of convolutional neural network s and the development of the enhancement of results by adding elastic distortions for image analysis were not always so popular. In [2], the author approaches this issue, in the early 21st century. Elastic distortion is the alteration of an image through simple geometric changes, such as rotations, translations or skewing. In the convolutional neural network system, the usage of elastic distortion, as the author describes, is used to recognize similar characters that have been slightly altered. In this case, we can also use elastic distortion to vary the training set of data to improve system performance.

Quantum neural networks are the culmination of physics, mathematics, and computer science [18]. Much research on them have relied upon simulations done on classical neural networks. These simulations are reliant on universal ideas of quantum computing, which use the setup of multi-layer perceptron to act as qubits.

3 Quantum Neural Networks

As the development of computers, including their underlying systems, grows at a faster pace, the design phase has begun to mirror that of our human brain model. Although how our human cognitive processes are designed has yet to be completely solved, the foundations of this has allowed engineers to propose methods that in simulations create neural networks that emulate human brain functions. The potential of brains and quantum computation seamlessly processing thought engages the imagination, although at this current moment, there is minimal practical reality of these thoughts [3]. This area of research is alive with quantum computing ideas, bringing out that quantum computing is using nature to do its work, as the quantum particles are found in nature, and seeking where this natural order could reveal more in neuro-computing and brain function [5].

Neural network systems retrieve input data, process the information, and provide an output outlined by the set parameters. This has much in common with the design of brain learning systems. Neurons in the human brain communicate via synapses to create consciousness and process information. The current approach in research is to model that the mass of living synapses need to be replicated in the digital environment in order to create a comparable model. Nayak is one of many researchers who believe that the achievement of quantum neural network computing will deliver a digital brain model that is very close in capability to the brain [18].

The proposed architectures for a quantum neural network vary by design, some that are similar to classical neural networks, while other exhibit their own unique qualities. In the proposed model by Ricks [21], a simple quantum neural network will continue to utilize layered perceptron. The perceptron would include an input layer, at least one (if not more) hidden layers, and an output layer. Layers are connected fully, in a method that provides a link between every perceptron in a layer to everyone in the subsequent layer. Quantum networks would require two inputs to generate weight that the system will classify [16].

The possibility that the field of classical artificial neural networks can be generalized to the quantum domain by eclectic combination of that field with the promising new field of quantum computing. Each factor suggests a new interpretation of brain and cognitive function. Furthermore, new abilities in information processing that may yet to be acknowledged or discovered. As a whole, we may consider quantum neural networks as the next step in the evolution of computing system, as we shift our focus of our attention on artificial instead of organic systems.

4 QNN Background

QNN is an Artificial Neural Network (ANN) that is included in a system with Quantum Computation. The reason researchers are attempting this is to develop more efficient algorithms in pattern classification or machine learning than what is available now in the capabilities of ANN. Quantum computation expands

the computational behavior of machines to exponential capacity, using quantum parallelism and the effects of interference and entanglement [16]. Quantum computation promises to expedite the training of classical neural networks and the computation of big data applications, generating faster results.

Quantum computing and quantum information ideas that are most powerfully driving research today originated with Richard Feynman, a physicist who realized that the development of future hardware would have to require quantum effects [3]. Feynman's observation was that the limitation of the capacity of classical hardware, gates and wires would be overcome by demand, and those classical systems would need to be replaced by computers that would consist of only a few atoms using quantum forces to do their information processing.

Some of the quantum concepts that help with understanding quantum neural networks are introduced below:

Linear superposition is analogous to the mathematical standard of linear combination of vectors. Quantum computing recognized through a wave function exists in a Hilbert space [19].

Coherence and decoherence has similarity with linear superposition. A quantum computer is coherent when linear superposition is in basis states, and if there is no coherence a quantum computer is in decoherence [19].

Operators work on transformation of states on the wave functions on a Hilbert space [11].

Interference works on wave occurrence and it is measured in amplitude. Phase interfere constructively are in phase while wave peaks amplifying each others amplitude. Phase interfere destructively will decrease or eliminate each others amplitude [1].

Entanglement is a special quantum state that does not exist in classical computers. This reveals correlations [19].

In quantum computing, quantum information processing is completed by going beyond the classical binary system for computers. In traditional computing, systems use binary numbers, ones and zeroes, to indicate an on or off state; quantum computing utilizes superposition and entanglement to enable a system to expand beyond the classical binary.

Superposition: An object can exist in both forms of a binary, in the case of computing, it is a zero and one. Superposition can be illustrated in computing by two states occurring over the variable amount of occurrences.

$$N_{\text{cubes}} = 2^N \text{ states}$$

Entanglement: An object that occurs in one state becomes paired with another state.

QNN advantages include the following [20]:

- memory capacity is in exponential
- lower number of hidden neurons provide higher performance
- learns fast
- eliminates catastrophic forgetting for not having pattern interference
- linearly inseparable problems for single layer network solutions

- wires not required
- processing speed (1010 bits/s)
- small scale (10^{11} neurons/mm 3)
- higher stability and consistency.

Figure 3 highlights the steps for classical neural networks and the quantum neural networks. The figure also shows the differences between Classical neural networks and QNN models.

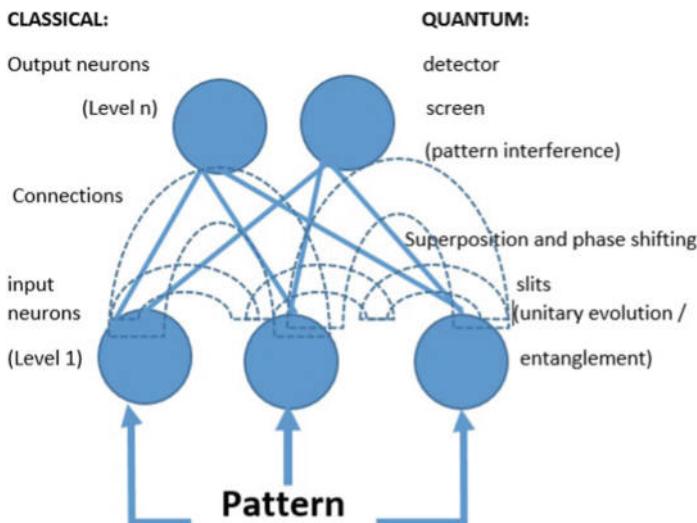


Fig. 3. Classical NN versus QNN [9]

Quantum Neural Computing (QNN) is the joining of a classical neural network, which could be of a type such as a convolutional neural network(CNN), a deep learning neural network or an artificial neural network (ANN), with some aspect of quantum computing. A previous obstacle to putting a quantum feature into a classical neural network system has been that quantum computation is a linear and unitary action, but all of the designs of the classical neural network depend upon nonlinear methods.

Hu [8] has designed a way to solve this. He found a way to have the nonlinear activation function needed for QNN to work inside the quantum neuron, even though it is the law that quantum computing operations must be both unitary and linear. He accomplished this by using a special quantum circuit, the Repeat-Until-Success (RUS) circuit.

Lewenstein [13] twenty-two years before Hu, published research in which he proposed quantum perceptron that would use linear algebra to receive input states, do a unitary transformation operation on them, and then deliver an output state. What is different about Hu is that he tested it on a simulator

(which was not available to Lewenstein) and he used a series of common gates to make his unique circuit.

Hu then followed up his success on the simulator by testing his idea on the IBM quantum simulator for a 5 qubit machine. In the RUS the action is repeated until a 0 is measured, which is found when a rotation with an angle is found in the quantum bit (qubit). When he tested his quantum perceptron creating circuit on the IBM real 5 qubit machine. It worked. This work on an actual quantum computer, instead of a simulator, is very exciting. At this point in time, Hus experiments work for input in the range of $[0, n/2]$ so the capacity is very limited [12]. One challenge is how much RUS will the qubits bear before they fall into quantum decoherence? Qubit states of coherence are perishable, and these quantum neurons must hold their states long enough to complete the neural network calculations. However, this RUS design as shown in Fig. 4 holds great promise for the future of Quantum Neural Networks.

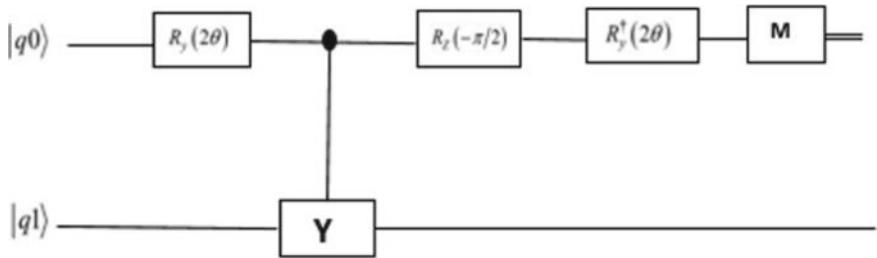


Fig. 4. RUS Quantum circuit [8]

The field of ANN contains multiple important ideas [26]:

- the concept of a neuron or processing element
- the transformation which is performed by this processing element (in general, input summation and nonlinear mapping of the result in to an output value),
- the interconnection construction between neurons,
- the network dynamics and the learning rule responsible for governing the modification of interaction strengths.

The main concepts of quantum mechanics are [1]

- wave function,
- superposition (coherence),
- measurement (decoherence),
- entanglement, and
- unitary transformations.

It is a major challenge to establish communication in the development of a model of QNN with ANN. Lewenstein [13] stated that in the quantum perceptron,

the classical weights perceptron are replaced as unitary operators to map input to output.

Chrisly proposed a design for feed forward artificial neural network using a double-slit experiment. Menneer and Narayanan [17] have extended Chrislys proposal for a single pattern quantum neural network.

- Multiple universes or superposition principle of quantum theory applied to neural computing.
- The architecture of double-slit experiment provides the basis for quantum artificial neural network.

In the extended model a quantum neural network can be modelled on the basis of double-slit experiment and it includes the following:

- the photon gun is equivalent to the input pattern,
- the slit is equivalent to input neurons,
- the waves between the slits and the screen is equivalent to the connections between the input and output neurons and the screen is equivalent to the output neurons.

Figure 5 shows the differences between ANN and QNN.

ANN	DOUBLESPLIT experimental setup	QNN
Pattern	Photon gun	Quantum register holds input Pattern
Input Neuron	Slits	Entanglement unitary evolution of input pattern
Connection	Waves	Superposition of waves created by input pattern
Output neuron	Detector screen	Pattern interference
Weight	Link	Phase shift

Fig. 5. ANN versus QNN

5 QNN Advantage Over ANN

In QNN, N-set of training patterns forms a set of N-homogeneous components and each training pattern is channeled to one component part and the set of weights connected to this component is changed to learn this training pattern. Due to the independent components for each training pattern, interference does not take place in the learning patterns. However, in a classical convolutional neural network, there is the possibility of pattern interference and the result is that the network unlearns, a form of catastrophic forgetting.

Using different components in QNN separates the pattern and prevents this catastrophe. It is required in a classical convolutional neural network, but in a quantum neural network there is no necessity of a decision plane to separate the patterns into classes. The components are trained as a superposition of networks. It is also possible to first train classical networks and then combine them into a superposition of classes [22]. With a quantum neural network, training time is minimal as compared to classical neural network.

6 QNN Obstacles to Implementation

- Quantum Coherence: The system needs to maintain coherence until the computation is complete. However, the Quantum system interacts with the environment; it is not possible to maintain the coherence.
- Connections: The connections is measured by entanglement of qubits. The measurement is the main obstacle for entanglement.
- The qubits need to form quantum perceptron in order to achieve an optimal quantum neural network.
- A training algorithm specifically for a quantum neural network needs to be utilized, instead of using training algorithms designed for classical computers.

7 Conclusion and Future Work

The Quantum Neural Network is still only theoretical. Since the three major obstacles to making it real have each had recent research breakthroughs, this powerful deep learning method will likely be possible in the near future. There are many areas for growth and change as these breakthroughs become translated into practical experiments on real quantum computers. Once we are able to build and configure the neural network machines with optimal learning behavior, we can revolutionize the execution of computing behavior in considerably less than exponential time. This will potentially change deep learning and all of the ways we use it. For example, there is the promise that researchers using quantum deep learning will soon solve previously unsolvable problems. For future research, our goal is to build and train a quantum computing deep learning neural network program utilizing the circuit design of the new quantum perceptron, test and improve the design of the quantum network training algorithm, and then run our neural network program on a real quantum computer available to researchers [24].

References

1. Carroll, S.M., Singh, A.: Mad-dog everettianism: quantum mechanics at its most minimal. arXiv preprint [arXiv:1801.08132](https://arxiv.org/abs/1801.08132) (2018)
2. Ezhov, A.A., Ventura, D.: Quantum neural networks. In: Future Directions for Intelligent Systems and Information Sciences, pp. 213–235. Springer, Berlin (2000)
3. Feynman, R.P.: Simulating physics with computers. *Int. J. Theor. Phys.* **21**(6–7), 467–488 (1982)
4. Fuller, S.: The brain as artificial intelligence: prospecting the frontiers of neuroscience. AI & SOCIETY, pp. 1–9 (2018)
5. Haroche, S.: Entanglement, decoherence and the quantum/classical boundary. *Phys. Today* **51**(7), 36–42 (1998)
6. Haykin, S.S.: Neural Networks and Learning Machines, vol. 3. Pearson Upper Saddle River (2009)
7. Hecht-Nielsen, R.: Neurocomputing: picking the human brain. *IEEE Spectr.* **25**(3), 36–41 (1988)
8. Hu, W.: Towards a real quantum neuron. *Nat. Sci.* **10**(03), 99 (2018)
9. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint [arXiv:1408.5882](https://arxiv.org/abs/1408.5882) (2014)
10. Konar, D., Kar, S.K.: An efficient handwritten character recognition using quantum multilayer neural network (qmlnn) architecture: quantum multilayer neural network. In: Quantum-Inspired Intelligent Systems for Multimedia Data Analysis, pages 262–276. IGI Global (2018)
11. Kribs, D.W., Pasieka, A., Laforest, M., Ryan, C., da Silva, M.P.: Research problems on numerical ranges in quantum computing. *Linear Multilinear Algebra* **57**(5), 491–502 (2009)
12. LeCun, Y.: The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/> (1998)
13. Lewenstein, M.: Quantum perceptrons. *J. Mod. Opt.* **41**(12), 2491–2501 (1994)
14. Marinescu, D.C., Marinescu, G.M.: Approaching Quantum Computing. Pearson/Prentice Hall (2005)
15. McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H.: Barren plateaus in quantum neural network training landscapes. arXiv preprint [arXiv:1803.11173](https://arxiv.org/abs/1803.11173) (2018)
16. Menneer, T., Narayanan, A.: Quantum-inspired neural networks. *Tech. Rep.* **R329** (1995)
17. Narayanan, A., Menneer, T.: Quantum artificial neural network architectures and components. *Inf. Sci.* **128**(3–4), 231–255 (2000)
18. Nayak, S.: An introduction to quantum neural computing. *Int. J. Global Res. Comput. Sci. (UGC Approved J.)* **2**(8), 50–54 (2011)
19. Nielsen, M.A., Chuang, I.L.: Quantum Computation and Quantum Information. Cambridge University Press, Cambridge (2010)
20. Pinheiro, P.H.O., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: 31st International Conference on Machine Learning (ICML), number EPFL-CONF-199822 in 0 (2014)
21. Ricks, B., Ventura, D.: Training a quantum neural network. In: Advances in Neural Information Processing Systems, pp. 1019–1026 (2004)
22. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**(6), 386 (1958)

23. Sak, H., Senior, A., Beaufays, F.: Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
24. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015)
25. Shin, C.-K., Yun, U.T., Kim, H.K., Chan Park, S.: A hybrid approach of neural network and memory-based learning to data mining. *IEEE Trans. Neural Netw.* **11**(3), 637–646 (2000)
26. Simard, P.Y., Steinkraus, D., Platt, J.C., et al.: Best practices for convolutional neural networks applied to visual document analysis. *ICDAR* **3**, 958–962 (2003)
27. Brooks, T., Kamruzzaman, A., Leider, A., Tappert, C.C.: A computer science perspective on models of the mind. Accepted for publication in IntelliSys 2018 (2018)



Hierarchical Modeling for Strategy-Based Multi-agent Multi-team Systems

D. Michael Franklin^(✉)

Kennesaw State University, Marietta, GA 30114, USA

mfranklin@kennesaw.edu

<http://ksuweb.kennesaw.edu/~dfrank15/>

Abstract. Modeling complex environments is a challenging problem that is compounded when there are multiple agents acting together as a team, and the team needs to maintain its own goals while allowing the agents to have some level of autonomy. We propose a modeling framework for strategy-based multi-agent multi-team simulation environments. For these types of environments it is necessary to have a modeling infrastructure that allows for high-level, high-complexity, hierarchical interactions where team goals are prevalent but individual needs are balanced. Such modeling is proposed in this paper – modeling that will avoid large, monolithic models while maintaining complexity of expression balanced with simplicity of operation.

Keywords: Artificial intelligence · Multi-agent systems · Strategy · Hierarchical modeling · Modeling

1 Introduction

In an effort to create a powerful, expressive, multi-layered methodology for describing, creating, implementing, and measuring multi-agent multi-team interactions we choose to implement strategy (i.e., to formulate complex systems of behavior into cohesive models). In this case, strategy refers to the ability to encapsulate policies within a super-policy that decides which policy should be in place. This decision is based on the current states, recent observations, and experience based on learning over time. Further, we wish to create this hierarchically so that an agent can have a strategy, yet be part of a team of agents, each with their own policy, that has an overall team strategy. This could continue up to more levels following the same guidelines. As a note, the models are uncoupled from the simulation framework, meaning that the models can change to suit the environment or to simulate different types of interactions, and the framework remains unchanged (i.e., it executes the models themselves, so it can follow the same steps for all models without having to change the simulation itself). While we claim that this is generally applicable to many different scenarios, we know that no such general model exists for all possible scenarios. As a further note, this modeling process can be inverted to infer another team's strategy through observations (similar to identifying a graph from the leaves).

2 Related Works

Though the context is different, the article from [8] elucidates the increasing complexities of simulations that require interactive-time interactions amongst many disparate elements. Additionally, in [9] this idea is furthered. The authors present several correlated instances where the simulation must simulate, detect, and adjust for multitudinous particles interacting in a variety of patterns, flows, and avoidances.

For background, [4] creates a simplified system that attempts to match formations with existing known formations and then choose a preferable formation for their own team. This approach does not include multi-agent considerations and tends to favor the single-policy approach common in multiple agent systems. Their desire to predict that optimality of match-ups, however, is informative and helpful to our work.

For some background details and general information on Decision Theory and Game Theory we have consulted [6]. This work, while general, is helpful in framing multi-agent systems and especially noteworthy for its considerations for learning in this environments. Further, it helps us to handle problems at scale based on their considerations.

Many seminal papers offer insight into this challenge even though they do not necessarily propose solutions. We find these papers helpful in understanding the challenges, the variety of approaches and their successes or failures, and in forming our problem with proper considerations. There are many such works, such as [2,3,5]; however, we find the aggregate study in [1] to be more helpful and have consulted it extensively for these considerations.

Stone and Veloso [7] offers an excellent introduction and exploration of the complexity of this type of scenario. Their paper explores role assignment, team management, and communications. Each of these considerations is vital to the proper execution of coordinated, strategic behavior at each level of the hierarchy. When considering stochastic games, this reference provides insight into teamwork and coordination that will work (and some that will not) within such a difficult scenario. There is, additionally, some insight into inference in multi-agent systems.

3 Strategy Modeling

There are a variety of approaches to solving this difficult artificial intelligence problem. Most of these approaches involve, ultimately, a statistical approach. This approach, while valid and useful, is limited when there are multiple agents working together as a team, and multiple teams working together and against each other in the same field. Statistical approaches tend to converge in environments where this one, fixed policy that you is to be found; however, the proposed scenario involves multiple agents that are sharing one overarching policy (what we refer to as a strategy, or a super-policy) but unique or varying individual policies. The resultant solution rapidly and exponentially increased the complexity

and causes issues, like decreasing performance or a required reduction in scale, for single-policy or statistical approaches. In contrast, we propose a model-based approach.

The model-based approach allows for each agent being managed to hold varied, flexible, expressive, and locally powerful behaviors. These models can be designed as Finite-State Automatons (FSAs) or Probabilistic Graphical Models (PGMs). The system architecture allows for these models to be represented as either a diverse set of graphs, where each one represents a policy as a walk through a graph of possible sets of actions or choices, or as a set of multiple isomorphic graphs where the weight of the edges encode the decision matrices. In either case the system allows for multiple layers of hierarchical reasoning and representation where, at each level, there is an overarching directional policy guiding the collected actions of the group of agents to accomplish a shared goal. This allows for a shared and distributed intelligence while carefully preserving the individual agent policies. Further, it can be seen that the large-scale policies flow downward (the coalition strategy flows down to the teams, the strategy directs each team uniquely, and the team policies direct each agent by assigning behaviors to each) while the information used for decisions flows upward (each agent collects observations and information and sends them to their team, the team collects, collates, and processes that information and sends higher-level observations upstream to the coalition).

Further, and most importantly, this multi-level direction and distributed policy network can be reverse engineered. In this case, another coalition can take their observations about the other team and their actions to infer the various behaviors, policies, and strategies that the other team is following. This is done by creating a belief network (BN) that contains individual beliefs about each element within the layer being reasoned about. The aggregate conclusion of the BN can be used to infer the most likely strategy being followed by the other coalition. At each node of the BN the possible actions of each known policy are compared with observed actions to create the probability match that generates that belief. The resultant BN creates a graph. This graph can be compared to known large-scale graphs derived from the known strategies. In this manner, the most likely strategy can be inferred quickly (in soft real-time or interactive time) without having to analyze every single element within the graph and without having to expand each node (as each node represents another sub-graph). This is highly important because the complexity of these systems increases rapidly, and the corresponding analysis proves taxing to even high-powered systems.

To power this model-based system we employ non-deterministic finite state transducers (NFSTs). These NFSTs advance each agent along their prescribed pathway according to the currently in-force policy assigned to them. At each stage the probability pathways are calculated with respect to the characteristics of their unique personalities (here, each agent is assigned individual values for each feature, like their allegiance, or their speed) to determine how each agent will interpret and act on the team's overall policy. This allows for independent and varied action for each agent while still adhering to a team direction, and this

occurs at each level of the hierarchy. This is how the system can be both centered around a single, cohesive policy and retain individual actions and behaviors. The NSFTs work will with FSAa to provide a combined hierarchy of reasoning and communication.

This work utilizes a FSA that allows for a similar action set for each agent but with customizable factoring (probabilistic progression through the model), shown, for example, in Fig. 1 (this is a sample agent model). In this example the agent has distinct phases that are competing for attention each turn. They can be thought of as Markov random fields with multiple variables. This is reflective of the mindset of many real-world agents. For instance, the model starts in a move state and then is directed via the probabilistic pathing towards each of the possible states (e.g., an offensive, a defensive, an observation, or a movement phase). The weighting (factoring) of each agent action is probabilistically determined by the individual agent's policy. This policy is given to them from the strategy the team is following. Thus the agent considers their action with their own probability (based on their agent type) which reflects their own 'personality'. This probability is then modified by the policy according to the overall policy goals. Finally, the team strategy weighs in on the probability. This gives the effect of individualized performance with overall short-term goals and team performance with match-wide long-term goals. *This is critical to the real-world performance of the system: it must emphasize individualized activities but constrain them (or at least influence them) by the team's overall goals (as realized in the team strategy).*

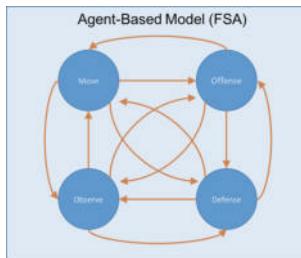


Fig. 1. Agent finite state automaton model

There are many ways that the actions of agents can be modeled, so our framework offers flexibility to adapt to these many situations. There are similar phases for many different types of models, and the sample we offer is one of those. The four phases of move, offense, defense, and observe can be used in many different scenarios from navigation (move means navigate towards destination, offense means seek sub-goals, defense means avoid obstacles, observe means gather intel), military (move means take action, offense means fire, defense means avoid, and observe means gather intel), to many others, such as autonomous driving, sports, or power management for buildings (move means adjust settings,

offense means take action to achieve goal, defense means to adjust environment to limit loss, observe means to take measurements). As should be clear, even this simple sample model has many applications. Further, the sub-elements of these phases can offer increased complexity.

In the sample FSA shown in Fig. 1, the agent starts in any of the states. By default, this is usually the move state. They will move through each of the states in any order (though the model could be modified or customized to have the agent process each phase differently). In each phase they have a probability of taking the designated action (i.e., invoking the underlying engine for each state of the agent model). This may mean choosing an offensive action (handled by the Targeting Matrix Model (TM)), choosing a defensive action (handled by the Threat Analysis Model (Threat)), making observations (handled by the SCAN Diagram Model (SCAN)), or choosing to move (handled by their Behavior Vector Model (BV)). The TM model scans through the ‘lanes’ visible from the agent’s location to see if there are any opposing agents in view. If so, they are targeted and selected to be fired upon. The Threat model scans the same ‘lanes’ to see if there are any incoming threats (either from the agent ‘taking fire’ or from opposing agents approaching the agent’s location). If there are threats, take cover. If not, choose the optimal posture (e.g., left, right, etc.) for the situation (e.g., if there are no targets on one side, switch to the other). The SCAN model searches the visible area for the agent to find any relevant observations (other agent locations, agent transitions, flag locations, teammate locations, etc.) and report them back to the FSA coordinating the agent behavior. They have a certainty of observation (meaning that there is not a probability on scanning, only on whether or not they will see something). This can be passive, where they note the probable zones from which they are taking fire), or active (by direct observations of the agents within their view). The key to the inference engine’s performance is the ability to directly observe the other agents, their transitions, and any other relevant information that can be derived from the other agents. Finally, the BV model is this particular agent’s subgraph of the Total Movement Dependency Diagram (the model of all possible movement for any agent in the simulation space). This model governs the agent’s movement within their assigned area of the simulation space. Each of these models can be considered sub-models of the agent’s FSA.

The basic agent FSA can then be expanded to show the sub-models (also FSAs) as shown in Fig. 2. These various models and sub-models are segmented for clarification in this figure. Each of these sub-models are also FSAs, though their specific models are not shown here. Moving up a level in the hierarchy, the agent models are assigned (as behaviors) to the agent by their policy.

The expanded policy model, shown in Fig. 3a, shows that the policy selects the behavior that is to be assigned to the agent from the set of available behaviors. This behavior is then assigned (i.e., implemented) in the agent. Finally, any evaluation data is sent back to the selection engine (this may include performance data on how this policy is working so far). The evaluation mechanism for the policy (i.e., it evaluates the quality and performance of the behavior) is the

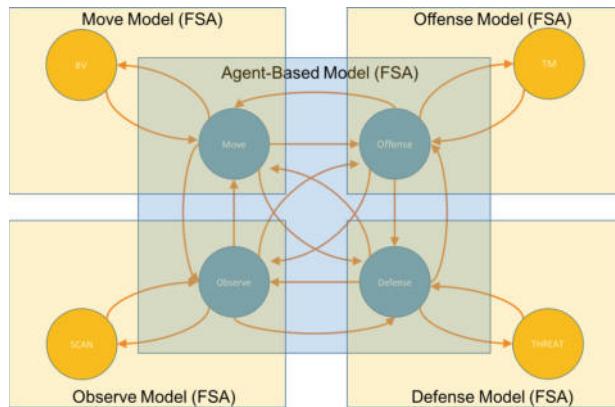


Fig. 2. Agent model FSA w/ sub-models

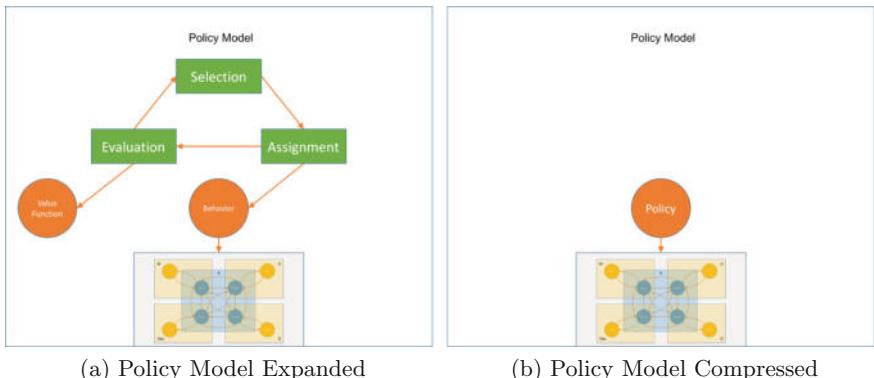
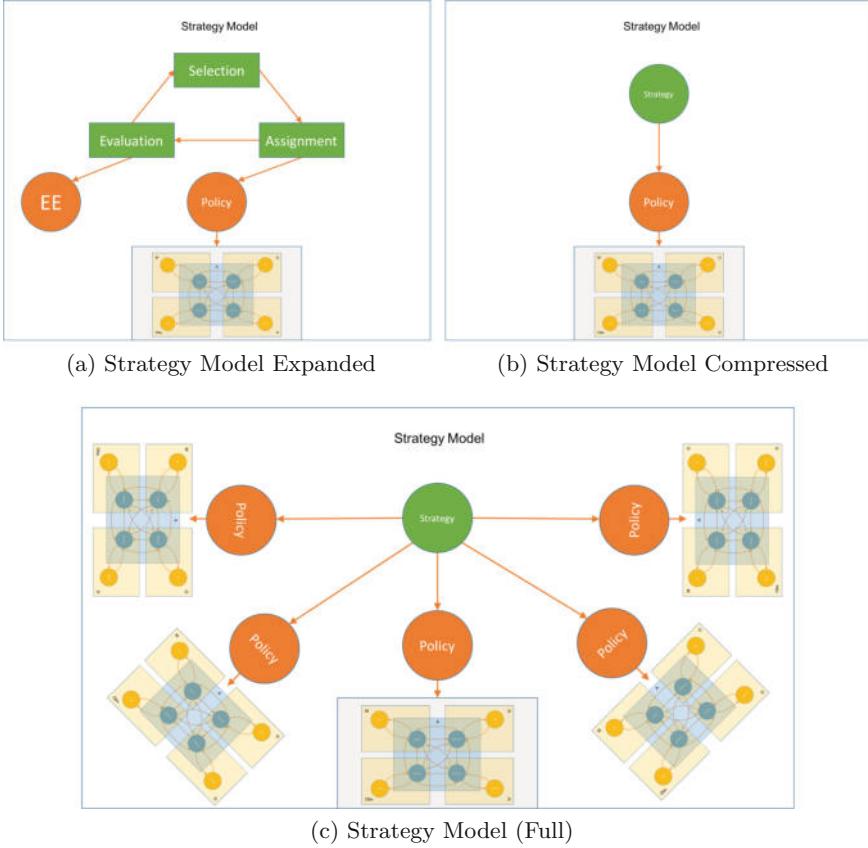


Fig. 3. Policy model

Value Function. This function may take on many forms, but it is a methodology of evaluating the efficacy of the behavior. This policy model can be compressed into a single element before going up to the next level of the hierarchy, as shown in Fig. 3b. The policy is assigned to the agent by the strategy.

The strategy model is shown in Fig. 4a. As with the policy model, the strategy model first selects the given policy from the subset of policies available to the strategy. It then assigns that policy to the particular agent during the assignment phase. This phase loads the policy into the policy model for that agent. The Evaluation Engine evaluates the performance of the policy and reports those findings to the strategy, and through the strategy to the Intelligent Strategy Selection Engine. This model can also be compressed, as in Fig. 4b, so that the full effect of the strategy can be seen. In Fig. 4c it can be seen that the strategy is assigning a policy to each agent on the team (as a reminder, the strategy can be viewed as a team policy). The strategy is assigned by the intelligence.

**Fig. 4.** Strategy model

The intelligence model is shown in Fig. 5a. Like the other layers, it starts with a selection phase that chooses the strategy from the set of strategies available to assign to each team. The assignment phase performs this operation and puts the strategy into effect for the team. The evaluation is performed by the Intelligent Strategy Selection Engine (ISSE). The ISSE monitors the performance of the strategy as well as determining the most likely strategy that other teams are following. This evaluation is forwarded to the intelligence in case a new selection is needed. As before, this intelligence model can be compressed as in Fig. 5b so that a larger view of the intelligence model can be viewed. This final view can be seen in Fig. 5c. This figure shows a sample model with three teams. In this example, these teams would be cooperating since they share a central intelligence model. If there were another alliance (i.e., another set of teams) then they would have their own intelligence model. This final figure gives some perspective for the entire model used in the simulation.

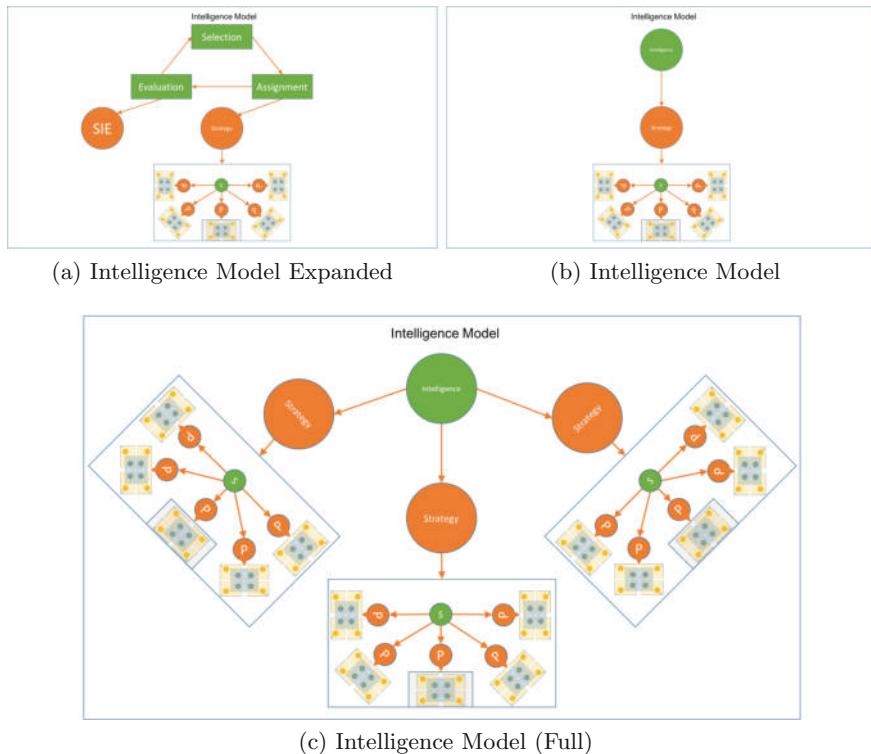


Fig. 5. Intelligence model

This hierarchical arrangement of models allows for the processing of information at each level of the performance model rather than having to utilize one monolithic policy for the entire structure.

As previously mentioned, the models can be customized based on the scenario being simulated, but these figures show how the hierarchical modeling works within the simulation framework. These models will also be used for strategy inference. In strategy inference each model forms a graph structure and the algorithm matches partial observations with known models to form a belief network about which of these various models the agents believe they are seeing.

4 Movement Dependency Diagrams

One critical element to understanding and deploying strategy within a simulated environment is the Movement Dependency Diagram (MDD). The MDD is a diagram that is the result of a search through the entire state space (in each system, states are defined by that system as locations, positions, or situations

where an agent is located and from which they can take actions to shift their state). Generically, the state space of an environment is the list of connected states that are reachable through every possible action. In strategic simulations this is the entire list of movements - that is, any and every action that moves an agent from one position to another. This creates a diagram that shows all possible movements and all possible subsequent movements from those, thus creating a Total Movement Dependency Diagram. The Total Movement Dependency Diagram can be made into sub-graphs where each sub-graph contains a particular set of connected moves within the larger set of moves. These sub-graphs can then be tied to certain strategic behavior (e.g., playing a particular position in soccer).

By way of example, consider the following input to the system. The simulation contains a model called the Total Movement Dependency Diagram (TMDD), of which a particular agent's MDD is a subset. The simulation is also given the set of all possible states - in this example, they are locations or positions within the simulation space. Given the agent's model, the simulation will iterate through the phases of the model. The behavior selection machine is concerned with the movement phase, so this example will focus on that phase but will inform other phases along the way for completeness. The MDD that forms the core of the behavior is provided the current state (part of the input into the system, initially s_0) and then informs the simulation what the next move (or moves in a non-deterministic environment) is (or are) from the current state. A transition function receives these moves and probabilities and processes them. If the probability threshold is not passed, the s' , or the next move, is the state output from choosing the movement selected. If the probabilities exceed the threshold for the move probability then the output is ε , an empty move (i.e., the next state equals the current state, or $s' = s$). As indicated, these moves, either the next state or the empty symbol, are output from the simulation. As a note for completeness, while the agent model is being processed, the agent is also making observations of other agents that it sees on the field or within the simulation environment. These observations are output from the simulation as well and sent to the other segments of the SiMAMT framework.

To consider the example practically, some sample data is provided. Figure 6 shows the TMDD for one side of a soccer field. This represents all the places an agent could move within this system. Figure 7 shows one example MDD for an agent playing defense on the soccer field. This MDD can be extracted, shown in Fig. 8, and shown as a Behavior model, as shown in Fig. 9. Placing an agent at the initial state (here s_1 , in Fig. 10), the probable moves become m_0 and m_2 (shown in Fig. 11). If movement m_0 has a probability of 0.50, and the pseudo-random number generator (PRNG) comes up with a 0.60, then no move occurs using this movement. Similarly, if movement m_2 has a probability of 0.40, and the PRNG comes up with a 0.65, then no move occurs using this movement, either. In this case, an ε is generated as output. The agent remains

at s_1 (or, more accurately, the transition model issues a move from s_1 to s_1). The next time the agent model considers a move, the PRNG generates 0.40 and the movement m_0 is taken. In this case, the output is m_0 , indicating a transition from s_1 to s_5 . As stated before, there are other such calculations being run for each phase of the agent model. The agent continues to move through the phases of this model until the simulation reaches an accepting (or final) state, like the end of a period in the soccer match.

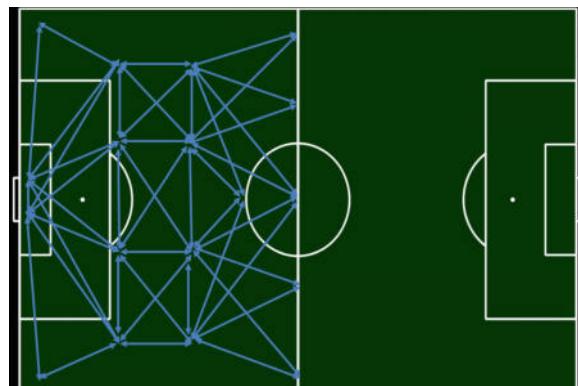


Fig. 6. Total movement dependency diagram for Soccer

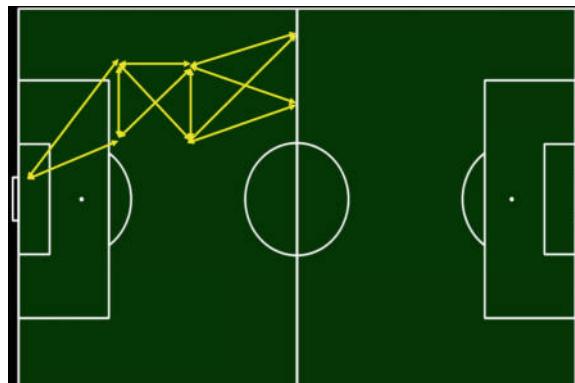


Fig. 7. Movement dependency diagram for a Soccer agent

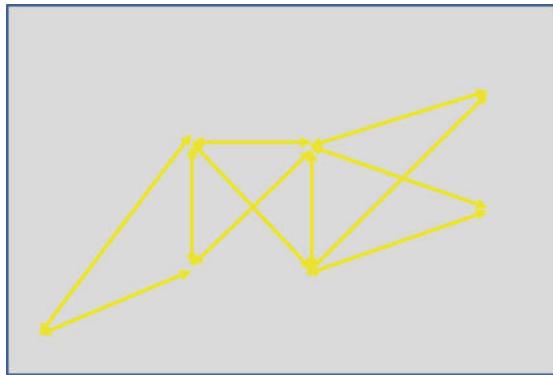


Fig. 8. Movement dependency diagram extracted

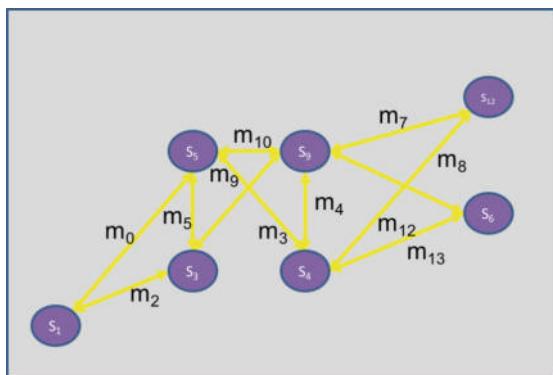


Fig. 9. Movement dependency diagram to behavior

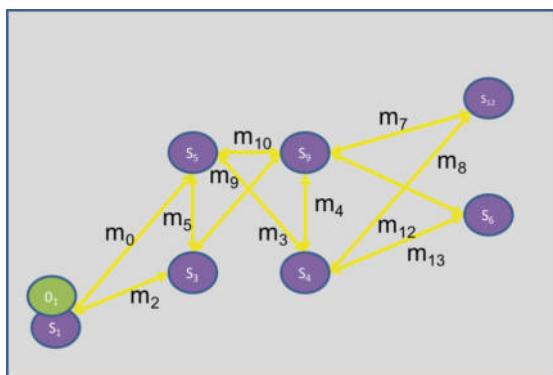


Fig. 10. Movement dependency diagram with agent

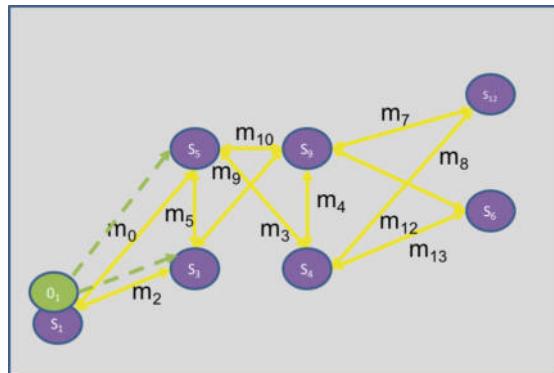


Fig. 11. Movement dependency diagram with agent moves

5 Example

The FSA for each agent controls their actions. The FSA makes decisions for each interval (the simulation can be run in either serial or parallel, but in either case the decision are made at each decision interval) during each round of the simulation. The configuration of the simulation can allow for round-robin processing, where each stage of the model driving the agent is considered in turn, or in parallel, where each action is processed and one is chosen to be performed at that specific time interval based on the weighting within the model. While we explain the steps here in order, this in no way precludes their parallel processing. First, we consider the **Move**. A probabilistic calculation is performed in accordance with the policy for that agent. The model holds a movement dependency diagram, and this model provides the next move, or the next set of moves and their probabilities. Once processed, a Movement is generated to provide a location for the move and a speed with which to move. If the agent is moving slowly they are more likely to be hit while fired upon, but their own aim probability increases. This is a balancing act that expresses the power of the features available to the agent (via their assigned behavior). In addition, the **Defense** phase is decided based on features of the agent's model like Aggression and Posture. These strategy features are considered during this phase while being factored by the individual agent's personality model. The agent may, once all probabilities have been factored, decide to stay under cover rather than make the intended move at the current interval. This defensive decision may overrule the Move phase decision depending on the agent's model or the team's overall strategy. Another phase of operation is **Offense**. This phase considers the similar agent model features as before, such as Aggression, and the defense phase's Posture consideration, in order to decide on the best offensive decision. The agent may decide during this phase to take an aggressive action, such as selecting a rate of fire or a change of tactic. Once a fire method and rate of fire have been selected the agent model works through the simulation's overarching probabilities of fire

accuracy, target selection, etc., to determine the efficacy of the agent's fire action. This interaction of multiple factors and features, different models and behaviors greatly increases the veracity of the simulation. Finally, the phase of **Observation** should be considered. In this phase the agent scans the alleys that they can observe for agents, agent movements or changes, or transitions. Any observations are then forwarded up the hierarchy to be aggregated across the team. If they cannot directly observe the enemy agents then they can infer their position from the inverse kinematic analysis of the enemy fire. This observation, as all observations are, are probabilistic and come with a measurement of fidelity, both of which factor into the upstream model considerations.

Once each phase is completed, either serially or in parallel, the agent can make a final decision as to which action should be taken. This is handled by the decision engine within the simulation. The next phase of the simulation is the Evaluation Engine (EE). This engine is designed to take the various observations from each agent and factor them to decide the next action that must be taken by the agent, the team, etc., depending on which level of the modeled hierarchy is being considered. Based on the observed information presented to the EE, the EE may decide that a change of policy is in order. This can be decided based on the disparity between the observed success of the current policy and the calculated success of each possible optional policy. If a change of policy is deemed necessary, this data is forwarded to the Strategy Inference Engine (SIE) to select the best matched strategy given the observed data.

To better understand the strategy factors involves examining them in context. For example, the Aggression speaks to how the strategy might overwhelm the policy to force a player to move where they might not have otherwise. It may also overwhelm their choice to take cover. Likewise, the Movement factor controls how likely they are to move and also in which direction. For example, the players may start to retreat (move backwards through their move list) as the number of active players remaining on their team diminishes. The Distribution often overrides move probabilities to move players closer to each other or farther apart depending on this setting. The Posture factor controls fire aggression, movement aggression, and likelihood to stand and fire vs. retreat. Finally, there is a Persistence factor that keeps the players on their current policy or frees them to move to another policy. These factors, as well as the subset of policies, differentiate the strategies one from another. It is important to note that the complexity of this environment is possible even with each agent in the simulation following the same model (though with different probabilities). If the models varied, the inference engines (both the EE and the SIE) would be even more effective because they would have more diversity in the observations provided for inference.

6 Conclusions

The modeling of hierarchical strategy-based multi-agent systems is indeed complicated; however, the provided models have achieved the desired goals. First, they offer high-fidelity — they accurately represent the real-world components

that they are modeling. Second, they are expressible — they create models and sub-models that allow for multitudinous combinations and permutations of hierarchical arrangements. Third, they are modular in their complexity — they match the complexity of the system by hierarchical arrangement of simpler modules rather than large, monolithic structures. Overall, the models are effective and perform well in multi-agent, multi-team environments.

7 Future Work

These models can be expanded further. The theory is presented here, with the hierarchical and aggregate modeling, to go to even more levels of layered progression, so that will be the next work. Additionally, there are always refinements to make with the models, so applying this work to other fields will help to expand the research.

References

1. Bowling, M., Veloso, M., et al.: Existence of multiagent equilibria with limited agents. *J. Artif. Intell. Res. (JAIR)* **22**, 353–384 (2004)
2. Greenwald, A., Hall, K., Serrano, R.: Correlated Q-learning. *ICML* **20**, 242 (2003)
3. Hu, J., Wellman, M., et al.: Multiagent reinforcement learning: theoretical framework and an algorithm. In: *Proceedings of the Fifteenth International Conference on Machine Learning*, vol. 242, p. 250. Citeseer (1998)
4. Lavers, K., Sukthankar, G., Aha, D.W., Molineaux, M., Darken, C., et al.: Improving offensive performance through opponent modeling. In: *AIIDE* (2009)
5. Littman, M.: Markov games as a framework for multi-agent reinforcement learning. In: *Proceedings of the Eleventh International Conference on Machine Learning*, vol. 157, p. 163 (1994)
6. Parson, Simon, Woolridge, Michael: Game theory and decision theory in multi-agent systems. *Auton. Agents Multi-Agent Syst.* **5**, 243–254 (2002)
7. Stone, P., Veloso, M.: Task decomposition and dynamic role assignment for real-time strategic teamwork. In: *Intelligent Agents V: Agents Theories, Architectures, and Languages*, pp. 293–308 (1999)
8. Treuille, A., Lewis, A., Popović, Z.: Model reduction for real-time fluids. In: *ACM SIGGRAPH 2006 Papers, SIGGRAPH '06*, pp. 826–834. ACM, New York (2006). 10.1145/1179352.1141962. <http://doi.acm.org/10.1145/1179352.1141962>
9. Treuille, A., Lewis, A., Popović, Z.: Model reduction for real-time fluids. *ACM Trans. Graph.* **25**(3), 826–834 (2006). 10.1145/1141911.1141962. <http://doi.acm.org/10.1145/1141911.1141962>



Bioinformatics Tools for PacBio Sequenced Amplicon Data Pre-processing and Target Sequence Extraction

Zeeshan Ahmed^{1,2(✉)}, Justin Pranulis¹, Saman Zeeshan³,
and Chew Yee Ngan³

¹ Genetics and Genome Science, School of Medicine,
University of Connecticut Health Center (UConn Health),
Farmington, CT 06032, USA
zahmed@uchc.edu

² Institute for Systems Genomics, School of Medicine,
University of Connecticut Health Center (UConn Health),
Farmington, CT 06032, USA

³ The Jackson Laboratory for Genomic Medicine,
Farmington, CT, USA

Abstract. Modern high throughput sequencing technologies are enormously contributing to the generation of heterogeneous genomic data of different sizes and kinds. In most of the cases, NGS data is first produced in the raw form, which is then demultiplexed into text based formats, representing nucleotide sequences i.e. FASTA and FASTQ formats for secondary analysis. One of the major challenges for the downstream analysis of amplicon data is to first demultiplex FASTQ files based on the different oligonucleotides barcode combinations. Match & Scratch Barcodes (MSB) are a set of interactive bioinformatics tools that support the analysis of PacBio sequenced long read amplicon data by detecting multiple forward and reverse end adapter sequences, generic adapters attached to the region specific oligos, multiple number of region specific oligos of variable length for the extraction of sequences of interest. These work with zero mismatch, retain only reads which map exactly to adapters and barcodes, report all sequences matched to both single and paired-end adapters and barcodes, and demultiplex FASTQ files based on the common and distinct barcodes combinations. The performance of MSB has been successfully tested using in-house sequenced non-published and external published datasets, which includes PacBio sequenced long read PDX (Patient-Derived Xenograft) amplicon data embedding multiple barcodes of variable lengths. MSB is user friendly and first interactively designed set of tools to empower non-computational scientists to demultiplex their own datasets and export results in different data formats (CSV, FASTA and FASTQ).

Keywords: Bioinformatics · PacBio · Amplicon data · Target sequences · Software

1 Introduction

Modern high throughput sequencing, also known as the next generation sequencing (NGS) technologies (e.g. PacBio, Illumina, 10xGenomics, Oxford Nanopore, ION Torrent, SOLiD etc.) have advanced and started enormously contributing to the generation of heterogeneous genomic data of different sizes, kinds, run types (single end, paired end and mate pair), quality scores (Phred Q), error rates and read lengths [1]. In most of the cases, NGS data is first produced in the raw form (e.g. BCL files for Illumina, Movie files for PacBio etc.), which is then demultiplexed into text based formats, representing nucleotide sequences i.e. FASTA (single definition line with description starting with a “>” sign, followed by one/multiple lines of sequence data) and FASTQ (structure based one/multiple sets of four lines: first begins with “@” sign and third with “+” sign and both are for optional description, second contains nucleotide sequence and fourth comprises of ASCII encoded Phred quality score for each nucleotide) formats for secondary analysis (Fig. 1). FASTQ has been one of the most widely used formats for NGS data quality checking, processing, secondary and downstream analysis.

Val	Char								
33	!	53	5	73	I	93]	113	q
34	"	54	6	74	J	94	^	114	R
35	#	55	7	75	K	95	-	115	S
36	\$	56	8	76	L	96	_	116	T
37	%	57	9	77	M	97	a	117	U
38	&	58	:	78	N	98	b	118	V
39	'	59	;	79	O	99	c	119	W
40	(60	<	80	P	100	d	120	X
41)	61	=	81	Q	101	e	121	Y
42	*	62	>	82	R	102	f	122	Z
43	+	63	?	83	S	103	g	123	{
44	,	64	@	84	T	104	h	124	
45	-	65	A	85	U	105	i	125	}
46	.	66	B	86	V	106	j	126	~
47	/	67	C	87	W	107	k		
48	0	68	D	88	X	108	l		
49	1	69	E	89	Y	109	m		
50	2	70	F	90	Z	110	n		
51	3	71	G	91	[111	o		
52	4	72	H	92	\	112	p		

Fig. 1. FASTQ file structure example (**a**), and ASCII encoded Phred quality score for each nucleotide (**b**).

While preparing libraries for sequencing (e.g. Genome sequencing, RNA-seq, ChIP-seq, ATAC-seq RIP-seq, methylation, targeted and amplicon sequencing etc.), based on different protocols, high quality commercial kits (e.g. SureSelect by Agilent Technologies, TruSeq by Illumina and SeqCap by Roche NimbleGen etc.) are used to fragment target sequences, transform target to double stranded DNA, ligate different oligonucleotides and quantify final product [2]. More and more recent NGS technology development uses generic adaptors or barcode sequences to maximize the utilization of

NGS throughput. All the steps are important but staying within the focus of this study, we are interested in analyzing combinations of oligonucleotides used for identifying specific genomic elements. In most of the cases, during library preparation, oligonucleotides are attached at the start and end of the sequences, and in both forward and reverse directions, which are then identified and trimmed out during secondary data analysis (bioinformatics).

Amplicon sequencing is ultra-deep sequencing of a targeted piece of DNA/RNA resulting from natural/artificial amplification [3]. It is widely used for the detection of genetic variations in downstream data analysis. The aim of this study is getting regions of interest sequenced by PacBio (sequencing technology, which can produce single end reads with maximum read length up to 40 kb, at <10 quality score and with error rates between 5 and 10%) by performing secondary data analysis on this amplicon data that contains multiple nucleotide indexes (adapters and barcodes) of different lengths and placements. While PacBio provides RS Long Amplicon Analysis (LAA) to perform barcode-demultiplexing (separating sequence reads into separate files for each index tag/sample) and de novo consensus sequencing of full-length amplicon, its currently available version is incapable to detect adapter sequences of variable lengths and randomly placed barcodes and primers of length lesser than 16 and 30 bp, especially if produced by the third-party vendors.

Many open source, freely available and published bioinformatics tools are available online, which are helpful in extensive evaluation of read index effects on NGS data, and identifying and removing contaminant oligonucleotide sequences [4] e.g. NGSSUtils [5], Trimmomatic [6], SeqPurge [7], Atropos [8], Flexbar [9], SeqAn [10], KNIME4NGS [11], FASTX-Toolkit, condetri, NextGen etc. However, most of the available tools are command line based, do not allow users to input multiple combinations of paired end oligonucleotides of various lengths and placements, designed for Illumina short reads, do not offer the identification of target sequences but clipping of sequence adaptors, not engaging user at each step of analysis for cross validation of results, do not offer cross platform graphical user interface for handling several layers of barcodes, and not solely designed for performing secondary analysis of PacBio LAA data. Most importantly, we were not able to find any tool that meets all our requirements.

2 Methods

The research and developmental (R&D) objectives of this study are to implement a new bioinformatics tool for the analysis of PacBio sequenced amplicon data with generic adapter sequences, forward and reverse end adaptors, and barcodes of multiple numbers and variable sizes for extracting the target sequence (Fig. 2). Supporting our R&D objectives, we have proposed new methodology and have implemented as a set of bioinformatics tools [12], which are Match Barcodes (MB) (Figs. 3 and 4) and Scratch Barcodes (SB) (Fig. 5).

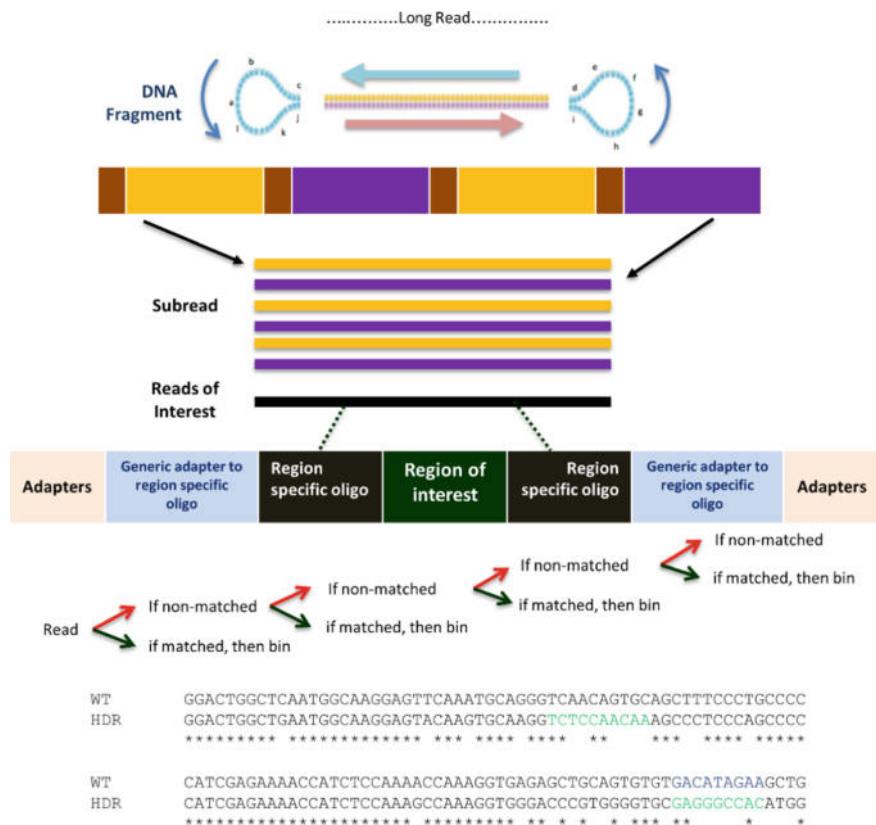


Fig. 2. Concept of PacBio sequenced amplicon data analysis using Match & Scratch Barcode. Multiplexed PacBio sequenced data containing oligonucleotides obtained following PacBio Reads of Insert protocol and analyzed using Match & Scratch Barcodes applications. The oligonucleotides in sequenced data are divided into single and paired end sequence adapters, generic adapters to region specific oligo, region specific oligos and regions of interests.

PacBio sequenced data (SMRT cells), demultiplexed into FASTQ format with the use of PacBio Reads of Insert protocol, containing distinct and paired end combinations of variable sequence adapters and barcodes combinations. Input and analyzed using MB and produced output contained information about unidentified (mismatched) sequences, and identified sequences with common and distinct single end sequence adapters, paired end sequence adapters, single end barcodes and paired end barcodes combinations. Sequences based on common and distinct combinations demultiplexed in FASTQ, FASTA and CSV files.

Proposed and implemented methodology of MB is based on different sequence data processing steps for the extraction of target sequence (Fig. 4). It takes FASTQ file as an input, parses and analyzes each sequence individual basis and performs complex comparative analysis. It gives zero mismatch reads by filtering out sequences without any match or mapped to incorrect adapter and barcode sequences. It detects user given

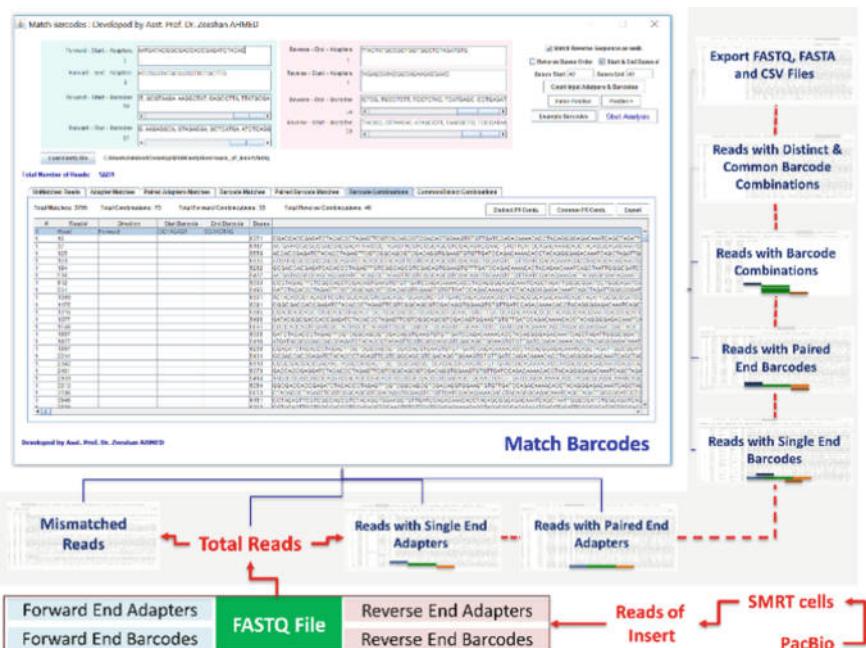


Fig. 3. Graphical user interface, data processing steps and outcome of Match Barcode. PacBio sequenced data (SMRT cells), demultiplexed into FASTQ format with the use of PacBio Reads of Insert protocol, containing distinct and paired end combinations of variable sequence adapters and barcodes combinations. Input and analyzed using Match Barcodes and produced output contained information about unidentified (mismatched) sequences, and sequences identified with common and distinct single end sequence adapters, paired end sequence adapters, single end barcodes, and paired end barcodes combinations. Sequences based on common and distinct combinations demultiplexed in FASTQ, FASTA and CSV files.

adapters and barcode sequences with global alignments i.e. all letters of the subject sequence must be aligned to the query sequence. Gaps are not permitted. It reports all sequences matched to the single or paired end adapters and barcodes, and demultiplexes FASTQ files based on the common and distinct barcodes combinations.

MB provides the following features: (1) adding forward and reverse sequence adapters at the start and end of sequence, (2) adding multiple (comma separated) forward and reverse barcodes, at the start and end of sequence, (3) loading FASTQ file (input, amplicon data) and press “Start Analysis” button to filter out mismatched sequences, matched sequences to any of the input sequence adapters, matched sequences to the paired end sequence adapters, matched sequences to any of the input barcode and matched sequences to paired sequence barcodes, (4) create combinations of extracted sequences based on common and distinct barcodes, and (5) demultiplex target regions in one or multiple “CSV”, “FASTA” and “FASTQ” files. Additionally, it allows user to calculate false-positive and actual positive number, count input adapters and barcodes, consider specific regions for analysis by setting start and end base pairs limit for forward and reverse directions.

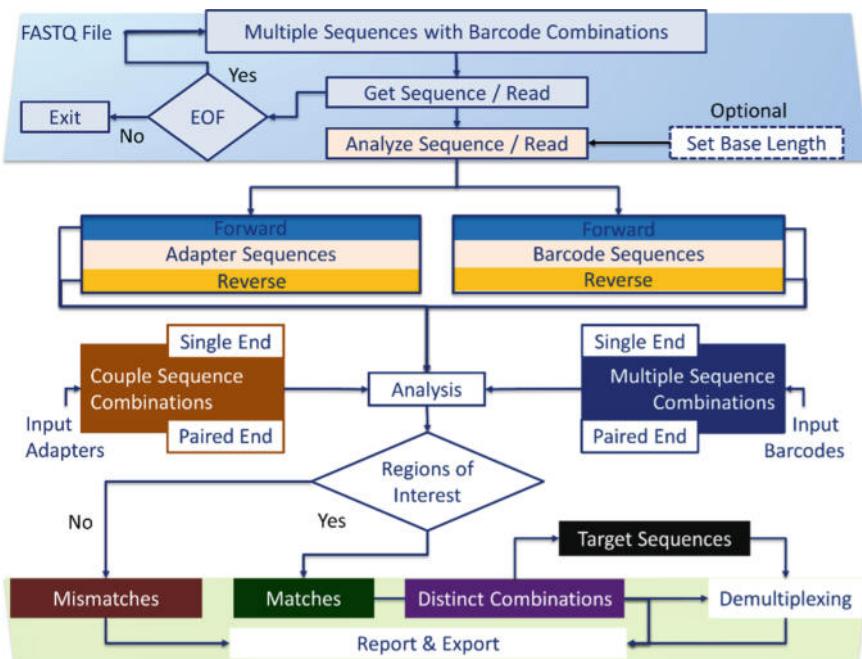


Fig. 4. Work flow of Match Barcode. FASTQ is the input to the MB, which parsed and each sequence (read) is separately analyzed, till the end of file (EOF). Complex comparative analysis is performed, based on the user input Individual as well as all possible combinations of sequence adapters and barcodes. Resultant mismatches and matches are reported to the user, with the given options (interactive graphical user interface) to further get distinct and common combinations to demultiplex target sequences.

SB is another handful tool (Fig. 5), which can read any number of FASTQ files as input and match maximum four combinations of adapter sequences. It can directly match sequences as well as match in an order, where it matches first set of combinations with reads and matches second with only mismatched reads and vice versa for others. SB provide features which allows user to add sequence adapters, load FASTQ files, analyze and report results, including total number of reads, total mismatched reads, matched reads with all four sequence adapters and export results in “CSV” format.

MSB are accessible through a graphical user interface and is designed by following software engineering principles for the sustainable bioinformatics software implementation [13–15]. MSB are programmed in Java programming language, built and tested Mac-OS-X and Windows platforms. To successfully run (JAR file), it requires Java Version 8 with Update 131 (build 1.8.0_131-b11), which can be downloaded from www.java.com. Best screen resolution for designed graphical user interface is with “1680 × 1050” OR “More Space” (System Preference → Display → More Space). MSB source code is freely available to download at https://github.com/drzeeshanahmed/match_scratch_barcodes/wiki.

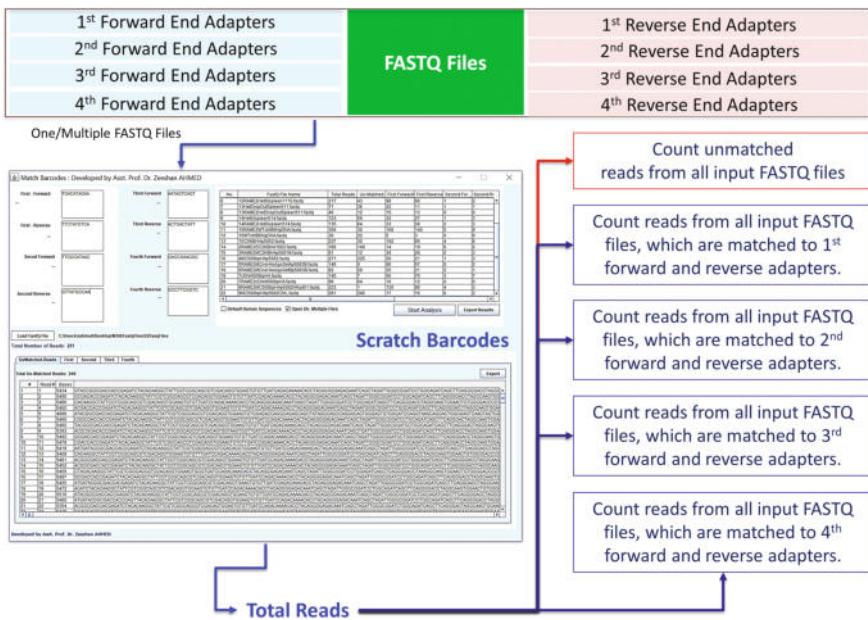


Fig. 5. Graphical user interface, data processing steps and outcomes of Scratch Barcode. Match Barcode's demultiplexed data (FASTQ files) input and analyzed using Scratch Barcodes, with the application for four different combinations of sequence adapters. Produced output contained information about the count of unidentified (mismatched) sequences as well as sequences identified with common and distinct single and paired end sequence adapters.

3 Results

The performance of MSB has been successfully tested using in-house sequenced non-published and external published datasets, which includes PacBio sequenced long read PDX (Patient-Derived Xenograft) amplicon data embedding multiple barcodes of variable lengths.

3.1 In-House Study

We experimented and produced PacBio sequenced long read PDX (Patient-Derived Xenograft) amplicon data and analyzed the generated FASTQ file (Reads Of Insert) using MB and SB. FASTQ file includes reads with customized barcodes in the beginning/end in forward orientation, and reads with barcodes in the beginning/end in reverse orientation. We analyzed FASTQ with MB and used customized barcodes, with focus on targeted (human and mouse sequences) barcodes identification and demultiplexing. During analysis, we considered input sequences both in forwards and reverse directions, and considered first 40 bases from start and end to identify adapter sequences.

This approach assists in systematic analysis of PacBio multiplex amplicon sequencing to allow precise identification of targeted regions. The MSB output for the tested data provides information at various levels of analysis. It gives the size of the construct with/without adaptors. It also gives the number of unmatched reads, reads with a single end matched barcode (forward or reverse), reads with a paired end matched barcodes (both forward and reverse) and the number of exact matched reads with specie specific distinct barcode combinations and mixed barcode combinations. Duplicate reads are removed to determine the unique reads for each set and each specie. Although it gives the maximum number of reads of a certain read length for a particular specie; however it does not identify and remove the chimera sequences (PacBio introduced sequencing artefacts).

MSB generates both matched and mismatched read sequences as an output, which can be exported in FASTQ format after trimming the barcodes/adaptors. The test results (Fig. 6) mainly includes total number of reads which are exact match to QC report generated by PacBio, number of Mismatched Reads, total number of Single End Adapters, total number of Paired End Adapters, total number of Single End Barcodes, (3705) total number of Paired End Barcodes, number of Distinct Combinations, both Forward (33) and Reverse, generated targeted output FASTQ/FASTA files. The results show that MSB uses a fast and exact method without any heuristics in finding all matches meeting the accepted criteria.

3.2 External Study

We have also used MB to analyze and report one of the publically available datasets. A published microbial study [16] presented a method that identifies microbe-microbe interaction found in plant communities, and claims for more efficient than traditional methods of microbial identification. The authors present a method that allows cross-referencing of their data with community-independent analyses that have already been established. The method advances the concept of community-based culture collections (CBC) [17], in which microbial colonies are isolated and picked, whether they include one or multiple organisms, and stored in a single plate well. This helps to maintain any necessary microbe-microbe interaction [18] some microorganisms rely on. For their study, authors picked colonies from primary platings [19] of microbial rich fractions of the rhizosphere [20], endophytic root, and endophytic stalk. 5137 colonies were picked and stored in 96-well plates [16].

To annotate the CBCs and microbes within them, the authors incorporate a multiplex strategy, which implements a 2-step Polymerase chain reaction (PCR) [21] (laboratory technique to make multiple copies of a segment of DNA) procedure where the initial 96-well plates were run in the first step and the amplicons of the plates were then pooled into a single 96-well plate and ran in the second step; the amplicons of which were then pooled into a single tube to be sequenced. To tag the microorganism and to allow the authors to know what plate and well the microorganism came from, specific primers, called barcodes, were added to the 16S ribosomal gene [22] during the PCR. This gene is highly conserved among microbes and is often used in research for identification. The first step of the PCR added Nextera transposase [23] forward and reverse primers which amplified the 16S sequence to enhance the quality of biological

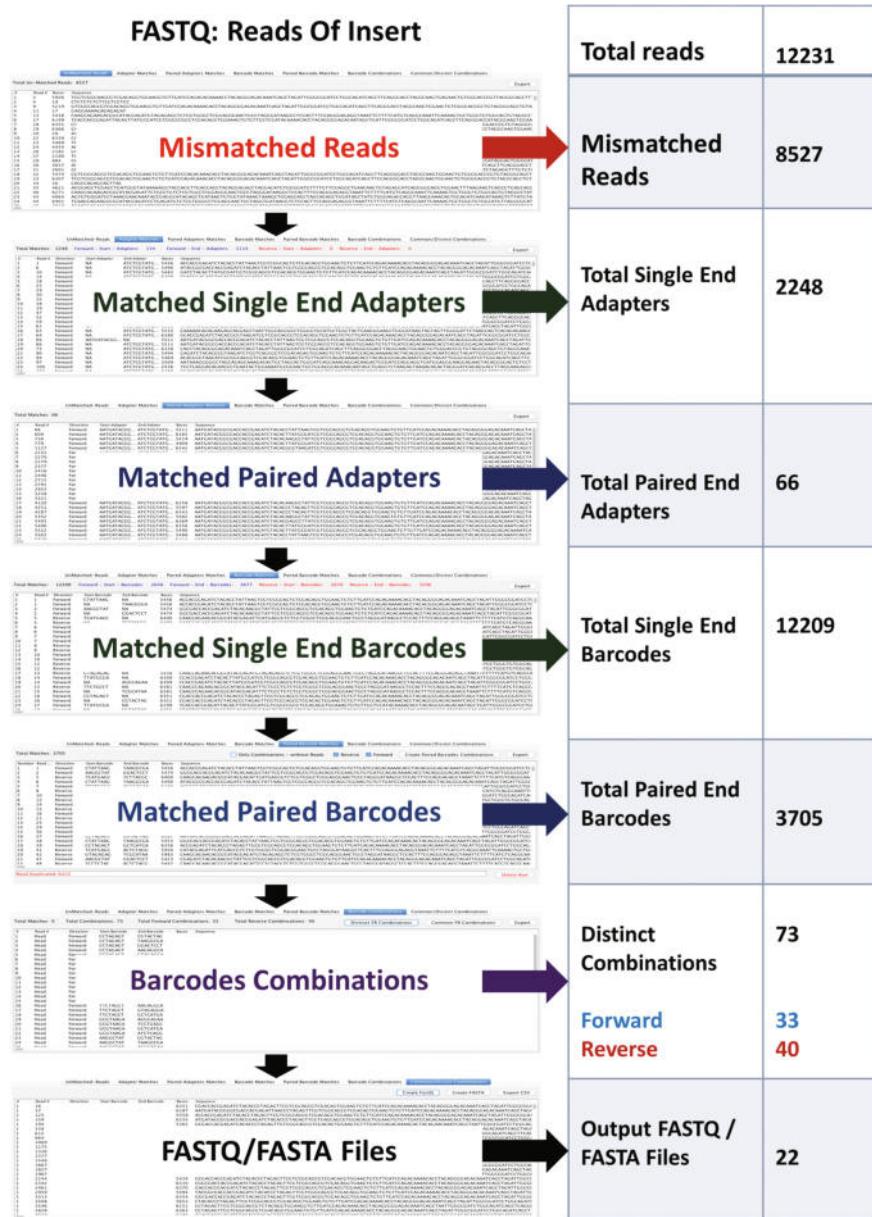


Fig. 6. Demultiplexing Reads of Insert with Match Barcode. FASTQ files given as input to the MB, compared and analyzed using two different paired end adapters and barcodes combinations.

information to improve the depth of taxonomic assignment. The reverse primer included a 9 bp barcode specific to each plate. The second step then added barcodes specific for the rows and columns to the single 96-well plate the amplicons of the first step PCR were pooled into. Illumina SMRT bell adapters were also added in the second

step which allowed the 16S rRNA strand to form a circular strand, called circular consensus sequences (CCSs) [24] that allows for multiple passes of the sequencing polymerase; the more passes the polymerase made around the strand, also referred to as coverage, the more accurate the CCS was considered. The amplicons of the second step were then pooled into a single tube for sequencing [16].

The amplicons of the second step were sequenced with the PacBio RS II platform. Implementing the RS_ReadsOfInsert protocol from the PacBio SMRT Portal v2.1.1, raw sequence data was generated and assembled into CCSs with the parameters: minFullPass2 and minPredictedAccuracy90. Raw data was then stored as FASTA files which allows the CCSs to be stored as software script with an identification number followed by the nucleotide sequence of the CCS. Raw CCSs consisted of the amplified 16S rRNA, plate, and row/column barcodes. The total number of assembled CCSs with at least $2 \times$ coverage was 27,220. After the initial polymerase raw sequencing, the data output was processed using ‘in-house’ pipeline set by the authors to filter the data. The raw data was initially filtered by CCS length. Authors discarded any sequences <100 bp as well as sequences >1600 bp, as these sequences were assumed to be chimeras since the expected length of the CCSs was \sim 1300 bp. Authors also implemented the UCHIME [25] software package, which is a program that takes the raw data and cross-references it with a high-quality chimera-free reference database to assist in further filtering any possible chimeric sequences. The amount of usable CCSs after this filtering step was 10,290, which correlates with 377 recovered wells of the initial 480.

As previously mentioned the authors implemented a multiplex strategy using 2-step PCR to accomplish this. Using their ‘in-house’ script as input the authors used various software packages/algorithms to demultiplex their samples/data. This allowed the authors to identify the specific well, plate of origin, and primary plate the CBC was collected that any given CCS came from. Using the UBLAST [26] algorithm, which searches a query for a local alignment of a low identity ‘target’ sequence. Setting the CCS dataset as their query the authors searched for the Nextera transposase and non-barcoded sequences. This allowed the authors to predict the location of the barcodes on the CCS. Using the search_oligodb algorithm [16], which is used to search a query for shorter sequences, or oligonucleotides, such as primers or probes, the authors were able to find the location of the barcodes on the CCS.

At this point the authors are left with the trimmed 16S sequence which was run through a reliability filter they devised. Using the usearch_global command from the USEARCH software package, the CCSs are cross-referenced with the Greengenes 16S rRNA gene database. CCSs with a hit against the database were considered reliable. CCSs that did not get a hit were then aligned with the authors initial CCS dataset. If the CCS received a hit against the dataset it was also considered reliable, whereas CCSs that did not receive a hit against the Greengenes [27] database and the CCS dataset were discarded. The number of usable CCSs at this point was 9978, with 375 recovered wells out of 480. The authors then clustered the CCSs within each well independently to form operational taxonomic units (OTUs) [28] (classify groups of closely related individuals) using the UPARSE [29] pipeline with a 97% identity threshold. The OTUs were used for the taxonomical assignment, as well the cross referencing of the culture collection and culture independent community analysis and the recovery estimate. The authors were able to form 521 reliable OTUs and identified 34 distinct genera of microbes. The OTUs

were cross-referenced with 20,731 OTUs from previously described sugarcane microbiota by a culture-independent community analysis to give an estimate of the extent of the sugarcane microbiota diversity recovered by the authors. They showed their OTUs comprise of 13.3, 14.8, and 29.1% of the total bacterial relative abundance in the sugarcane rhizosphere, endophytic root, and endophytic stalk respectively. They also showed that their collection contained microbes that represent 12.2, 39.5, and 50.8% of the exophytic stalk, exophytic leaf, and endophytic leaf respectively. The authors also conclude that most of their isolates represent groups of bacteria that have not yet been studied for their association with plants.

We downloaded authors published dataset (FASTQ file) and their provided barcodes (illumine adaptors) at MB, and concluded with found unmatched reads, reads matched to single and paired barcodes, and adapters (Fig. 7). While analyzing FASTQ file with Nextera primers, we found matched reads to all the adapter sequences (Fig. 8). In review the authors present to the reader a primary paper that is relatively short, yet very descriptive and full of details on their procedures and protocols. The drawback of having so much detail in such a short report makes the reader loose connectivity as they skim through the primary report, materials and methods section, and have to rely on the supplementary data to get a clearer picture of the study and the order of events.

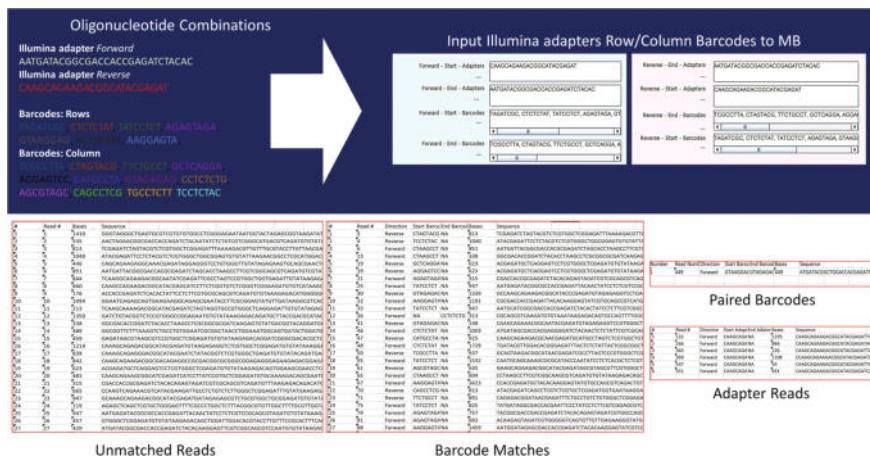


Fig. 7. MB applied to analyze dataset (FASTQ file) with provided Illumina adapters row/column barcodes.

In conclusion the authors did present a more efficient method for storing, isolating, and identifying microbial communities that is ideal for high-throughput and large-scale research studies. The authors also advance on the CBC concept and allow for the cross-referencing of data between the culture collections in question and culture-independent community analysis. Altogether their method helps to avoid the laborious, time consuming, and costly aspects that accompany the traditional methods of microbial identification, as well as give a better representation of true microbial communities found within microbiomes.

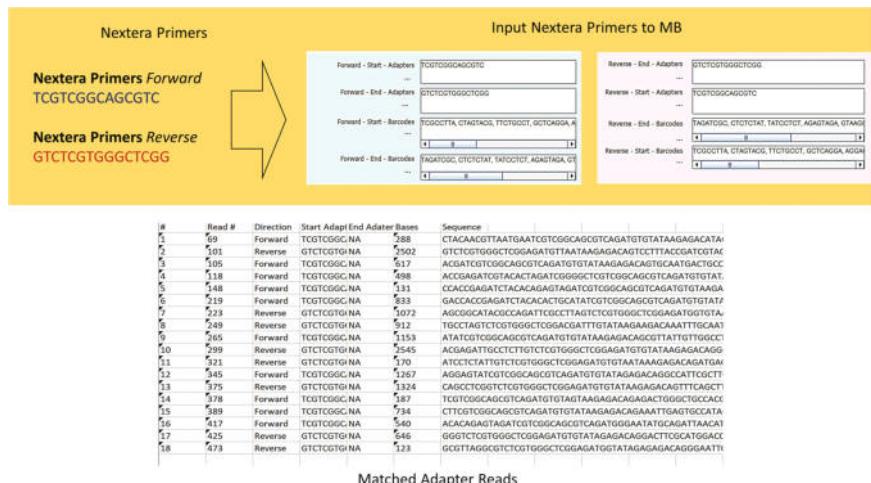


Fig. 8. MB applied to analyze dataset (FASTQ file) with provided Nextera primers.

4 Discussion

Whether its long or short reads, NGS data in most of the cases is first produced in the raw formats, which is then converted into nucleotide sequences and further demultiplexed into FASTA and FASTQ formats for secondary analysis and downstream analysis. During secondary analysis, it is very important to first recognize NGS data based on its type e.g. total-RNA-seq, mRNA-seq, tRNA-seq, microRNA-seq, WGS, WES, ChIP-seq, ATAC-seq, PDX-WGS, PDX-RNA-seq Iso-Seq, Single cell, CRISPR-CAS9, Amplicon data etc. Before starting sequencing using any modern technology (e.g. Illumina, PacBio SMRT-Cell, Oxford Nanopore, 10xGeomics and Ion Torrent etc.), based on the NGS data type, different protocols are used in wet lab/bench to create libraries of samples. While creating libraries different oligonucleotides barcode combinations are embedded into data. These barcodes help in recognizing data at secondary and even during downstream analysis.

Amplicon data is one of the complex kind of data. One of the major challenges for the downstream analysis of amplicon data is to first demultiplex FASTQ files based on the different oligonucleotides barcode combinations inside. Meeting set research objectives, it is very important to properly recognize and extract sequences from the amplicon data, to get help in finding the target sequences based on different barcodes inside, as the quality of end results of downstream analysis can highly depend upon this step as well. While doing our research and development, we found it extremely important and one of the difficult tasks to manually demultiplex, as most of the available software tools and technologies were not suitable for the conditions. Moreover, those were all command line based and we were not able to track complex demultiplexing process at each step. This motivated us in developing our own software

with user friendly graphical interface, which even a scientist with non-computational background can use. Our software helpful in secondary analysis and tracking the whole process of demultiplexing, where user can easily backtrack and trouble shoot as well.

5 Conclusion

We have presented bioinformatics method to identify adapter sequences and barcodes in FASTQ files and divide the reads accordingly. We have successfully tested MB and SB on PacBio sequenced long read amplicon data. Input to the MB was the FASTQ file generated using PacBio Reads of Insert protocol. Valuable results produced from both MB and SB, including quantitative figures and FASTQ files, were easily utilized for further downstream analysis. While these tools have been implemented and well tested with PacBio sequenced amplicon data, they are not technology dependent and can be applied to perform secondary analysis at sequence data produced by other sequencing technologies e.g. Illumina.

6 Competing Interests Statement

Authors have no competing interests to declare.

7 Contributorship Statement

ZA and CYN equally participated in defining methodology. ZA produced results which were analyzed and validated by CYN. ZA did all work on the software aspects of MSB. JP and SZ performed MSB validation and analysis at public dataset. All authors contributed in writing and review of manuscript.

Acknowledgements. We thank Ahmed lab, Department of Genetics and Genome Sciences, Institute for Systems Genomics (ISG), School of Medicine, University of Connecticut Health Center (UConn Health), and The Jackson Laboratory for Genomics Medicine for their support to ZA, SZ and CYN. We also thank Partnership for Innovation and Education (PIE) and Technology Incubation Program (TIP) for supporting JP at UConn Health. We appreciate all colleagues, who have provided insight and expertise that greatly assisted the research and development.

References

1. Escalona, M., Rocha, S., Posada, D.: A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* **17**, 459–469 (2016)
2. Head, S.R., Komori, H.K., LaMere, S.A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D.R., Ordoukhianian, P.: Library construction for next-generation sequencing: overviews and challenges. *Biotechniques* **56**, 2 (2014)

3. Boeva, V., Popova, T., Lienard, M., Toffoli, S., Kamal, M., Le Tourneau, C., Gentien, D., Servant, N., Gestraud, P., Rio Frio, T., Hupé, P., Barillot, E., Laes, J.F.: Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics* **31**, 3443–3450 (2014)
4. Del Fabbro, C., Scalabrin, S., Morgante, M., Giorgi, F.M.: An extensive evaluation of read trimming effects on Illumina NGS data analysis. *Plos One* **8**, e85024 (2013)
5. Breese, M.R., Liu, Y.: NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* **29**, 494–496 (2013)
6. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinformatics* **30**, 15 (2014)
7. Sturm, M., Schroeder, C., Bauer, P.: SeqPurge: highly-sensitive adapter trimming for paired-end NGS data. *BMC Bioinf.* **17**, 208 (2016)
8. Didion, J.P., Martin, M., Collins F.S.: Atropos: specific, sensitive, and speedy trimming of sequencing reads. *PeerJ* **5**, e2452v3 (2017) (Preprints)
9. Dodt, M., Roehr, J.T., Ahmed, R., Dieterich, C.: FLEXBAR—flexible barcode and adapter processing for next-generation sequencing platforms. *Biology* **1**, 895–905 (2012)
10. Döring, A., Rocha, S., Posada, D.: SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinf.* **9**, 11 (2008)
11. Hastreiter, M., Jeske, T., Hoser, J., Kluge, M., Ahomaa, K., Friedl, M.S., Kopetzky, S.J., Quell, J.D., Werner Mewes, H., Küffner, R.: KNIME4NGS: a comprehensive toolbox for next generation sequencing analysis. *Bioinformatics* **33**, 1565–1567 (2017)
12. Ahmed, Z., Ngan, C.Y.: Match & Scratch Barcodes: tools for the demultiplexing and extraction of target sequences from PacBio amplicon data. *Nat. Methods* (2017)
13. Ahmed, Z., Zeeshan, S., Dandekar, T.: Developing sustainable software solutions for bioinformatics by the “Butterfly” paradigm. *F1000Research* **3**, 71 (2014)
14. Ahmed, Z., Zeeshan, S.: Cultivating software solutions development in the scientific academia. *Recent Pat. Comput. Sci.* **7**, 54–66 (2011)
15. Ahmed, Z.: Designing flexible gui to increase the acceptance rate of product data management systems in industry. *Int. J. Comput. Sci. Emerg. Technol.* **2**, 100–109 (2011)
16. Armanhi, J.S.L., de Souza, R.S.C., de Araújo, L.M., Okura, V.K., Mieczkowski, P., Imperial, J., Arruda, P.: Multiplex amplicon sequencing for microbe identification in community-based culture collections. *Sci. Rep.* **6**, 29543 (2016)
17. Armanhi, J.S.L., de Souza, R.S.C., Damasceno, N.D.B., de Araújo, L.M., Imperial, J., Arruda, P.A.: Community-based culture collection for targeting novel plant growth-promoting bacteria from the sugarcane microbiome. *Front. Plant Sci.* **8**, 2191 (2017)
18. Wolin, M.J., Miller, T.L., Stewart, C.S.: Microbe-microbe interactions. In: Hobson, P.N., Stewart, C.S. (eds.) *The Rumen Microbial Ecosystem*. Springer, Dordrecht (1997)
19. Sanders, E.R.: Aseptic laboratory techniques: plating methods. *J. Visual. Exp: JoVE* **63**, 3064 (2012)
20. McNear Jr., D.H.: The rhizosphere—roots, soil and everything in between. *Nat. Educ. Knowl.* **4**(3), 1 (2013)
21. Bartlett, J.M., Stirling, D.: A short history of the polymerase chain reaction. *Methods Mol. Biol.* **226**, 3–6 (2003)
22. Janda, J.M., Abbott, S.L.: 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J. Clin. Microbiol.* **45**(9), 2761–2764 (2007)
23. Kia, A., Gloeckner, C., Osothprarop, T., Gormley, N., Bomati, E., Stephenson, M., Goryshin, I., He, M.M.: Improved genome sequencing using an engineered transposase. *BMC Biotechnol.* **17**, 6 (2017)

24. Grohme, M.A., Soler, R.F., Wink, M., Frohme, M.: Microsatellite marker discovery using single molecule real-time circular consensus sequencing on the Pacific Biosciences RS. *Biotechniques* **55**, 253–256 (2013)
25. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R.: UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**(16), 2194–2200 (2011)
26. Edgar, R.C.: Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010)
27. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L.: Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**(7), 5069–5072 (2006)
28. Sokal, R.R., Sneath, P.H.A.: Principles of numerical taxonomy. W.H. Freeman, San Francisco (1963)
29. Edgar, R.C.: UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10**, 996–998 (2013)



SimTee: An Automated Environment for Simulation and Analysis of Requirements

Saad Zafar^(✉), Musharif Ahmed, Taskeen Fatima, and Zohra Aslam

Faculty of Computing, Riphah International University, Islamabad, Pakistan
`{saad.zafar, musharif.ahmed}@riphah.edu.pk`

Abstract. The mathematical nature of Formal Methods make them more amenable for machine assisted analysis. However, the exhaustive model-checking and theorem-proving of the complete specification remains an elusive target due to the state-explosion problem. Simulation or execution of formally specified requirements provides us a less expensive alternate to understand, analyze and validate requirements early in the development phase. In this paper, we present SimTree simulator that carries out requirements analysis. SimTree code is generated from automatically transforming Behavior Trees (BT)—a graphical formal notation—using ATLAS Transformation Language (ATL). During the step-by-step execution of BT, Datalog code is also generated. The Datalog queries are used to further analyze the stored state-space of the executed requirements. These features of the simulator are illustrated using a published case study. The simulator was useful for identifying and rectifying logical defects in the specification.

Keywords: Behavior trees · Simulation · Datalog · Requirements engineering · Formal methods · ATLAS transformation language

1 Introduction

The mathematical nature of formal methods provides us the foundation of more rigorous analysis of requirements. However, due to this very quality they are difficult for analyst and users to understand and use [1]. A significant level of expertise and effort is required to model and analyze formal specification. A number of graphical notations with formal semantics have been introduced to improve the usability and acceptability of formal methods [2–4]. The graphically specified requirements are generally perceived to be easy to construct, communicate and validate. But the real power of the formal method is in machine assisted analysis [5]. With the advent of more powerful and more inexpensive hardware, the use of formal methods has gained renewed importance. Nonetheless, hardware still remains a serious bottleneck in performing exhaustive verification and analysis of complete specification due to the ever persistent state-explosion problem [6]. To overcome the problem, researchers have proposed limited use of lightweight formal methods where only the critical aspects of the specification are formally modeled and analyzed [7]. Requirements execution or

simulation provides a middle ground where the formal basis of the notion is exploited to execute the requirements in order to gain an insight into the system being modeled. The simulation can not only be useful for early defect detection but can also play a crucial role in requirements elicitation and validation [8].

In this paper we present a tool environment that is developed for executing requirements. We use Behavior Trees (BT)—a graphical formal notation—for specifying requirements and executing them through automatically generated code using a simulator called SimTree. The tree-like notation is developed with the aim of improving readability and understandability of the specification. BT notation is part of an overall requirements development methodology referred to as Behavior Engineering (BE) [9] in which the requirements are systematically translated from informal language into its formal representation, thereby, preserving the intent and traceability between the two representations [9]. The limited industrial trials of the approach report a high yield in early defect detection due to the systematic translation approach of BE [10]. Different tools have been developed [11, 12] that help in developing BT models and automatically translating them into other formalisms for automated analysis like theorem proving, model-checking and for verification of other properties of interest [13, 14].

SimTree is also used to generate Datalog [15] code. Datalog is a rule based query language for deductive databases. We translate the BT specification into facts and rules during run time using the simulator. The resultant database allows us to query for different properties of interest and for generating traces of our choice. We illustrate the use of these features using a simple case study called one-minute microwave oven [4]. Our main contribution is the SimTree tool that does automatic translation of BT into Datalog. The simulator enriches the BT formal analysis provided by previous approaches by adding simulation capability. The simulator's environment is interactive and has the ability to configure number of complete runs for the execution of the specified requirements. Rest of the paper is structured as follows. Section 2 provides the required background. Section 3 outlines the model transformation approach used for generating the simulator code. Section 4 provides detail of the SimTree tool environment using a published case study as a running example. Section 5 concludes the paper along with overview of our future work.

2 Background

2.1 Requirement Simulation

“Simulation is a proven technique and an intuitive means to assess the behavior of a system” [16]. Such formal verification approaches do not require advanced mathematical training and theorem proving skills [17]. The simulators vary from being tools for executing the formal specifications to the consistency checkers for checking the well-formedness of the specification. One of the advantages of simulation is that it is inexpensive and can be of major help in pinning down requirements defects early in the product lifecycle [18]. By means of simulation analyst can know about the behavior of the system by interacting with the simulator. The interaction can take the form of

specifying safety properties or entering of probability values to the events. Apart from that the simulation, traces can be generated that can later on be stepped through to inspect the system for the causes of errors and defects. Moreover, simulation can play an instrumental role during the specification process in the step wise refinement [19]. The results of different simulation runs can then be analyzed to look for defects and to deduce information about the structure of the described behavior. Also simulation is a valuable tool in defining the initial requirements of the system by validation and verification process and can be used to test out alternate modifications before the implementation phase [15]. Simulation which is a light weight approach is straight forward and easy, its analysis is however shallow due to its inability to extrapolate all the states and behaviors of the system [16].

2.2 Datalog

Datalog is a subset of Prolog with its syntax very similar to it [17]. It is a rule based query language for deductive databases. It is based on first order logic and is declarative [18]. Datalog is simple due to which it supports rich analysis. It has universally agreed syntax and its clean semantics allow for the better reasoning about the problem specifications. On the other hand its neat formulations allow for a better understanding when it comes to using recursive predicates [17]. It provides more expressivity and allows for better program maintenance. Datalog program consists of rules and facts.

2.3 Behavior Trees

Behavior Trees is a graphical formal specification notation that has been introduced in the literature with the aim to reduce the informal-formal gap [20]. It advocates Behavior Engineering (BE) process that builds specifications out of its requirements. Each requirement in natural language is translated into corresponding Behavior Tree word by word as a single Requirement Behavior Tree (RBT). The translation involves identification of components, states and their behavior. This simple process helps in moving from requirements to design in a piece wise incremental way [9]. Once all the requirements have been translated, they are joined together into a single Integrated Behavior Tree (IBT). The process involves satisfaction of pre and post conditions of RBTs such that errors are rectified and integration defects are removed [21]. This integrated view is very helpful for the modeler in visualizing any change during the requirements correction. As per industrial trials, it has been found to be useful in the defect detection [10].

Behavior Trees is a graphical formal specification notation that has been introduced in the literature with the aim to reduce the informal-formal gap [22]. It advocates Behavior Engineering (BE) process that builds specifications out of its requirements. Each requirement in natural language is translated into corresponding Behavior Tree word by word as a single Requirement Behavior Tree (RBT). The translation involves identification of components, states and their behavior. This simple process helps in moving from requirements to design in a piece wise incremental way [9]. Once all the requirements have been translated, they are joined together into a single Integrated Behavior Tree (IBT). The process involves satisfaction of pre and post conditions of

RBTs such that errors are rectified and integration defects are removed [23]. This integrated view is very helpful for the modeler in visualizing any change during the requirements correction. As per industrial trials, it has been found to be useful in the defect detection [10].

A BT node is shown in Fig. 1, consisting of *component name*, *behavior*, *tag* and *operator*. Generally a node is composed of a *component* that realizes a *behavior*. The behavior of a component can be; state realization, selection, event, guard, internal input, internal output, external input and external output. The state realization (Fig. 2a) is written as C [s] showing that the component C realizes the state s. Selection (Fig. 2b) works similar to an *if statement* and is written as C?s?—that is to say the control proceeds to the next node if the selection becomes true otherwise the control on the branch is terminated. An event (Fig. 2c) is the occurrence of an external event—C??e?? The control is passed to the subsequent nodes only when the external event happens. On the other hand, a guard (Fig. 2d) is represented by triple question mark—C???s??? It waits for the internal state s to become true in order to pass control to the next nodes. Internal input (Fig. 2e) C>s< waits until the message s is received while internal output (Fig. 2f) C ><s>< sends the message s inside a BT. Similarly external input (Fig. 2g) C >>s<< waits until the message s is received while external output (Fig. 2h) C <<s>> sends the message s to the environment that is outside the BT. Node operators are used to control the concurrent threads. Node operators are defined in the originating node which matches a destination node having same *component*, *behavior* and *behavior type*. The reversion operator (Fig. 2i) “^” indicates that the flow of the control reverts back to the matching ancestor nodes higher up in the same branch; however the siblings branches are terminated. The reference operator (Fig. 2j) “=>” allows the control to pass on the destination node, however the destination node must be in the alternative branch to the source node. The branch kill operator (Fig. 2k) represented by “-“ is used to terminate all the branches after the destination node. The destination node should be in the alternative branch to the source node. The synchronization operator (Fig. 2l) that is represented by “=” allows the control flow to continue after the destination node has been executed.

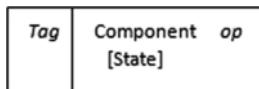


Fig. 1. A BT node

The nodes are joined together using edges which capture the control flow throughout the BT. A sequential edge (Fig. 2m) is represented by an arrow originating from one node to the other. It allows for interleaving of concurrent thread in between the sequentially composed nodes. Nodes can be composed atomically (Fig. 2n) and are represented by two nodes conjoined together without an edge. The atomically composed nodes form a single group in which the control flow is uninterruptable and interleaving of concurrent threads is not possible. Nodes in multiple branches are also composed together using either parallel or alternate branching. Branches which are

composed in parallel branching (Fig. 2o) allow for the creation of concurrent threads while those composed in alternate branching (Fig. 2p) allow for only one branch to succeed non-deterministically. Once, one branch is succeeded in getting the control flow, all other branches are terminated in alternate branching. Once a complete IBT has been constructed, it can be translated into other formal specification languages with the intention to carry out formal analysis.

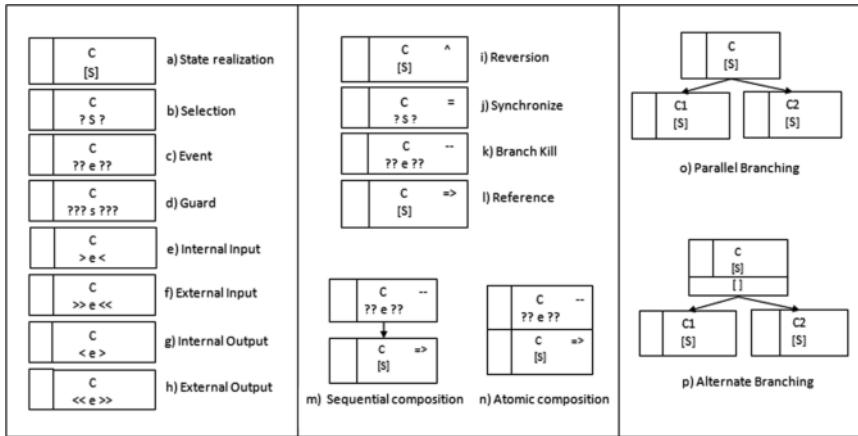


Fig. 2. Elements of BT notation

2.4 Related Work

BTs have been translated into couple of model checkers and theorem provers. One example is the translation into input language of Symbolic Analysis Laboratory (SAL) [13]. SAL supports theorem proving, deadlock checking and checking for well-formedness along with model checking. Analysis support for Failure Divergent Refinement (FDR) model checker is carried out by translating BT into Communicating Sequential Processes (CSP) [24]. Failure Mode and Effect Analysis (FMEA) can be performed [22], which ascertain the types of hazards that can occur during component failure. Similarly timed FMEA [23] can be carried out using UPPAAL model checker. Translation into PRISM model checker gives support for carrying out probabilistic FMEA [25]. Behavior Trees have been translated into State machines with the intent to extend the software analysis BT already provides and provide support for automatic generation of test cases [14]. Slicing strategy has been proposed [26] to circumvent the state-explosion problem in BT, however the slicing criteria is dependent on property being checked. Our simulator SimTree is different from the previous approaches in that it adds full fledged simulation flavor to the Behavior Trees analysis. So far model checking is dominant in BT with a number of model checkers available that offer different facets of state space exploration (see Fig. 3). Simulators which are already available either provide theoretical framework [11] or are geared towards carrying out automatic test case generation [14]. Table 1 shows the comparison of previous

approaches on the basis of formal analysis spectrum and shows the problems associated with each. Our simulator enriches the analysis available for BT by adding simulation capabilities. Simulation helps carry out formal analysis inexpensively (in terms of time and resources) as compared to model-checking that demands not only more computational resources but also requires significant mathematical background for it to be of any practical use.

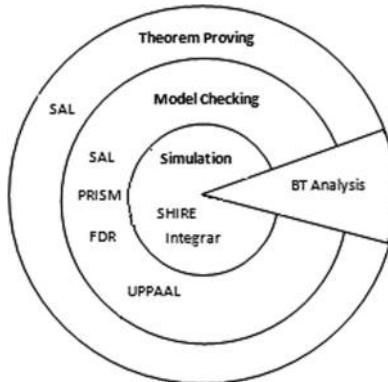


Fig. 3. BT formal analysis spectrum

Table 1. Comparison of BT formal spectrum

Tools	Simulation	Model checking	Theorem proving	Problems
SAL	—	Y	Y	State explosion
PRISM	—	Y	—	State explosion
FDR	—	Y	—	State explosion
UPPAAL	—	Y	—	State explosion
SHIRE	Y	—	—	Test case generation
INTEGRARE	P	—	—	Theoretical framework

3 Transformations

The simulation of BT is achieved by automatically transforming the core components of a given Behavior Tree into the corresponding components of SimTree. For this purpose ATLAS Transformation Language (ATL) is used [27]. The predefined TextBT model [12] is used as a source model in ATL. We define rules to match and navigate the source model elements and then initialize various elements of the simulator defined in the target SimTree model. The simulator code is generated through Model-to-Model (M2M) and Model-to-Text (M2T) transformations [28]. The target model is then used to produce java code for the simulator. The simulator can also be configured to produce Datalog code. Figure 4 outlines the transformation process.

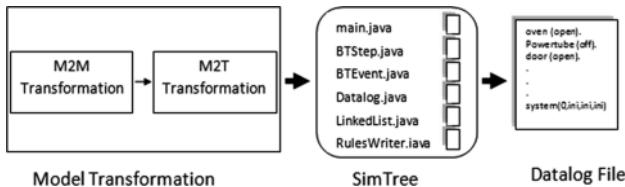


Fig. 4. Transformation process

As a first step, translation rules written in ATL are defined to transform TextBT source model into simulator SimTree target model. The input to the ATL is the TextBT model of a given tree and the result of the execution of ATL is the SimTree model. The translation consists of the following rules; (1) entry point called rule, (2) one matched rule, (3) five lazy rules and (4) four called rules. The entry point rule is so called because it is executed between the initialization and the matching phases in the translation. Matched rules are applied to match a given set of elements using a standard rule. Lazy rules are typically referred from other rules and may be applied many times for each match. In addition, ATL allows for those rules that enable to explicitly generate target model elements directly from code. These rules are referred to as called rules. Our transformation consists of one matched rule, five lazy rules along with few called rules. The matched rules and called rules are used for generating the model elements of SimTree metamodel. Our model also makes use of number of helpers and attribute-helpers to assign values to different properties of SimTree metamodel elements. In the second step, code generation is carried out using M2T transformation. We use Java Emitter Templates (JET) [29] for carrying out the code generation. JET uses templates, XPath expressions and model handlers to navigate the input model and generates the desired code. In our M2T transformation the output is the java code which we save in main.java file. The automatically generated main.java file then becomes the part of the simulator SimTree in the third step. In the third step, simulator is executed to step through the TextBT model for a predetermined number of runs. The simulator can be configured to generate Datalog code for the number of runs defined by the user. The Datalog code consists of rules and facts. These facts and rules are generated on the fly as the simulator steps through each BT node. The code is saved in a text file that can be used in any of the Datalog tool environment for further analysis of the specification.

The BT nodes that are of *state* type are asserted as facts in the code. The node's *component* name becomes the predicate symbol while its *behavior* becomes the term. These facts are asserted only once and become the header in the file. For the nodes that are of *event* type, a fact is asserted with the predicate symbol "events". The arity of this fact is equal to the total number of BT nodes which are of *event* type. For instance, if there are two BT nodes having behavior as *event* then the arity of "events" fact is two. The terms of the "events" predicate symbol are the Boolean values. These values represent the BT nodes having *event* type behavior captured during that instance of the execution. The order of the terms differentiates the different *event* type BT nodes. On the other hand, a Datalog rule is asserted with the predicate symbol as "system". The arity of "system" predicate is two more than the total number of state type BT nodes.

The position of the terms distinguishes one BT node *component* with the other. The first term is reserved for the step number while the last term is reserved for any event occurring during the execution. The head of the rule is the “system” predicate capturing the current executing BT node while the body is the conjunction (represented by comma) of the previously executed BT node as “system” literal along with other literals. The other literals depend upon what type of BT node is being executed. If the executing BT node has behavior of type *event* then the literal is “events” otherwise it is the literal capturing the currently executing BT node. The literal’s predicate symbol is the *component* of the node while the term is the *behavior* of the node. Once the execution of the SimTree is complete and the generated Datalog rules and facts are stored in the file, we can carry out analysis. The Datalog file can be used via Datalog queries for this purpose. The polynomial termination time of the queries guarantees an efficient analysis. The detailed transformation can be found in [30]. In the next section we describe the working of the simulator SimTree.

4 The SimTree Tool

SimTree is based on existing integrated development called the Eclipse Modeling Framework (EMF) [31]. The abstract level workflow diagram of the tool environment is illustrated in Fig. 5. The tool takes BT in XMI format as input. The validated file is loaded and displayed on the screen (Fig. 6). The user can initiate M2M transformation using ATL. The JET module enables generation of the corresponding java code. The BT execution module is used to execute the java code. The graphical interface of the tool allows us to configure the number of runs the simulator should execute. Another interesting aspect of the tool is that some basic safety properties can be graphically configured (Fig. 7). The tool generates a warning and a trace path if any violation of the safety property occurs. The safety property takes the form: If component A is in state x then component B should not be in state y. Since the simulator runs are not always exhaustive, therefore, absence of violation of the property does not conclusively rule out a failure. However, the violation and the corresponding trace in a limited number of runs allows for easy and timely feedback for the analyst. The user can graphically select the components and their states from the list along with the events of interest.

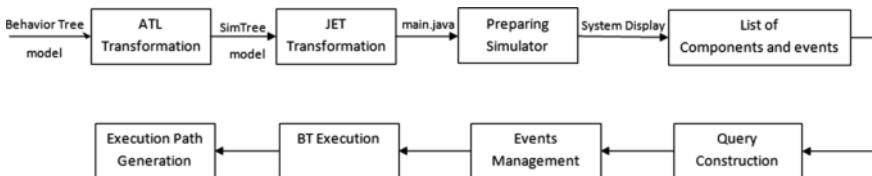


Fig. 5. Block diagram of SimTree simulator

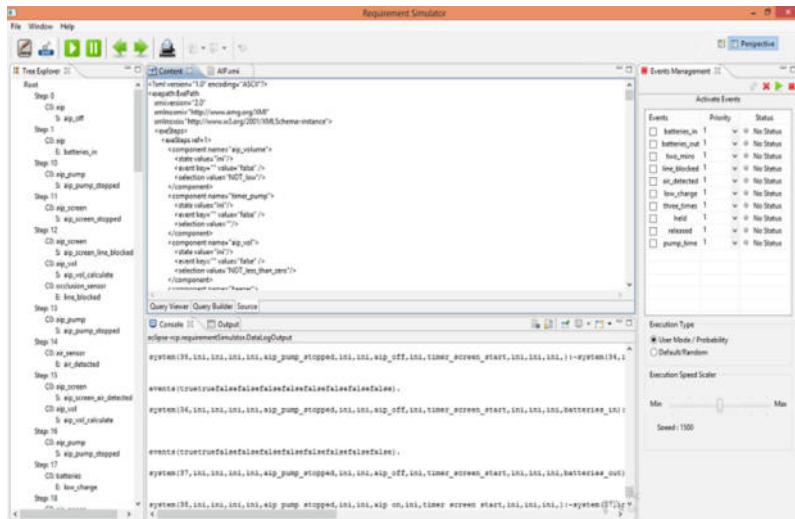


Fig. 6. Datalog rules execution and file generation

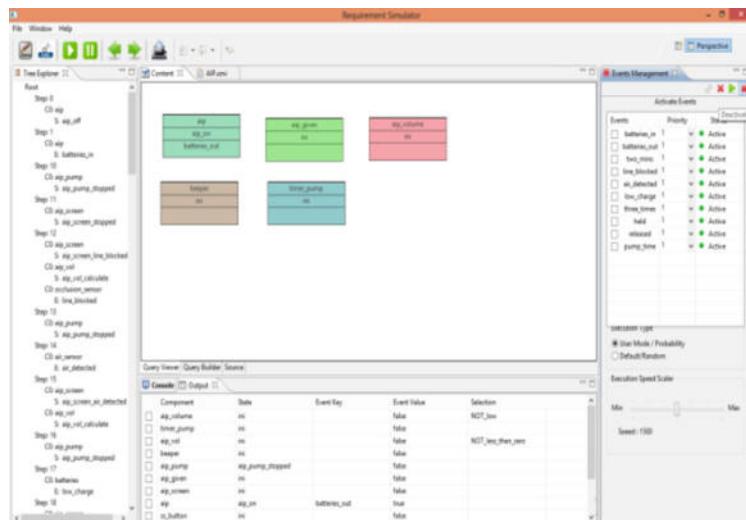


Fig. 7. Graphical queries construction and analysis

The Event Manager module (Fig. 8) in the tool can be used to assign varying priorities to the event to make the execution of the simulator more meaningful. The user has the option of configuring events according to pre-determined probabilities. This serves two important purposes. One is that the low probability events do not get priority over high priority events and vice versa. Otherwise, the simulator results may not only be biased but uninteresting traces will also be generated. The other advantage is that full coverage of the tree is ensured in reasonable amount of time by covering all

the leaf nodes with realistic values for all the events of the tree. Additionally, some events can be executed or omitted through brute force to reach or avoid interesting parts of the tree during analysis. Snap shots of the SimTree tool environment are shown in Figs. 6 and 7. At each step during execution, the simulator records the system states as Datalog facts and rules in a separate text file. The Datalog file can be opened in any of the available environment. We use a free and open source environment called Datalog Educational System (DES) [32] for querying the database for the properties of interest. The database of facts and rules are useful for generating traces, validating liveness properties and safety properties in a given set of runs. The exhaustive search of system states is not possible because all the possible system states cannot be guaranteed in a user defined execution of the simulator. While this is a major limitation, we find the ability to identify the defects that would have passed through to the next stage or caught only through more expensive and exhaustive model-checking to be a useful feature of the Datalog code.

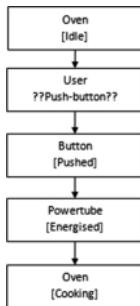


Fig. 8. Requirements translation

5 Requirements Execution Using SimTree

We now introduce the microwave oven case study that we have used to illustrate the usefulness of SimTree. The requirements of microwave oven have been adapted from [9]. The requirements have been simplified for the convenience of time and space. Microwave is a simple system that cooks for one minute on push of a button. The system can cook for subsequent minutes if the user keeps on pushing the button after a lapse of one minute. The cooking function of the one-minute microwave can be interrupted if the door of the oven is opened. These simplified requirements are listed in Table 2. The system must also exhibit some safety properties. The first safety property requires that the power tube of the microwave should not be energized when the door is open (1). While the second property requires that the microwave stops cooking as soon as the door is opened (2). These properties are stated using Linear Temporal Logic (LTL) [33] in the following equations, where the global operator G stands for *it should always hold*.

Table 2. Requirements of Microwave Oven

Requirements description	
R1	If the oven idle with the door closed and the user pushes the button the oven will start cooking
R2	If the button is pushed while the oven is cooking it will cause the oven to cook for an extra-minute
R3	Opening the door stops the cooking
R4	If the oven times-out the power-tube is turned off and a beeper emits a sound to indicate that cooking has finished
R5	If the oven is idle and the user opens the door the power-tube is turned off

$$G(Door = Open \Rightarrow X(\neg(PowerTube = energized))) \quad (1)$$

$$G(Door = Open \Rightarrow X(\neg(Oven = Cooking))) \quad (2)$$

5.1 Requirements Translation

The requirements of microwave oven were systematically translated into BT specification using the Behavior Engineering approach. The requirements listed in Table 2 were translated one by one into individual Requirements Behavior Trees (RBTs). Translation defects are addressed before composing together all the RBTs into a single Integrated Behavior Tree (IBT). Again, all the integration defects were removed during this step. Figure 9 illustrates translation of the first requirement R1 with its defects removed. Figure 10 shows the final microwave BT specification as a single IBT. The complete IBT is not meant to be readable but it is presented here only to give the reader a general idea of how the tree may look like. The BT model was developed using the TextBE tool. After specifying the BT requirements as a complete IBT, M2M and M2T transformations are applied on it. The result of which produces the code of the simulator SimTree. SimTree is configured for the desired number of runs and the probability values for the events.

5.2 Requirements Simulation

The SimTree after configuration is set up for the execution of microwave oven requirements. The execution allows for stepping through each BT node according to the execution semantics of BT and capturing that information as Datalog rules and facts. The SimTree code consists of threads that represent each step of the Behavior Tree. Each thread consists of conditions, node values, events and program counter values. The conditions need to be fulfilled for that particular thread to be executed. The node and event values contain the information about a particular BT node, while the program counter values are set to control the execution of the threads. Each thread is executed only when the program counters values in the condition hold. Once a step is executed, its corresponding program counter is incremented to show the control flow and the node and event values are set. The concurrency is handled by setting the



Fig. 9. Behavior tree of microwave oven

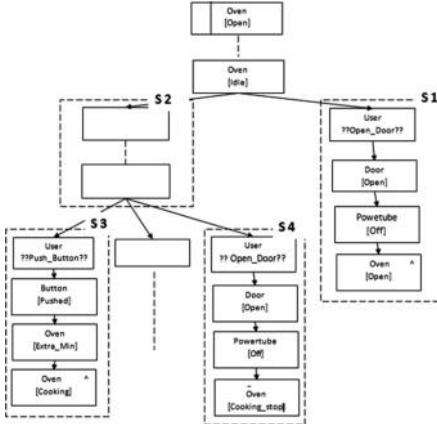


Fig. 10. Defects in microwave BT

program counter values of the children BT nodes (that are composed together in parallel) to one, of the corresponding parent node thread in the SimTree code. The simulator non-deterministically executes any of the child-nodes that are composed together in parallel. The interleaving of other threads is possible in between those threads that represent the parallel composition. The critical regions are preserved by precluding the interleaving of other SimTree threads. This is possible because atomically composed nodes are captured as a single thread. It is executed as a single step and hence acts as a single BT node that contains multiple node and/or event values.

Similarly the alternate composition of BT nodes is captured by setting the program counter values of the sibling nodes to zero, thereby terminating their execution. The execution of SimTree threads at the back end saves the state changes of components as Datalog rules and facts as a Datalog file. The Datalog file contains the information about the states of the different components at any given time which can later be used to carry out formal analysis.

We simulated the microwave oven BT in the SimTree simulator. The state space of the microwave captured as Datalog code is later put subject to formal analysis. We did 10 simulation runs during the execution of BT. One run in simulator is counted when all the leaf nodes in a microwave BT have been visited. It may be noted that some leaf nodes may be visited multiple times due to nondeterministically chosen paths during the execution. The simulator generated Datalog text file during its execution in a step by step manner. We use the Datalog code to analyze the simulation results and to verify the safety properties in the generated state space.

5.3 Requirements Analysis

We use the DES environment to perform the analysis. The captured state space of microwave oven as Datalog rules and facts are loaded into the DES environment. The loaded rules and facts are then checked for violation of any safety property. Datalog provides constraints to check for violations of such properties over the loaded state space. We present here the result of checking the two safety properties of the system. The first property (1) states that it should always hold that whenever the door is open the power tube is not energized. To query for the property to see if it holds or not we make use of the Datalog *constraint*. The Datalog *constraint* (3) states that when B is True, A should not be true. To check the first property we translate it as illustrated in (4) below. It states that no such tuple should be computed where door is opened and the power tube is energized. Similarly, the second safety property when specified using Datalog constraint (5) can be read as no such tuple should be computed when the door is open and oven is cooking.

$$:- A, B \quad (3)$$

$$:- \text{system}(A, B, C, \text{door_open}, \text{Powertube}, F), \text{ Powertube} = \text{powertube_energized} \quad (4)$$

$$:- \text{system}(A, B, \text{Oven}, \text{door_open}, E, F), \text{Oven} = \text{oven_cooking} \quad (5)$$

We used the DES environment to execute the queries (4 and 5) to check for the desired properties in the microwave system. On execution, results of both the queries failed as undesirable tuples were found in both the cases. On careful examination of the generated traces we identified root cause of the failure. As illustrated in Fig. 11 the parallel execution of the tree caused the undesirable interleaving of the BT nodes in branches (S1, S2, S3, S4). In S1 branch the component Door is in open state followed by Powertube being in off state. However, in branch S2, the Powertube is being energized in parallel. Due to the interleaving in these parallel branches the system fails

to hold first safety property. Similarly in S4 branch the component Oven is in Cooking_stop state when the component Door is in Open state. However due to the parallel branch S3 the Oven is Cooking, hence the violation of Datalog constraint (5). For rectification of the incorrect modeling of microwave oven BT we make the composition of nodes in S1 and S4 branches as atomic. The atomic composition of nodes: (i) Door [Open], (ii) Powertube [Off] in S1 ensures that Powertube is turned off as soon as microwave Door is opened. Similarly the three nodes: (i) Door [Open], (ii) Powertube [Off], (iii) Oven [Cooking_stop] are made atomic in branch S4. This guarantees that as soon as door is opened the powertube is turned off and the cooking is stopped immediately by preventing the interleaving of nodes in S2 and S4 branches.



Fig. 11. Defects in microwave BT

6 Conclusion

Simulation of requirements is recognized as an intuitive way of developing an insight into the system being modeled. Simulators can play wide-range of roles from checking for well-formedness of requirements to more complex analysis like checking for safety properties, liveness and deadlocks in system specifications. More importantly they can also play an important role in early defect detection, requirements elicitation and requirements validation. In this paper we introduce the first version of requirements simulator SimTree. Main contribution of our work is the automatic translation of BT into SimTree. The simulator executes the requirements modeled as BT to carry out formal analysis. The current version of the simulator is configured to produce Datalog code as it steps through each BT node in a given run. The usefulness of the simulator is illustrated using a simple one-minute microwave case study. The simulator was useful for detecting a logical error in the specification during the checking of safety properties. The limitation of the tool include that the Datalog code generation is not optimized

making the analysis slow and time consuming. We plan to rectify the problem using the time slicing strategy used to optimize the automated translation of BT into SAL specification [26]. Furthermore, we plan to incorporate exhaustive simulation runs for verification of different properties. We also aim to improve the graphical specification of safety properties of a more complex nature.

References

1. Mauco, M.V., Leonardi, M.C., Riesco, D.: Deriving formal specifications from natural language requirements (2009)
2. Jensen, K., Kristensen, L.M., Wells, L.: Coloured Petri nets and CPN tools for modelling and validation of concurrent systems. *Int. J. Softw. Tools Technol. Transf.* **9**(3–4), 213–254 (2007)
3. Wagner, F., Wolstenholme, P.: Modeling and building reliable, re-useable software. In: *Engineering of Computer-Based Systems*, 2003. Proceedings. 10th IEEE International Conference and Workshop on the, pp. 277–286. Alabama, USA (2003)
4. Dromey, R.G.: From requirements to design: Formalizing the key steps. In: *Software Engineering and Formal Methods*, 2003. Proceedings. First International Conference, pp. 2–11. Brisbane, Australia (2003)
5. Lamsweerde, A.: Formal specification: a roadmap. In: *Proceedings of the Conference on the Future of Software Engineering*, pp. 147–159. Limerick, Ireland (2000)
6. Hasan, O., Tahar, S., Abbasi, N.: Formal reliability analysis using theorem proving. *Comput. IEEE Trans.* **59**(5), 579–592 (2010)
7. Jackson, D.: Lightweight formal methods. In: *FME 2001: Formal Methods for Increasing Software Productivity*, pp. 1–1. Springer, Berlin, Germany (2001)
8. Schmid, R., Rysen, J., Berner, S., Glinz, M., Reutemann, R., Fahr, E.: A survey of simulation tools for requirement engineering. Universität Zürich, Institut für Informatik (2000)
9. Dromey, R.G.: Formalizing the transition from requirements to design. *Math. Framew. Compon. Softw. Models Anal. Synth.* 173–205 (2006)
10. Powell, D.: Requirements evaluation using behavior trees—findings from industry. In: *Australian Software Engineering Conference (ASWEC'07)*, Melbourne, Australia (2007)
11. Wen, L., Colvin, R., Lin, K., Seagrott, J., Yatapanage, N., Dromey, G.: ‘Integrale’, a collaborative environment for behavior-oriented design. In: *Cooperative Design, Visualization, and Engineering*, pp. 122–131. Springer (2007)
12. Myers, T.: TextBE: a textual editor for behavior engineering. In: *Proceedings of the 3rd Improving Systems and Software Engineering Conference (ISSEC)*, Sydney, Australia (2011)
13. Grunske, L., Winter, K., Yatapanage, N.: Defining the abstract syntax of visual languages with advanced graph grammars—a case study based on behavior trees. *J. Vis. Lang. Comput.* **19**(3), 343–379 (2008)
14. Kim, S.-K., Myers, T., Wendland, M.-F., Lindsay, P.A.: Execution of natural language requirements using state machines synthesised from Behavior Trees. *J. Syst. Softw.* **85**(11), 2652–2664 (2012)
15. Christie, A.M.: Simulation: an enabling technology in software engineering. *CROSSTALK —J. Def. Softw. Eng.* **12**(4), 25–30 (1999)
16. Rushby, J.M.: Model Checking and Other Ways of Automating Formal Methods Position Paper Panel Model Checking Concurrent Programs. *Software Quality Week*, San Francisco (1995)

17. Sáenz-Pérez, F.: Outer joins in a deductive database system. *Electron. Notes Theor. Comput. Sci.* **282**, 73–88 (2012)
18. Besson, F., Jensen, T.: Modular class analysis with datalog. In: *Static Analysis*, pp. 1075–1075. San Diego, California (2003)
19. Hoffmann, V., Licher, H.: A model-based narrative use case simulation environment. In: *ICSOFT*, pp. 63–72. Athens, Greece (2010)
20. Dromey, R.G.: Architecture as an emergent property of requirements integration. In: *STRAW'03 Second International Software Requirements to Architectures Workshop*, p. 77. Oregon, USA (2003)
21. Zafar, S., Dromey, R.G.: Managing Complexity in Modelling Embedded Systems. In: *Systems Engineering/Test and Evaluation Conference SETE2005*. Brisbane, Australia (2005)
22. Grunske, L., Lindsay, P., Yatapanage, N., Winter, K.: An automated failure mode and effect analysis based on high-level design specification with behavior trees. In: *Integrated Formal Methods*, pp. 129–149. The Netherlands, Eindhoven (2005)
23. Colvin, R., Grunske, L., Winter, K.: Timed behavior trees for failure mode and effects analysis of time-critical systems. *J. Syst. Softw.* **81**(12), 2163–2182 (2008)
24. Winter, K.: Formalising behaviour trees with CSP. In: *Integrated Formal Methods*, pp. 148–167 (2004)
25. Grunske, L., Colvin, R., Winter, K.: Probabilistic model-checking support for FMEA. In: *Quantitative Evaluation of Systems, 2007. QEST 2007. Fourth International Conference*, pp. 119–128. Edinburgh, Scotland (2007)
26. Yatapanage, N., Winter, K., Zafar, S.: Slicing behavior tree models for verification. In: *Theoretical Computer Science*, pp. 125–139. Springer (2010)
27. Allilaire, F., Bézivin, J., Jouault, F., Kurtev, I.: ATL-eclipse support for model transformation. In: *Proceedings of the Eclipse Technology eXchange workshop (eTX) at the ECOOP 2006 Conference*, vol. 66. Nantes, France (2006)
28. Mens, T., Van Gorp, P.: A taxonomy of model transformation. *Electron. Notes Theor. Comput. Sci.* **152**, 125–142 (2006)
29. Templates, J.E.: Part of the Eclipse Modeling Framework, see JET Tutorial by Remko Pompa at <http://eclipse.org/articles>, Article-JET2/jetTutorial2.html, vol. 69
30. Zafar, S., Farooq-Khan, N., Ahmed, M.: Requirements simulation for early validation using Behavior Trees and Datalog. *Inf. Softw. Technol.* **61**, 52–70 (2015)
31. Steinberg, D., Budinsky, F., Merks, E., Paternostro, M.: *EMF: Eclipse Modeling Framework*. Pearson Education (2008)
32. Sáenz-Pérez, F.: DES: a deductive database system. *Electron. Notes Theor. Comput. Sci.* **271**, 63–78 (2011)
33. Emerson, E.A.: Temporal and modal logic. *Handb. Theor. Comput. Sci.* **2**, 995–1072 (1990)



SuperDense Coding Step by Step

Lewis Westfall and Avery Leider^(✉)

Pace University, Pleasantville, NY 10570, USA
[{lw19277w,aleider}](mailto:{lw19277w,aleider}@pace.edu)@pace.edu

Abstract. Scholars of quantum computing all become familiar with Alice and Bob when learning about superdense coding and entanglement. However, in every research book and video that we found, the assumption is made that the student will automatically understand how those two classical bits at the end come to their values when they started as two qubits. This vagueness was unavoidable when quantum computers were purely theoretical. After exhaustive search of every quantum superdense coding Bob and Alice example in the research literature since late 2017, we found not one that presented evidence from a real quantum computer. However, moving from theory to practice is necessary. Today, using results from a real IBM Q Experience quantum computer, we illustrate each step of the Bob and Alice qubit journey and make it all crystal clear.

Keywords: SuperDense coding · Quantum computing · Classical bits · Qubits · Entanglement · IBM · Q experience · Bob and Alice

1 Introduction

Superdense coding is the ability of entangled qubits to carry more information than a classical bit allows [1]. This is possible because of quantum entanglement, which conveys a quantum state from one qubit to another [2]. The use of superdense coding by entanglement is a quantum cryptography method of great promise to defeat any eavesdropper [3]. Frequently the example is of Alice sending two classical bits of information in one qubit to Bob. In our case, Alice wants to tell Bob that the weather is clear or cloudy and cold or warm. The process starts when a third party, Eve, starting with two separate qubits, entangles the pair by applying a Hadamard (H) gate to the first qubit and then a Controlled Not (CNOT) gate to both qubits, where the first qubit is the control and the second qubit is the target. This entanglement is a Bell state [4], $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Once the CNOT gate is applied, entanglement occurs between the two qubits. It is important in this exercises is to remember that quantum gates are reversible, unlike classical gates. The CNOT gate can go both ways - it can entangle, and it can disentangle. One qubit is given to Alice while the other is given to Bob.

Thanks to the IBM Faculty Award that made this research possible.

© Springer Nature Switzerland AG 2020

K. Arai and R. Bhatia (Eds.): FICC 2019, LNNS 70, pp. 357–372, 2020.

https://doi.org/10.1007/978-3-030-12385-7_28

The qubits of Alice and Bob will remain entangled although one of the qubits is subjected to more gates than the other qubit.

Table 1 shows what message options Alice has to send to Bob, the binary code for that message, and the quantum gates she will use to encode this message information to her qubit. Because of the entanglement, when Alice does her quantum gate operations on her qubit, she can not use a standalone gate because she has an entangled qubit. The gate she will use will be the tensor product of the desired gate with the Identity matrix to create a 4×4 matrix. Although the qubits are entangled, while the qubits are in the possession of their original owner, only Alice can perform gate operations on her qubit, and only Bob can perform gate operations on his qubit. Once Alice's qubit has been encoded by operations of the quantum gates, it is transmitted to Bob.

Table 1. Alice's conditions and actions

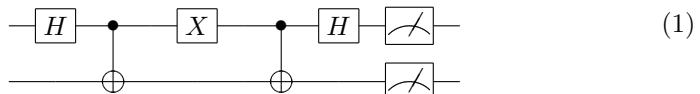
Temp	Sky	Code	Quantum gate
Cold	Clear	00	I
Cold	Cloudy	01	Z
Warm	Clear	10	X
Warm	Cloudy	11	ZX

Bob, with Alice's qubit now in his possession, as well as his own qubit, can perform quantum gate operations on Alice's qubit to discover the message in it.

Bob first applies a Controlled-Not gate operation to both entangled qubits, where Alice's qubit is the control and Bob's qubit is the target. Applying the CNOT to an entangled pair of qubits causes them to become disentangled and break into two independent qubits. Bob first measures the second qubit, the formerly target qubit that was Bob's qubit. Measurement ends the qubit's activation life, but it reveals data, which is either cold or warm.

Bob then applies an H gate to Alice's original qubit, the first qubit, the former control qubit of the now disentangled pair. The H gate operation extracts the second bit of information in the message, i.e. clear or cloudy. When the qubit is measured the binary code that reveals the message that Alice encoded in her qubit is displayed. The measurements end the process [5].

Circuit 1 shows the complete quantum circuit diagram for Alice using the X gate to encode warm and clear. We see the Hadamard and CNOT gates used by Eve to entangle, the X gate used by Alice to encode the message, and finally the CNOT and Hadamard gates used by Bob to extract the message.



One example of how this is explained is Michael Nielsen's YouTube video [6]. A beginner, or even a more experienced student of quantum computing might

not understand what is going on because the video skips some steps that are assumed to be understood. Figure 1 shows the diagram that is drawn while explaining the process.

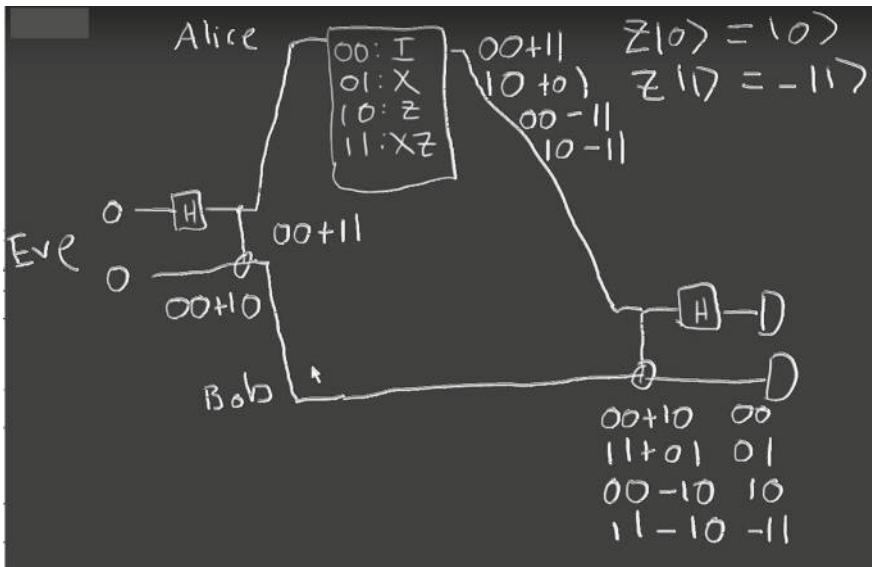


Fig. 1. Michael Nielson describes SuperDense coding [6]

Our goal is to explain each step in careful detail so nothing is assumed other than linear algebra and basic quantum code notation [7]. We have run the quantum program, of superdense coding step by step, on the IBM Q experience quantum processors. We will show how each step changes the values of the two qubits using histograms from the runs and linear algebra to show the manipulation of the qubits.

Section 2 gives a review of the literature on explaining superdense coding using Bob and Alice.

Section 3 describes the methodology used to create the examples.

Section 4 gives an overview of the entire process, a step by step explanation of how each step affects the two qubits and ends with the histograms and a discussion of results.

Section 5 discusses the conclusion and future work.

2 Literature Review

In the middle of 2017 came a great divide, as in that time frame, IBM made available to the public real quantum computers over the Internet. IBM has continued this service, and the real quantum computers they are offering to the free

service are in a state of continuously increasing qubit capacity. Research published before that time did not have the opportunity to test their equations on a commonly available real quantum computer. Research published after that time could have included results from the nascent real quantum computers. However, few do this. Alice and Bob and their superdense example presents just the right simplicity for analysis on a real quantum computer.

We did an extensive search of the literature on quantum computing since mid-2017 that expressed examples of Bob and Alice. We found not a single paper presented evidence of using a real quantum computer in their studies. The researchers mostly stayed in the familiar theoretical area of mathematical proofs. We found intriguing research on quantum cryptography where Alice locks her bit into a safe, and later gives Bob the key, with the magic of superdense coding entanglement [8], a study of pure theory, no experiments. We also found where Alice and Bob each modify their quantum bit particles with operators to reveal their superposition features and on the path to each other, the bits bouncing, deflecting, off of interference from a unitary operator in their communication path that influences their values - this is exceptionally interesting, yet the researchers did not use a real quantum computer or a quantum simulator to test their ideas [9].

Even when the researchers on a government funded study visited the engineers on the site of the real IBM quantum computers and made note of that in their paper, they still did not conduct a real experiment to test their ideas about Alice and Bob [10]. We did find a real experiment, using mirrors, photon detectors, and other optic equipment, of sending a single photon on a two-way journey from Alice to Bob carrying two classical bits [11]. However, although real, they did not use a real quantum computer. Best in our review was cryptography research that used Alice and Bob, joined by Eve the eavesdropper and Charlie the observer, in quantum key distribution that used quantum optics simulations that show real results that could be practical [12].

We determined at that point that we must try to match quantum theory to its test in real quantum computer output with this simple example of Bob and Alice. We then quickly experienced chaos as we attempted to execute that plan. It became obvious that this new ability to test linear equations of gates on real quantum computers did not have firm sequential steps for newcomers. Real quantum computer programming is raw, new, full of phenomena not well defined. We found in that chaos the second objective of this study: making our findings clear enough that they could be easy for our readers to duplicate and follow in the real environment of quantum computing.

Although some cryptographers dislike using the terms “Alice” and “Bob”, or, in fact, any human name, in quantum computing mathematics, and ask that researchers instead use symbols such as “A” or “B” [13], the 1978 seminal paper establishing RSA famously created Alice and Bob, and they remain enduring tools to understand complex security communications protocols [14] including superdense coding, which is a purely quantum action - there is no classical equivalent. Alice and Bob are arguably as ubiquitous as Schrodinger’s cat [15].

In 2018, Alice and Bob continue to appear in new quantum computer articles in publications from Russia [16] to China [17]. Alice and Bob are universal. We determined to use them, as they are the standard.

3 Project Requirements

To duplicate our study, refer to our development environment setup in Fig. 2. Needed will be the Anaconda environment [18], and a Jupyter Notebook containing QISKit, which can be downloaded with Anaconda from QISKit. Each individual user will need an API key obtained at no cost from IBM's Q Experience [19].

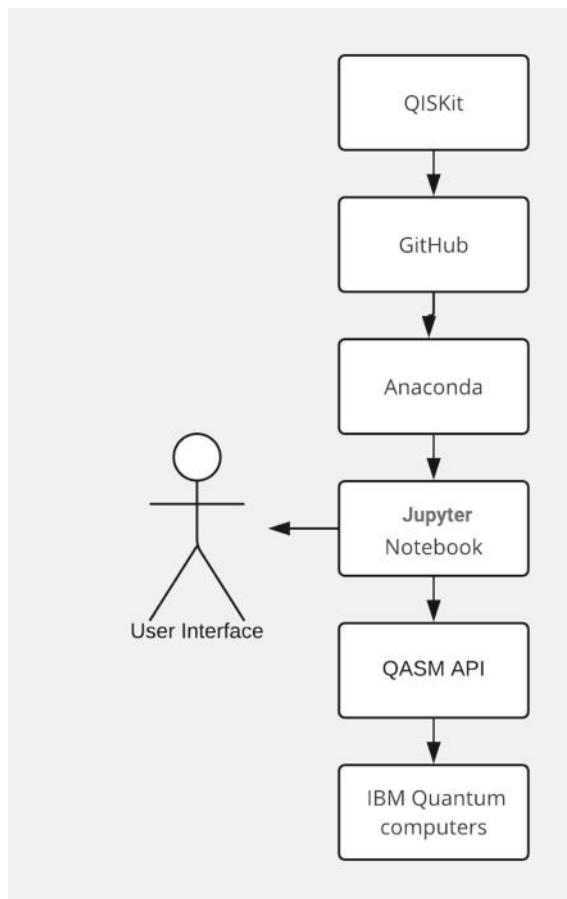


Fig. 2. Setting up the development environment

The following histograms used in this study are all direct output from our Jupyter Notebook connected to the the IBM Q experience quantum computer [19]. The Jupyter Notebook implements the open source software development kit (SDK), QISKit, that includes a Python API [20] that translates the Python into Quantum Assembly Language (QASM) [21], which is then processed by an IBM quantum computer [22].

At the GitHub site [23] for the QISKit-core there is a risk; the default branch “master” may include untested code. On the left side of the QISKit-core GitHub page is a pull-down. Select the “stable” branch. This one step will save a lot of frustration.

4 Methodology

Superdense coding has been presented [1, 6] in a theoretical manner using a quantum diagram and notation of qubit values. We will be using Python programs that are written in Jupyter Notebooks. The program code and the resulting histograms will be discussed. Then we will provide the linear algebra manipulations done by the program.

The steps will be to start with the unmodified qubits and add the steps one at a time until the complete process for one of the classical bit information communication messages has been described. Finally we will discuss the other 3 classical bit information sets that can be transmitted.

The five quantum gates we will use are the identity (I), Pauli-X (X), Pauli-Y (Y), Pauli-Z (Z), and the Hadamard (H).

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}$$

$$Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

$$H = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

We use one multiple qubit gate, the controlled not or CNOT gate. This gate has a control qubit and a target qubit. If the control qubit is 0 then the target qubit is not affected but if the control qubit is 1 then the value of the target qubit is flipped, $0 \rightarrow 1$ and $1 \rightarrow 0$. The matrix operation for the CNOT is

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

and the quantum diagram for the CNOT is



The ket values $|0\rangle$ and $|1\rangle$ correspond to the column vectors $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ respectively.

5 Results

One thing to be very acutely aware of is that, by the conventions we are using, when we write the value of two entangled qubits, a and b, in ket form $|ab\rangle$ that a is the top qubit, $q[0]$, and b is the bottom qubit, $q[1]$, but in the histogram produced by IBM Q Experience they read the other direction. For example in the histogram Fig. 3 the ket values from left to right are $|00\rangle$, $|10\rangle$, $|01\rangle$, and $|11\rangle$. This naming convention mix-up is because it has not yet been settled in quantum computing, if we are reading left to right, or right to left.

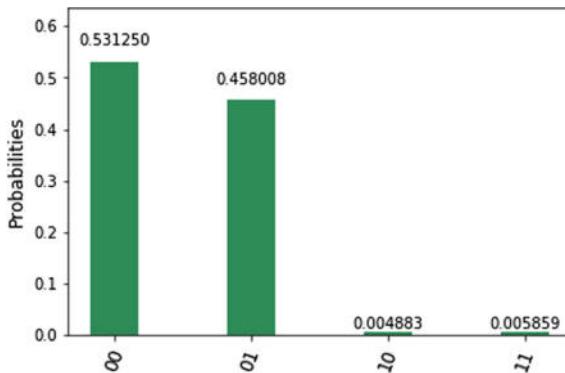


Fig. 3. Histogram after applying the H gate to $q[0]$

However if you look at the histogram x-axis values from top to bottom, they match the order of the qubits in the quantum diagram. Then IBM's histogram convention makes more sense.

The quantum circuit starts with 2 qubits in the base state $|0\rangle$ or in matrix format $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

The quantum diagram is rather unexciting but it is the starting point.

$$\begin{array}{c} |0\rangle \xrightarrow{\quad} \\ |0\rangle \xrightarrow{\quad} \end{array} \quad (2)$$

Note that the upper qubit, $q[0]$ in Eq. 2 is the first or low order qubit and the lower qubit, $q[1]$ is the second or high order qubit. This naming convention will help keep things clear in later steps of the process.

The first operation Eve performs is a Hadamard (H) gate on qubit 0.

$$\begin{array}{c} |0\rangle \xrightarrow{\boxed{H}} \\ |0\rangle \xrightarrow{\quad} \end{array} \quad (3)$$

Applying the H gate to qubit 0 results in it being put into superposition.

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

See Fig. 3.

The second step that EVE performs is a CNOT gate

$$\begin{array}{c} |0\rangle \xrightarrow{\boxed{H}} \bullet \\ |0\rangle \xrightarrow{\quad} \oplus \end{array} \quad (4)$$

Since the CNOT is a multiple qubit operation the two qubits must be combined by a tensor product

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Then applying the CNOT gate gives

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad (5)$$

The two qubits are now entangled and have the ket value of $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, which is a Bell state. See Fig. 4. We will be using this vector as the starting point for Alice to apply her gates. Here is the code in Python we used to create and entangle the two qubits:

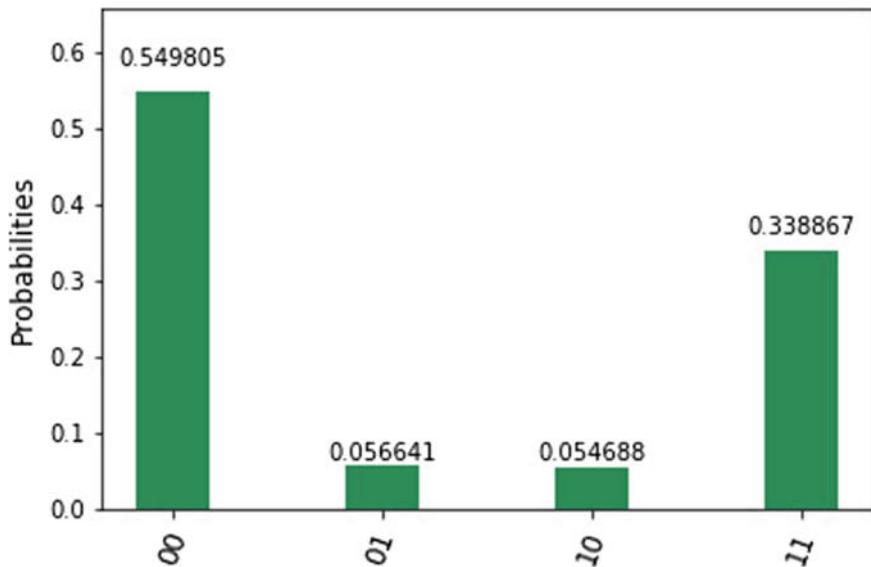


Fig. 4. Histogram after the qubits are entangled

```

q = QuantumRegister(2)
        # Define 2 bit quantum register
c = ClassicalRegister(2)
        # Define 2 bit classical register
qc = QuantumCircuit(q, c)
        # Define quantum circuit
qc.h(q[0])
        # Hadamard on qubit 0
qc.cx(q[0], q[1])
        # CNOT qubit 0 control-qubit 1 target
qc.measure(q, c)
        # Measure qubits
    
```

Listing 1.1. Entangling with the H and CX gate

Finally Eve sends qubit 0 to Alice and qubit 1 to Bob.

Alice determines the information she wants to send to Bob. We will use Cold and Cloudy, bit pattern 01, see Table 1. She applies the Z gate to her qubit to encode the information. The quantum diagram now looks like



Since the qubits are entangled at this point, we can not apply a simple Z gate but must tensor product the Z gate with the I gate producing the following

4×4 matrix.

$$\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

The matrix manipulation for applying the Z gate looks like this.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} \quad (7)$$

Here is the code in Python we used to apply the Z gate:

```

q = QuantumRegister(2)
    # Define 2 bit quantum register
c = ClassicalRegister(2)
    # Define 2 bit classical register
qc = QuantumCircuit(q, c)
    # Define quantum circuit
qc.h(q[0])
    # Hadamard on qubit 0
qc.cx(q[0], q[1])
    # CNOT qubit 0 control-qubit 1 target
qc.z(q[0])
    # z on qubit 0
qc.measure(q, c)
    # Measure qubits

```

Listing 1.2. Adding the Z gate

If you compare the histograms before the Z gate was applied and after, Figs. 4 and 5, you will see that they look very similar. However if you compare Eqs. 5 and 7 for the same states you will see they are quite different. Equation 5 has the vector $1, 0, 0, 1$ while Eq. 7 has the vector $1, 0, 0, -1$. It appears that when the quantum state is measured the minus sign is lost. The thing to note is that the histogram results can be misleading about the actual quantum state.

Now Alice sends her qubit to Bob and he performs a CNOT. The following is the matrix manipulation to perform the CNOT.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix}$$

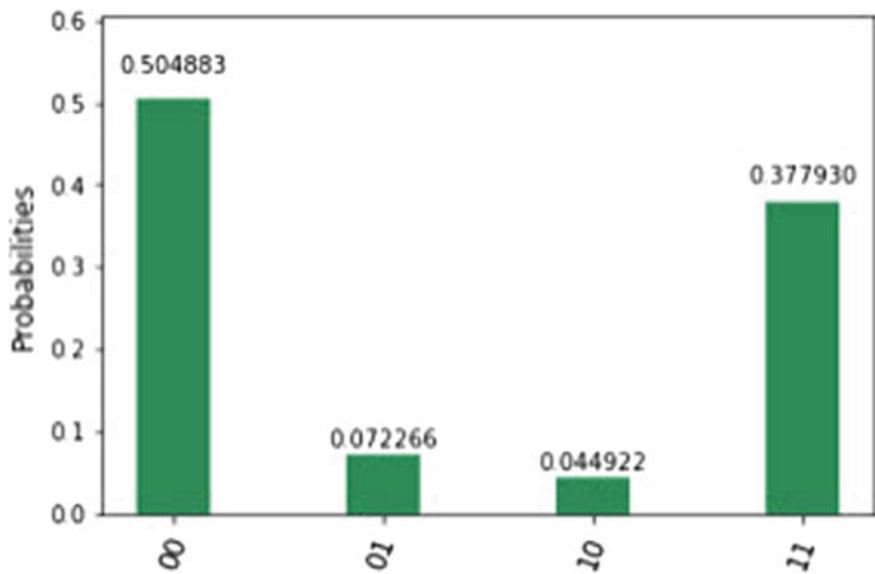


Fig. 5. Histogram after Alice applies Z gate

Here is the code in Python we used to apply Bob's CX (CNOT) gate:

```

q = QuantumRegister(2)
    # Define 2 bit quantum register
c = ClassicalRegister(2)
    # Define 2 bit classical register
qc = QuantumCircuit(q, c)
    # Define quantum circuit
qc.h(q[0])
    # Hadamard on qubit 0
qc.cx(q[0], q[1])
    # CNOT qubit 0 control-qubit 1 target
qc.z(q[0])
    # z on qubit 0
qc.x(q[0])
    # x on qubit 0
qc.cx(q[0], q[1])
    # cx on qubit 1
qc.measure(q, c)
    # Measure qubits

```

Listing 1.3. Applying the CX (CNOT) Gate

The CNOT can change two disentangled qubits to become entangled. The reverse is also true. A CNOT applied to an entangled pair of qubits can disentangle them [5]. See Fig. 6.

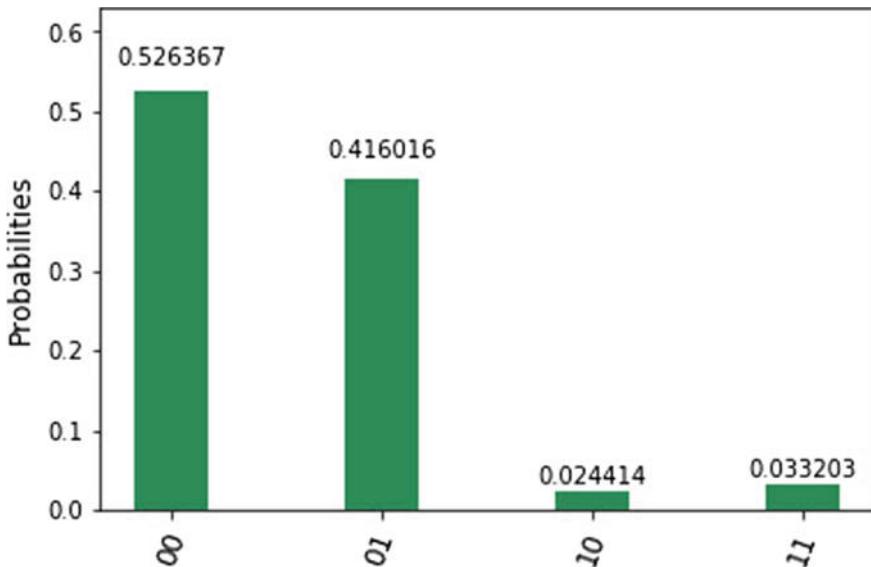


Fig. 6. Histogram after Bob's CNOT

We do not fully understand the matrix manipulation that is performed on the entangled qubits by Bob. The definition of a CNOT gate is that if the controlling qubit is 1 then the target qubit is flipped, otherwise the target qubit is unaffected. The way the process is described in superdense coding is that the CNOT is performed on Alice's qubit while the quantum circuit diagram shows Alice's qubit controlling the CNOT targeting Bob's qubit. In fact the CNOT matrix manipulation is performed on the state that matches Alice's qubit and this gives the expected result. This would make sense if operating on Alice's qubit with the Z gate also changed Bob's qubit in the same way or as Albert Einstein called it 'spooky action at a distance' [24]. If this distant action does not take place, then what exactly are we doing? A further question is how does the CNOT operate if the control qubit is not 0 or 1, but $-1, i, -i$, or some value in-between?

This process extracts the classical bit $c[1]$ from Alice's qubit, which in this case is a 0.

Now Bob applies an H gate to Alice's qubit which extracts the other classical bit, which is a 1. Giving the final result of 10 (reading right to left) as desired or read the other way 01 to match Table 1, see Fig. 7. We actually used the $H \otimes I$ gate to allow for the 4×1 column vector that represents Alice's qubit.

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$$

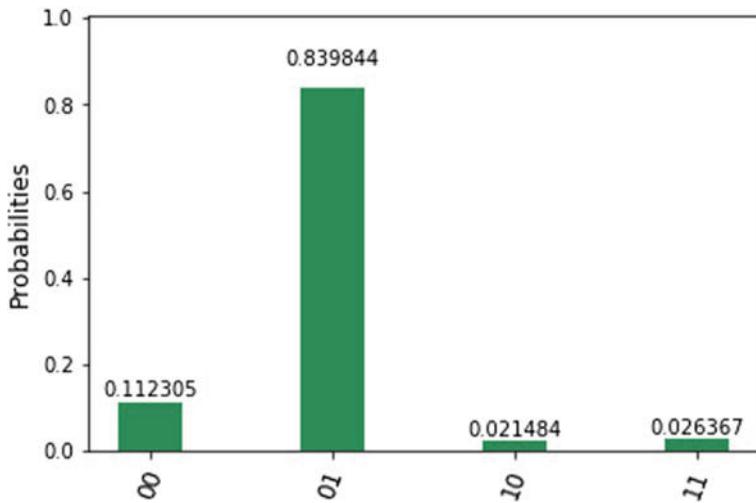


Fig. 7. Histogram Z final result

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = |10\rangle$$

The histograms in Figs. 8, 9 and 10, show the results for the three other operations that Alice can perform to encode her information. Another way to

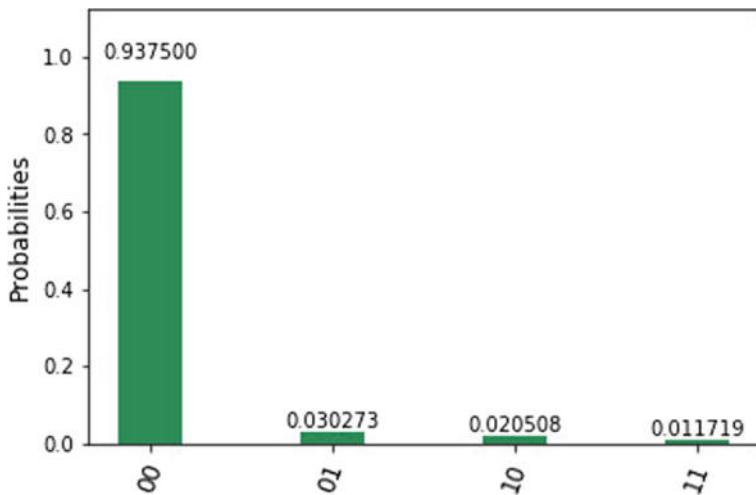


Fig. 8. Histogram I final result

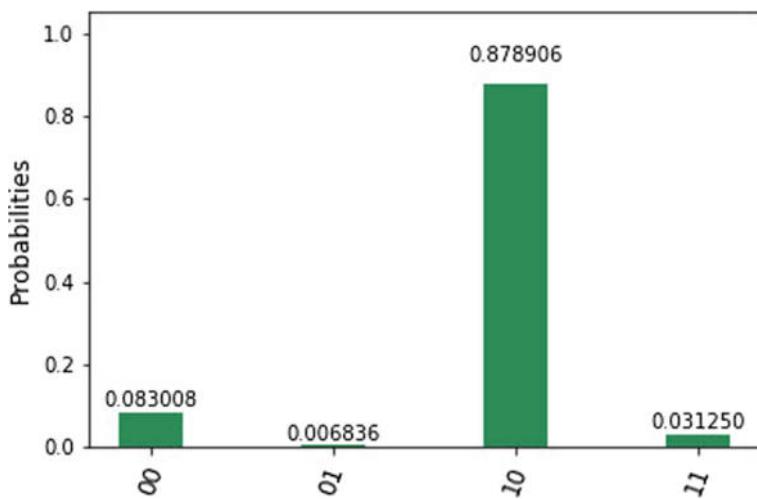


Fig. 9. Histogram X final result

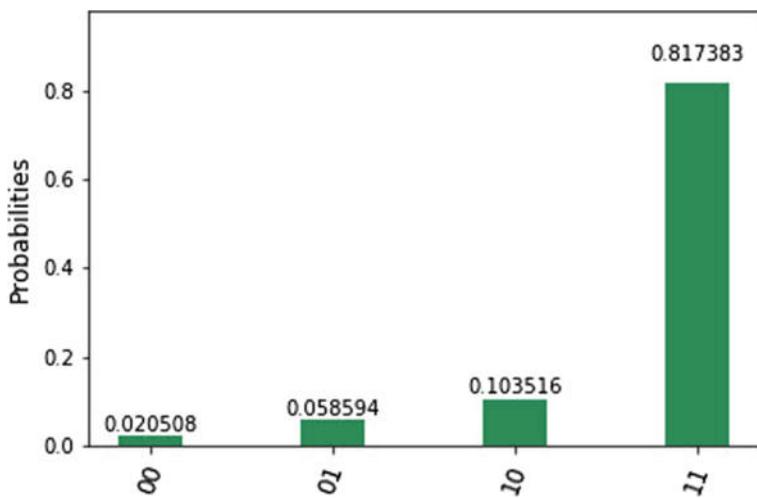


Fig. 10. Histogram ZX final result

encode the 11 condition is to use a iY gate instead of the Z gate followed by the X gate [25].

For the quantum circuit we used only the Y gate giving the histogram displayed in Fig. 11.

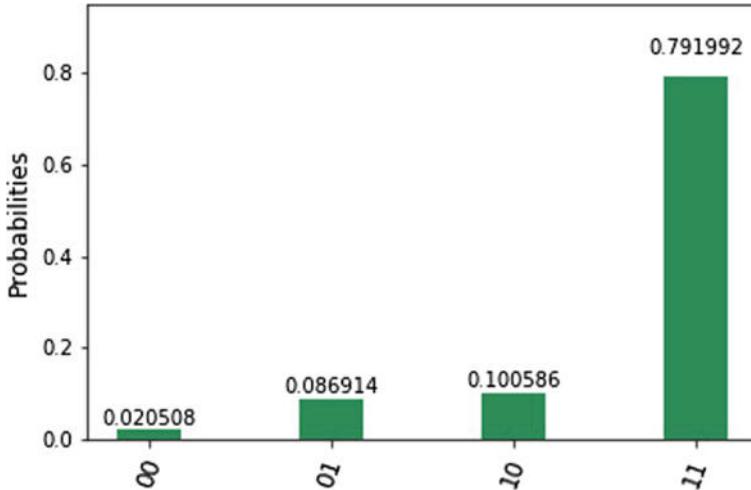


Fig. 11. Histogram Y final result

6 Conclusion and Future Work

We have shown that running superdense coding in the Alice and Bob scenario on a real quantum computer gives the desired results, namely that two classical bits can be carried by one qubit.

We have also shown that measuring a qubit and returning a classical bit does not always give a completely accurate depiction of the state of the qubit.

We have observed that we do not completely understand the matrix manipulation of how applying a CNOT to two entangled qubits works, even though it gives the expected results.

The further investigation of how the CNOT gate is applied to two entangled qubits is one area that needs further investigation along with a more complete analysis of how a manipulation on Alice's qubit affects Bob's qubit.

This research was performed on the $|\Phi^+\rangle$ Bell state, $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$. Future work should be done to see if superdense coding will work with the other three Bell states.

References

1. Nielsen, M.A., Chuang, I.: Quantum computation and quantum information (2002)
2. Zeilinger, A.: Quantum teleportation, onwards and upwards. *Nat. Phys.* **14**(1), 3 (2018)
3. Ekert, A.: Quantum cryptography: the power of independence. *Nat. Phys.* **14**(2), 114 (2018)
4. Kaye, P., Laflamme, R., Mosca, M.: An Introduction to Quantum Computing. Oxford University Press, Oxford (2007)
5. Rieffel, E., Polak, W.: An introduction to quantum computing for non-physicists. *ACM Comput. Surv. (CSUR)* **32**(3), 300–335 (2000)
6. Nielsen, M.: Superdense Coding: How to Send Two Bits Using One Qubit (2010). Available at <https://youtu.be/w5rCn593Dig>
7. Westfall, L., Leider, A.: Teaching quantum computing. In: *Future Technologies Conference* (2018) (in press)
8. Nagy, M., Nagy, N.: An information-theoretic perspective on the quantum bit commitment impossibility theorem. *Entropy* **20**(3), 193 (2018)
9. Del Santo, F., Dakić, B.: Two-way communication with a single quantum particle. *Phys. Rev. Lett.* **120**(6), 060503 (2018)
10. Horodecki, P., Horodecki, M., Horodecki, R.: Zero-knowledge convincing protocol on quantum bit is impossible. *Quantum* **1**, 41 (2017)
11. Massa, F., Moqanaki, A., Del Santo, F., Dakic, B., Walther, P.: Experimental two-way communication with one photon. (2018) arXiv preprint [arXiv:1802.05102](https://arxiv.org/abs/1802.05102)
12. Simon, G.K., Huff, B.K., Meier, W.M., Mailloux, L.O., Harrell, L.E.: Quantification of the impact of photon distinguishability on measurement-device-independent quantum key distribution. *Electronics* **7**(4), 49 (2018)
13. Oppliger, R.: Disillusioning alice and bob. *IEEE Secur. Priv.* **5**, 82–84 (2017)
14. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **21**(2), 120–126 (1978)
15. Schrödinger, E.: The present situation in quantum mechanics. *Naturwissenschaften* **23**, 844–849 (1935)
16. Podoshvedov, S.A.: Quantum teleportation of unknown qubit beyond bell states formalism. (2018) arXiv preprint [arXiv:1801.09452](https://arxiv.org/abs/1801.09452)
17. Wang, K., Yu, X.-T., Cai, X.-F., Zhang, Z.-C.: Probabilistic teleportation of arbitrary two-qubit quantum state via non-symmetric quantum channel. *Entropy* **20**(4), 238 (2018)
18. Anaconda, Inc.: (2018). <https://www.anaconda.com/>
19. IBM. Q Experience.: (2018). <https://quantumexperience.ng.bluemix.net/>
20. Python Software Foundation.: (2018). <https://www.python.org/>
21. Cross, A.W., Bishop, L.S., Smolin, J.A., Gambetta, J.M.: Open quantum assembly language. (2017) arXiv preprint [arXiv:1707.03429](https://arxiv.org/abs/1707.03429)
22. Open Source Quantum Information Science Kit.: (2018). <https://qiskit.org/>
23. QISKit GitHub.: (2018). <https://github.com/QISKit/>
24. Einstein, A., Podolsky, B., Rosen, N.: Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47**(10), 777 (1935)
25. Tappert, C.: Lecture Slides from Quantum Computing Course at Pace University (2018). Available at <http://csis.pace.edu/ctappert/cs837-18spring/index.htm>



Artificial Swarming Shown to Amplify Accuracy of Group Decisions in Subjective Judgment Tasks

Gregg Willcox¹, Louis Rosenberg¹⁽⁾, David Askay², Lynn Metcalf², Erick Harris², and Colin Domnauer³

¹ Unanimous AI, San Francisco, CA 94115, USA

Gregg@Unanimous.ai, louis@unanimousai.com

² California Polytechnic State University, San Luis Obispo, CA 93401, USA
{daskay, lmetcalf}@calpoly.edu

³ University of California Berkeley, Berkeley, CA 94702, USA
Colin@Unanimous.ai

Abstract. New technologies enable distributed human teams to form real-time systems modeled after natural swarms. Often referred to as Artificial Swarm Intelligence (ASI) or simply “human swarming”, these real-time systems have been shown to amplify group intelligence across a wide range of tasks, from handicapping sports to forecasting financial markets. While most prior research has studied human swarms with 20–100 members, the present study explores the ability of ASI to amplify accuracy in small teams of 3–6 members. The present study also explores if conducting multiple swarms and aggregating by taking a “vote of swarms” can further amplify the accuracy. A large set of 66 small teams were engaged in this study. Each team was given a standard subjective judgement test. Participants took the test both as individuals and real-time swarms. The average individual scored 69% correct, while the average swarm scored 84% correct ($p < 0.001$). In addition, aggregation of multiple swarms revealed additional amplifications of accuracy. For example, by randomly selecting sets of 3 swarms and aggregating by plurality vote, average accuracy increased to 91% ($p < 0.001$). These results suggest that when small teams make subjective judgements as real-time swarms, they can be significantly more accurate than individual members, and that their accuracy can be further amplified by aggregating the output across small sets of swarms.

Keywords: Swarm intelligence · Collective intelligence · Artificial intelligence · Human swarming · Wisdom of crowds · Artificial swarm intelligence · ASI

1 Introduction

In the natural world, Swarm Intelligence (SI) enables social organisms to rapidly aggregate their collective insights and reach optimal decisions by forming real-time closed-loop systems that converge in synchrony. Swarm Intelligence has been deeply studied across many social species, from schools of fish and flocks of birds, to swarms of honey bees. In recent years, advances in networking technology and artificial

intelligence have enabled human groups to form similar systems online, moderated by AI algorithms. Often referred to as Artificial Swarm Intelligence (ASI) or simply “human swarming,” this novel approach has been shown to significantly increase the accuracy of group decisions across a wide variety of tasks, from handicapping sporting events to forecasting financial markets [1–7].

While ASI has been found to significantly amplify decision-making accuracy in human groups comprised of 20–100 members, few studies have investigated the use of swarming among small groups on the order of 3–6 members. This is important to scholarship and practice as many important decisions are made by small teams.

2 Enabling “Human Swarms”

Unlike birds, bees and fish, humans have not evolved the natural ability to form real-time swarms, as we lack the innate mechanisms used by other species to form closed-loop systems. Schooling fish detect vibrations in the water around them. Flocking birds detect high-speed motions propagating through the group formation. Swarming bees generate complex body vibrations called a “waggle dance” that encode assessment information. To enable networked human groups to form similar closed-loop systems, a cloud-based platform called “swarm.ai” was developed, as shown in Fig. 1. It enables distributed human groups, connected over the internet, to make collective predictions, decisions, and assessments by working together as closed-loop swarms.

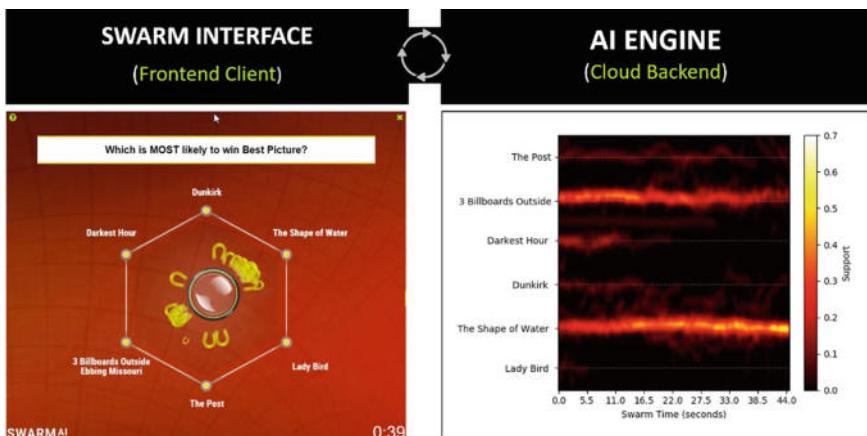


Fig. 1. Architecture of the **swarm.ai** platform with graphical client and cloud-based AI engine

When using the swarm.ai platform, networked human teams answer questions by collaboratively moving a graphical pointer to select from a set of answer options. Each participant provides input by manipulating a graphical magnet with a touchscreen or mouse. By adjusting the position and orientation of their magnet with respect to the moving puck, participants express their intent, not as a discrete vote, but a stream of vectors that varies freely over time. Because all members adjust their individual intent

continuously in real-time, the swarm explores the decision-space, not based on the input of any single member, but based on the emergent dynamics of the full system. The complex behavioral interactions among the full population are processed by swarming algorithms in real-time, empowering the unified system to converge on solutions that maximizes the collective confidence and conviction of the group.

It is important to note that participants not only vary the direction of their intent but also modulate the magnitude of their intent by adjusting the distance between their magnets and the pointer, which is commonly represented as a graphical puck. Because the graphical puck is in continuous motion across the decision-space, users need to continually move their magnets so that they stay close to the puck's rim. This is significant, for it requires that all participants, regardless of group size or composition, to be engaged continuously throughout the decision process, evaluating and re-evaluating their intent in real-time. If a participant stops adjusting their magnet with respect to the changing position of the puck, the distance grows and the participant's influence on the group's decision wanes.

Thus, like bees vibrating their bodies to express sentiment in a biological swarm, or neurons firing to express conviction levels within a biological neural-network, the participants in an artificial swarm must continuously update and express their changing preferences during the decision process or lose their influence over the collective outcome. This is generally referred to as a "leaky integrator" structure and common to both swarm-based and neuron-based systems. In addition, intelligence algorithms monitor the behaviors of swarm members in real-time, inferring their relative conviction based on their actions and interactions over time. This reveals a range of behavioral characteristics within the population and weights their contributions accordingly.

3 Accuracy Study

To assess the ability of ASI technology to amplify the accuracy of team decisions in subjective judgement tasks, a large study was conducted across a set of 66 working groups (i.e. business teams), each of 3–6 members. Each of these teams were already engaged in a long-term project together and had already established a working relationship among themselves. In total, 330 subjects participated in this study. All were college students in business, communication studies, and engineering courses, for which the team project was a significant component.

To rigorously measure accuracy in a standardized subjective judgement task, a widely used instrument was employed known as the "Reading the Mind in the Eyes" or RME test [8]. The test includes 35 questions, each of which provides a facial image cropped so that only a narrow region around the eyes is shown. A set of four options are provided that describe the emotion expressed by the person in the image, requiring participants to assess the emotional state based only on the eyes. An example question from a standard RME test is shown in Fig. 2. Four options are provided, only one of which accurately represents the emotion of the depicted individual.

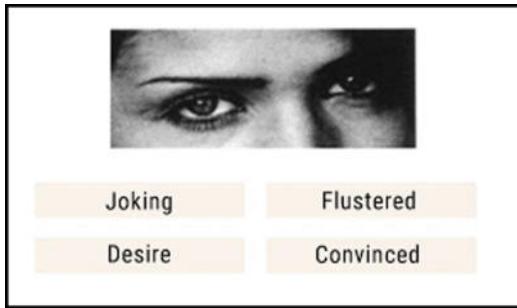


Fig. 2. Sample question from standard RME test

Prior studies have shown that the RME test is a reliable instrument with strong internal consistency and test-retest stability [9]. The performance in this subjective judgement task has been shown to indicate the Social Sensitivity of the test taker and is generally used for that purpose [10–12]. To test whether real-time swarming enabled small working groups to amplify their performance in the RME task, a two-stage process was employed. First, each of the 330 study participants were administered a 35-question test individually through an online survey. To limit bias and knowledge of correct answers, individual scores were not disclosed. In the second stage, each of the 66 teams was administered the RME test through an online swarming platform. This enabled each team to converge on each subjective judgement by working together as a real-time system, moderated by swarming algorithms. Teams were discouraged from communicating with each other verbally during the assessment, instead relying only on the closed-loop interaction afforded by the platform.

For each of the 35 subjective judgements in the test, the platform displayed one of the 35 facial images to all members of each team, along with the four potential assessments of that image. Each team was allotted up to 60 s to coverage upon an answer as a real-time swarm. Figure 3 is a snapshot of a team member's screen during a real-time swarm response. The magnets in the image represents the pull of each teammate at one instant in time. It should be noted that to discourage conformity, participants did not see the magnets during the actual swarming session.

4 Data and Analysis

The RME test was administered to 330 individuals across 66 teams and produced four unique datasets. First, we received fully completed individual assessments from 283 participants (86% response rate) totaling over 9000 item responses. These responses were used to calculate individual RME scores for each participant. Second, these same responses were aggregated by team to generate a “plurality vote” RME score for each question. This was calculated by assessing the most popular answer among the team for each question. For questions where the vote was split evenly across multiple answers (i.e. there was no plurality winner) a “deadlock” was determined and classified as incorrect. This provided a dataset of over 2000 plurality vote responses. Third, a swarm

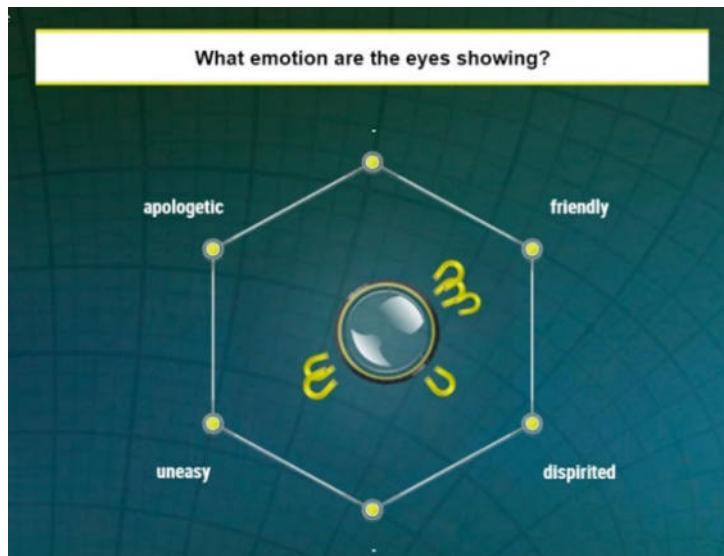


Fig. 3. Swarming group responding to RME question

RME score for each team was calculated from the responses collected through the online swarming platform. This provided a dataset of over 2000 swarm-based responses. For questions where the swarm could not converge upon an answer within the 60 s time limit, a “deadlock” was determined and classified as incorrect.

Finally, a “vote of swarms” RME score was generated by selecting random grouping of swarms for each question from the set of 66 teams and determining the final decision by plurality vote across the grouping. This was performed using a bootstrapping technique across a range of groupings of size $S = 3$ to $S = 10$ and repeated 1000 times for each size. For example, for $S = 3$, random groupings of three swarms were selected from the dataset and an RME score was generated based on a plurality vote across those 3 swarms. This process was repeated 1000 times for groupings of 3 swarms.

5 Results

Across the set of 330 subjects, each participating in one of 66 teams, a comparison was performed among four conditions:

1. **Individuals**—participants taking RME test alone
2. **Votes**—teams taking RME test by plurality vote
3. **Swarms**—teams taking RME test as real-time systems
4. **Votes of Swarms**—taking plurality vote among swarms.

Mean scores and error rates for RME were calculated for the individual, plurality, and swarm generated scores. As shown in the table of Fig. 4, the average individual RME score was 24.3, which corresponds to an error rate of 30.6%.

Testing Method (Deadlocks as Errors)	Mean # Correct	Error Rate	95% Error Rate Confidence Interval	95% Error Rate Difference to Swarm CI
<i>Individual Average</i>	24.3	30.58%	[27.84, 33.55]	[-20.44, -11.95]
<i>Plurality Voting</i>	25.45	27.27%	[24.11, 30.61]	[-18.98, -8.23]
<i>Swarm AI</i>	29.39	16.02%	[14.05, 18.06]	N/A

Fig. 4. Error rates and confidence intervals

The average of each team's plurality RME score was 25.45, which corresponds to an average error rate of 27.3%. When enabling the teams to work together as a swarm, the average RME score increased to 29.4, which corresponds to an average error rate of 16.0%. In other words, by working together as a swarm, the 66 teams, on average, reduced their error rates by 41%. This demonstrates that working as a swarm can significantly increase accuracy in subjective judgement tasks as compared to both individual performance and team performance by plurality vote.

To assess statistical significance, a bootstrap analysis of the error rate for each method was performed across 10,000 trials. The 95% confidence intervals and p -values were calculated for the difference between individual RME, plurality RME, and swarm RME scores. The results show that the swarm significantly outperforms both individual ($\mu_{\text{difference}} = 14.6\%$ error, $p < 0.001$) and plurality scores ($\mu_{\text{difference}} = 11.3\%$ error, $p < 0.001$). The bootstrapped error comparison is shown in Fig. 5.

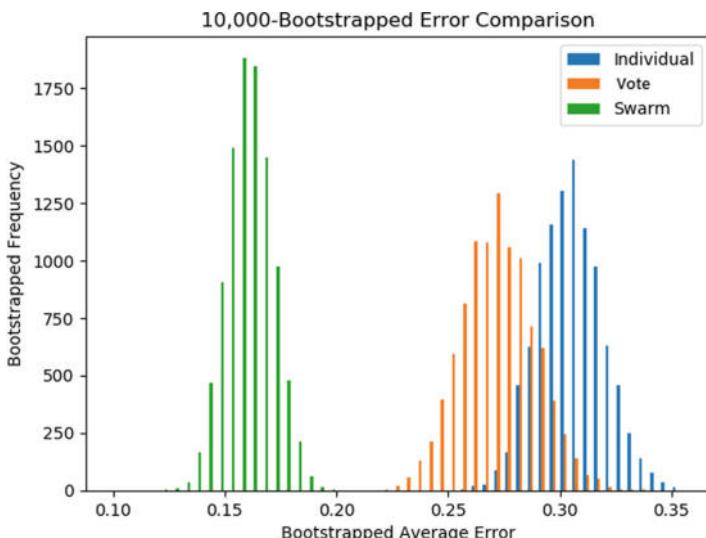


Fig. 5. Bootstrapped average error rate

With respect to deadlocks, a comparison was made between the rate of deadlocks determined by plurality vote as compared to the rate of deadlocks reached by swarms. Across the 66 working groups, plurality voting resulted in deadlocks in 14% of questions. Across those same groups, when working together as swarms, the rate of deadlocks dropped substantially to 0.6% of questions. This is a significant improvement, reducing the need for further steps to resolve undecided groups.

In addition, an analysis was performed assuming deadlocked votes were resolved by giving partial credit for tied answers that include a correct response: half credit for a two-way tie, third credit for a three-way tie, etc. To balance this, deadlocked swarms were given the chance to resolve immediately following a deadlock in another 60 s swarm, with the answer chosen in this second round selected as the final answer. There were no swarms that deadlocked twice in a row.

As shown in the table of Fig. 6, when deadlocks were resolved using partial credit, plurality vote averaged an RME score of 27.9 (an error rate of 20.4%). When enabling the swarms to work together as real-time systems and resolve their deadlocks in a follow-up swarm, the swarm RME score increased to 29.4 (an error rate of 15.9%). In other words, even when giving partial credit for deadlocks in group responses determined by plurality vote, the swarm outperformed.

Testing Method (Deadlocks Resolved)	Mean # Correct	Error Rate	95% Error Rate Confidence Interval	95% Error Rate Difference to Swarm CI
<i>Individual Average</i>	24.3	30.58%	[27.84, 33.55]	[-20.53, -12.08]
<i>Plurality Voting</i>	27.87	20.38%	[18.41, 22.46]	[-9.2, -2.47]
<i>Swarm AI</i>	29.43	15.9%	13.95, 17.93]	N/A

Fig. 6. Error rates and confidence intervals with deadlocks resolved

To assess statistical significance, a bootstrap analysis of the error rate for each method was again performed across 10,000 trials. We find that the swarm outperforms both the plurality vote ($\mu_{\text{difference}} = 4.5\%$ error, $p < 0.001$) and individuals ($\mu_{\text{difference}} = 14.7\%$ error, $p < 0.001$). The bootstrapping of the error rate confidence intervals is shown in Fig. 7.

In addition to comparing against the average individual, the swarm can be compared against all individuals. On average, swarms are in the 92nd percentile of individuals, indicating that an average swarm scores better than 92.2% of individuals taking the test alone. The histogram of user performance and average swarm performance is shown in Fig. 8.

Finally, we explored the aggregation of swarm responses by plurality vote to assess if the accuracy on the subjective judgement RME test could be further amplified as compared to individual swarm responses. This process, referred to herein as “aggregations of swarms,” was conducted by bootstrapping the average error rate of aggregations of swarms across a range of aggregation sizes from $S = 3$ to $S = 9$, with 1000 iterations of randomly selected aggregations performed for each aggregation size. The results are shown in Fig. 9. The single swarm case ($S = 1$) is bootstrapped and shown to depict how error rate decreases as the number of swarms aggregated increases.

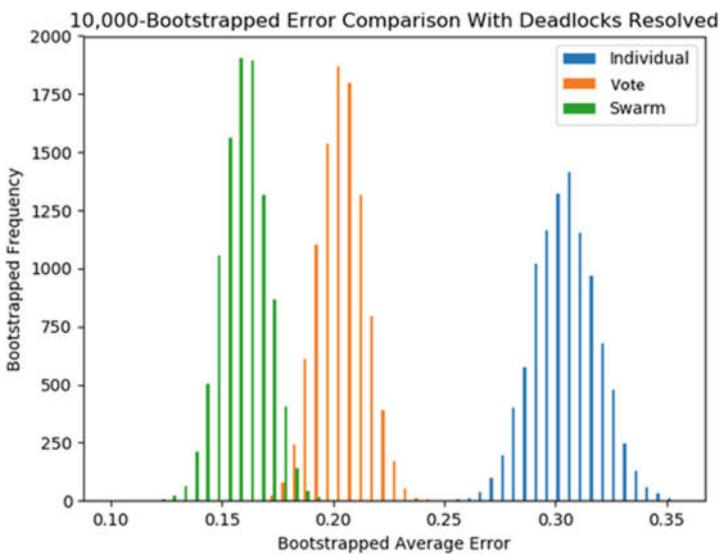


Fig. 7. Bootstrapped average error rate

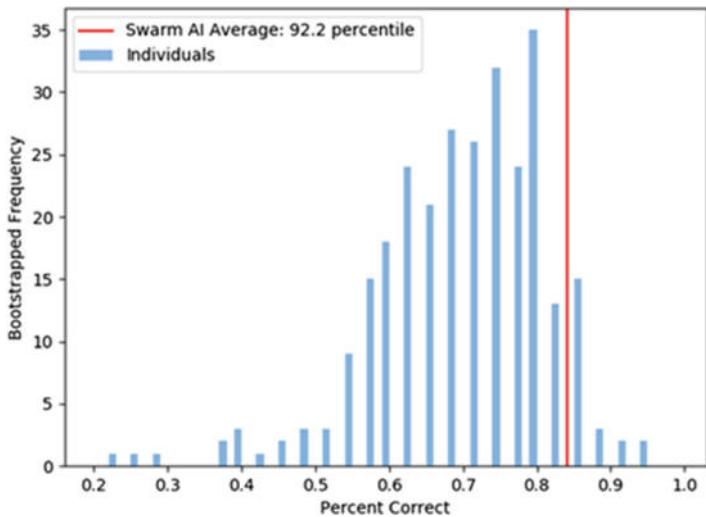


Fig. 8. Bootstrapped average error rate

We find that increasing the number of swarms aggregated decreases the error rate. In addition, the variation in performance decreases as the number of swarms aggregated increases. Not only do votes of swarms become more accurate as more swarms are aggregated, but they also become more consistent. The aggregation of as few as three swarms significantly outperforms single swarms ($\mu_{\text{difference}} = 6.9\%$ error, $p = 0.007$)

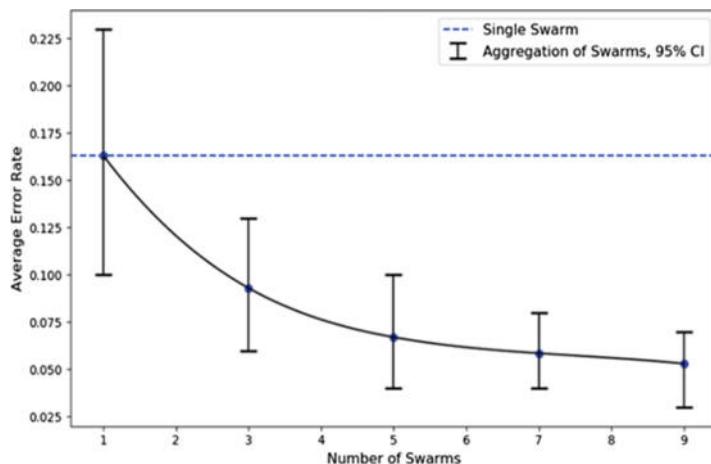


Fig. 9. Accuracy when swarms are aggregated by plurality vote

and the aggregation of five swarms significantly outperforms the aggregation of three swarms ($\mu_{\text{difference}} = 2.3\%$ error, $p < 0.044$).

The bootstrapped average error histogram created for individuals, swarms, and aggregations of three swarms is shown in Fig. 10. We find that an aggregation of three swarms outperforms individuals by an average error of 21.3%. So, by working together in swarms, and then aggregating three swarms together, the average error is reduced by 70% as compared to individual performance.

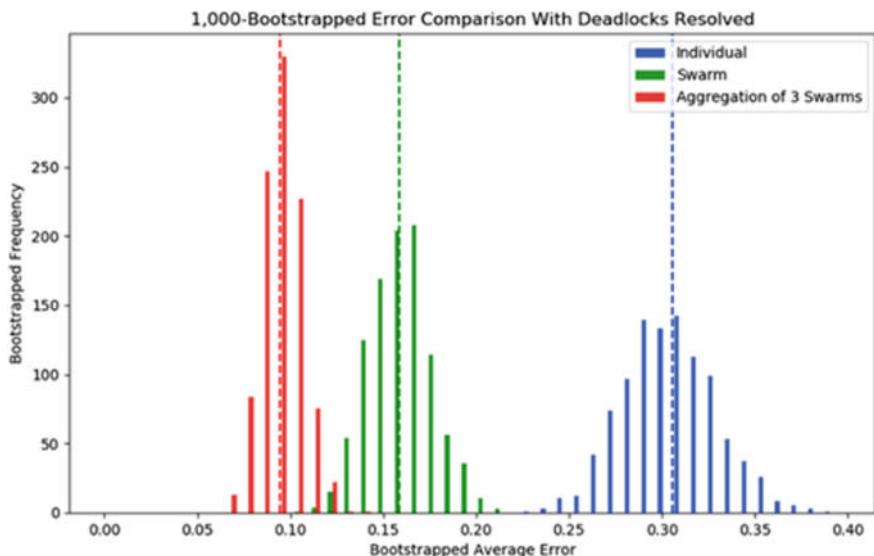


Fig. 10. Individual versus swarm versus aggregation of swarms

6 Conclusions

The results of this study suggest that small “human swarms” are significantly more accurate than individuals in subjective judgement tasks. As demonstrated across 66 working groups, each of 3–6 members, subjective judgement accuracy increased from 69% correct to 84% correct when participants worked together as real-time swarms. This corresponds to a reduction in error rate by 41%. In addition, the results of this study suggest that small human swarms are significantly more accurate than those same groups reaching subjective judgements by plurality vote, which demonstrated 73% accuracy. The probability that the swarm outperformed the individuals and the group vote by chance was very low ($p < 0.001$).

In addition, results of this study suggest that by aggregating the output from multiple human swarms, we can further increase accuracy on subjective judgment tasks. A range of aggregation sizes were explored from $S = 3$ to $S = 9$. Even when aggregating only three swarms at a time ($S = 3$), a significant increase in accuracy was observed, boosting performance from 84% correct for single swarms to 91% for aggregations. In other words, by having small human groups perform subjective judgement tasks as swarms, and then aggregating small sets of swarms, individual performance was increased from 69% accuracy (50th percentile) to 91% accuracy (98th percentile). These are a very significant results and suggests that real-time swarming may be a powerful method for boosting team performance, even among small teams of only 3–6 members.

References

- Rosenberg, L.B.: Human swarms, a real-time method for collective intelligence. In: Proceedings of the European Conference on Artificial Life 2015, pp. 658–659
- Rosenberg, L.: Artificial swarm intelligence vs human experts. In: 2016 International Joint Conference on Neural Networks (IJCNN). IEEE
- Rosenberg, L., Baltaxe, D., Pescetelli, N.: Crowds vs swarms, a comparison of intelligence. In: IEEE 2016 Swarm/Human Blended Intelligence (SHBI), Cleveland, OH, pp. 1–4 (2016)
- Baltaxe, D., Rosenberg, L., Pescetelli, N.: Amplifying prediction accuracy using human swarms. In: Collective Intelligence 2017, New York, NY (2017)
- Rosenberg, L., Pescetelli, N., Willcox, G.: Human swarms amplify accuracy in financial predictions. In: IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY (2017)
- Rosenberg, L., Willcox, G., Halabi, S., Lungren, M., Baltaxe, D., Lyons, M.: Artificial swarm intelligence employed to amplify diagnostic accuracy in radiology. In: 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, 2–4 Nov 2018
- Rosenberg, L., Willcox, G.: Artificial swarms find social optima: (late breaking report). In: 2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), pp. 174–178, Boston, MA (2018)
- Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., Plumb, I.: The “reading the mind in the eyes” test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry* **42**, 241 (2001)

9. Vellante, M., Baron-Cohen, S., Melis, M., Marrone, M., Petretto, D.R., Masala, C., Preti, A.: The “reading the mind in the eyes” test: systematic review of psychometric properties and a validation study in Italy. *Cogn. Neuropsychiatry* **18**, 326–354 (2012)
10. Fiske, S.T., Taylor, S.E.: Social Cognition: From Brains to Culture. Sage, London, UK (2013)
11. Kunda, Z.: Social Cognition: Making Sense of People. The MIT Press, Cambridge, MA (1999)
12. Frith, C.D., Singer, T.: The role of social cognition in decision making. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3875–3886 (2008)



High-Resolution Streaming Functionality in SAGE2 Screen Sharing

Kazuya Ishida^{1(✉)}, Daiki Asao², Arata Endo¹, Yoshiyuki Kido³,
Susumu Date³, and Shinji Shimojo³

¹ Graduate School of Information Science and Technology,
Osaka University, Suita, Japan

{kazuya.ishida,endo.arata}@ais.cmc.osaka-u.ac.jp

² School of Engineering, Osaka University, Suita, Japan

asao.daiki@ais.cmc.osaka-u.ac.jp

³ Cybermedia Center, Osaka University, Suita, Japan

{kido,date,shimojo}@cmc.osaka-u.ac.jp

Abstract. Visualization on a Tiled Display Wall (TDW) is an effective approach for sharing large quantities of scientific data among researchers in collaborative research. SAGE2 (Scalable Amplified Group Environment) is a popular middleware for organizing multiple monitors into a TDW. SAGE2 has a useful function, Screen Sharing, which allows the user to utilize existing applications on a TDW without redevelopment. However, the current Screen Sharing has a problem in that it displays the application window at the same resolution as the monitor devices connected to the user's PC. Such insufficient resolution degrades the visibility of scientific data visualized on a TDW. In this paper, we incorporate a frame-streaming method we designed into Screen Sharing to realize the functionality of high-resolution streaming. Our evaluation demonstrates that our method enables Screen Sharing to display the application window at an arbitrary resolution on a TDW. Our method is also effective in improving the frame rate of Screen Sharing. For example, 19.2 fps is achieved when displaying a 4K application on a TDW.

Keywords: Visualization · Tiled display wall · SAGE2 · Screen sharing

1 Introduction

The collaboration of researchers is essential for scientific discovery in modern science. To deal with difficult and complex tasks, the collective effort of many researchers has often been required. For example, in the KBDD (K supercomputer-based drug discovery) project [1], researchers from various companies and academic institutions have worked together to discover chemical compounds for novel drugs through molecular dynamics simulations performed by the K supercomputer.

In collaborative research, researchers involved in a project often have discussions based on the scientific data they acquire. As a consequence of the recent performance enhancement of computers and measuring equipment, the opportunities for researchers to treat large quantities of scientific data have been increasing. To facilitate research discussions using such large data, the data should be intuitively visualized and arranged on a high-resolution screen so that it can be easily understood and compared with other data by the researchers. In the KBDD project, for example, the results of the molecular dynamics simulations are supposed to be visualized in the discussions on the druggability of chemical compounds. Since the researchers discuss druggability while comparing the targeted compound with many different types of other compounds, the multiple visualized simulation results need to be laid out on a high-resolution screen [2]. For this purpose, visualization on a TDW (Tiled Display Wall) [3] is an appropriate approach. A TDW is a scalable visualization system, which can provide a high-resolution virtual screen by combining multiple sets of computers and monitors. A large-scale TDW allows a large number of researchers to observe the visualized data simultaneously and to exchange ideas with each other on the spot.

To construct a large-scale TDW, various middleware have been developed, such as SAGE2 (Scalable Amplified Group Environment) [4], CGLX (Cross Platform Cluster Graphics Library) [5] and DisplayCluster [6]. In particular, SAGE2 has been accepted in various research fields because of its advantageous features for both co-located and remote collaborative work. In reality, 91 or more universities/institutes all over the world have adopted SAGE2 [7].

The major reason why SAGE2 is suitable for collaborative research is that SAGE2 provides Screen Sharing, which is an optional function to leverage existing visualization applications on a TDW. Most middleware-based TDW can display only the applications developed with the special APIs provided by the middleware. To use an existing application on such a TDW, the user is required to add the API calls in the source code and re-compile it. In contrast, a SAGE2-based TDW can perform the mirroring of an application window generated on the user's PC by using Screen Sharing. This function allows the user to leverage a wide range of existing applications in a non-invasive manner.

However, the current Screen Sharing has an undesirable problem in that it cannot display an application window at an arbitrary resolution on a TDW. This is because Screen Sharing is realized by capturing video frames on the user's PC and streaming them to the TDW. The resolution of these video frames is determined by the size of a framebuffer on the user's PC. A framebuffer is a memory space to store image data drawn by applications. The available sizes for a framebuffer are confined to the specific values supported by the monitor devices connected to the user's PC video card device. Therefore, if the resolution which the connected monitor devices support is much lower than the TDW, the application window is displayed on the TDW with its visibility degraded.

In this paper, we propose a frame-streaming method which enables Screen Sharing to display a high-resolution application window regardless of the lim-

itations of the connected monitor devices. In addition, our method improves the efficiency of the processing in streaming video frames to achieve a better frame rate. The remainder of this paper is structured as follows. In Sect. 2, we present the overview of SAGE2 and Screen Sharing. In Sect. 3, we describe the technical problems and issues we tackled. In Sect. 4, works related to our research are introduced. Section 5 shows the frame-streaming method for realizing high-resolution streaming functionality in Screen Sharing. The experiments for evaluating the validity and scalability of the proposed method are described in Sect. 6. In Sect. 7, we conclude this paper and note our future work.

2 SAGE2 Screen Sharing

2.1 Architecture of SAGE2

SAGE2 (Scalable Amplified Group Environment) [4] is an open-source middleware, which has been developed at EVL (the Electronic Visualization Laboratory) at the University of Illinois, Chicago (UIC). SAGE2 can build a large virtual desktop on a TDW and arrange multiple application windows on it just like a window manager. In addition, SAGE2 has a function to distribute the visualized contents to other TDWs in geographically separated sites via a WAN (Wide Area Network).

SAGE2 is cloud and web-browser-based technology, which is implemented with HTML5, JavaScript and WebGL. Figure 1 shows the architecture of a general SAGE2-based TDW. A SAGE2-based TDW consists of three components: the *display clients*, the *interaction client* and the *SAGE2 server*. The display clients are the HTML pages opened by the full screen web browsers on all the monitors of the TDW. Each display client renders the particular domain of the virtual desktop corresponding to its own client ID. The interaction client is the web page with the JavaScript programs to provide SAGE UI, which is a graphical user interface of SAGE2. Through the SAGE UI, the user can move and resize the windows on the virtual desktop. The SAGE2 server is a web server that plays the core role of controlling the entire TDW. The SAGE2 server reflects the results of user operations on SAGE UI and synchronizes the other components by exchanging the control messages through a WebSocket protocol. The control message is composed of a message name and several data (e.g. window size and image data). When each component receives a control message, a callback process corresponding to its message name is executed on the component.

2.2 Screen Sharing

Screen Sharing is the function of SAGE2 for sharing the user's desktop contents among the members of the project. This function is realized by streaming video frames of the application window to the TDW with several browser APIs such as *getUserMedia()*, *drawImage()* and *toDataURL()*. Thanks to Screen Sharing, SAGE2 allows the user to leverage existing applications on a TDW without modification.

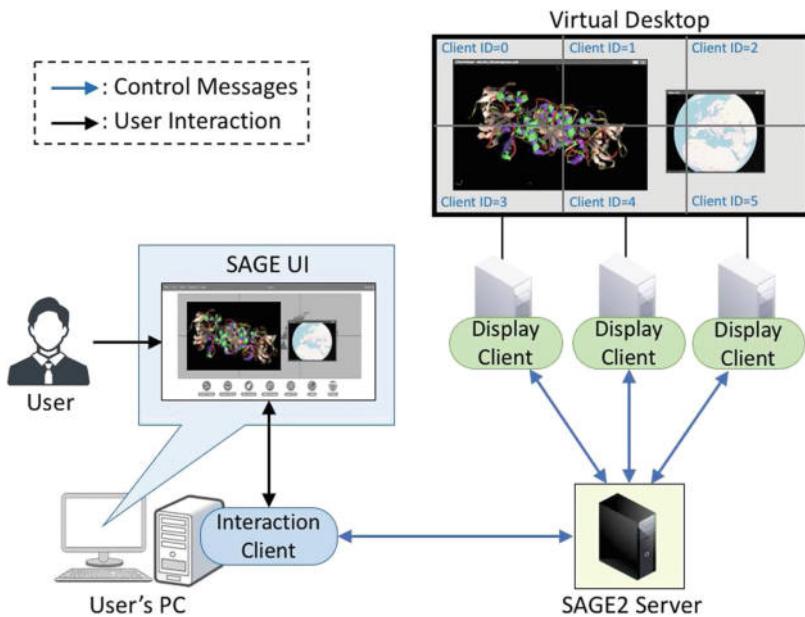


Fig. 1. Architecture of the SAGE2-based TDW

Screen Sharing is the function of SAGE2 for sharing the user's desktop contents among the members of the project. This function is realized by streaming video frames of the application window to the TDW with the several browser APIs such as `getUserMedia()`, `drawImage()` and `toDataURL()`. Thanks to Screen Sharing, SAGE2 allows the user to leverage existing applications on a TDW without modification of them.

Figure 2 illustrates how Screen Sharing works. If the user launches Screen Sharing through SAGE UI, the WebSocket connection is established between the user's PC and the SAGE2 server by exchanging the `requestToStartMediaStream` message and the `allowAction` message. Next, the interaction client starts capturing video of the application window in the background process by using a `getUserMedia()` API. This API captures video from a framebuffer, which is memory space allocated on the graphics memory so that an X server can store image data drawn by X applications. The captured video is streamed to the hidden video element of the interaction client. Then, the loop for streaming the frames of the captured video is iterated continually until Screen Sharing is terminated by the user. This loop is composed of the following four steps ((1)(4) in Fig. 2).

- (1) *Extraction:* The interaction client extracts the latest frame from the captured video, and places this frame onto the hidden canvas element by using the `drawImage()` API.

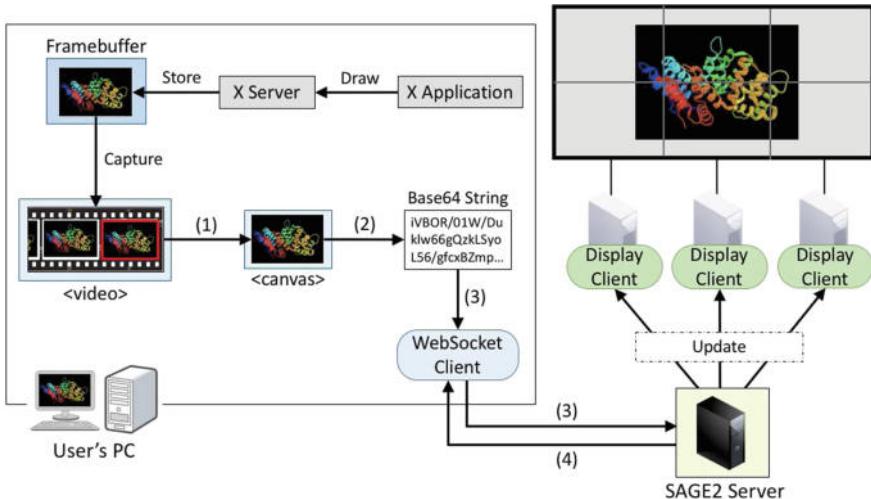


Fig. 2. System architecture for screen sharing

- (2) *Conversion*: The extracted frame is converted to a JPEG image in Base64 format with the `toDataURL()` API.
- (3) *Transmission*: The Base64 string of the frame is transmitted as the `update-MediaStreamFrame` message to the SAGE2 server.
- (4) *Synchronization*: After the SAGE2 server updates the frame to all the display clients, the SAGE2 server passes the `requestNextFrame` message to the user's PC.

3 Problems and Issues

The current Screen Sharing of SAGE2 has the following two problems: *Resolution Constraint* and *Conversion Delay*. In this section, we describe these problems and the technical issues required to overcome them.

3.1 Resolution Constraint

As mentioned in Sect. 2.2, the current Screen Sharing captures video frames of the application window from the framebuffer on the user's PC. The resolution of these video frames is determined by the size of the framebuffer on the user's PC. Generally, since the size of a framebuffer is fixed to the specific values which are supported by the monitor devices connected to the user's PC video card device, there is a limit to the resolution at which Screen Sharing can display the application on a TDW.

This *Resolution Constraint* problem causes the visibility of the application on a TDW to deteriorate when there is a large difference in the number of pixels

between the user's PC and the TDW. An example of such a case is depicted in Fig. 3. In this example, the screen resolution of the user's PC is HD (1280 × 720) and the total resolution of the TDW is 7680 × 3240. In this case, Screen Sharing displays the application as a relatively small sized window on the TDW ((a) in Fig. 3). Although the user can resize this window to make it larger through SAGE UI, the enlarged window will become a rough image ((b) in Fig. 3). Such poor visibility becomes a hindrance to gaining information and insights from visualized scientific data. To improve the visibility of the application using Screen Sharing, a technical issue arises: namely, how a flexible framebuffer can be prepared on the user's PC without being constrained by the connected monitor devices.

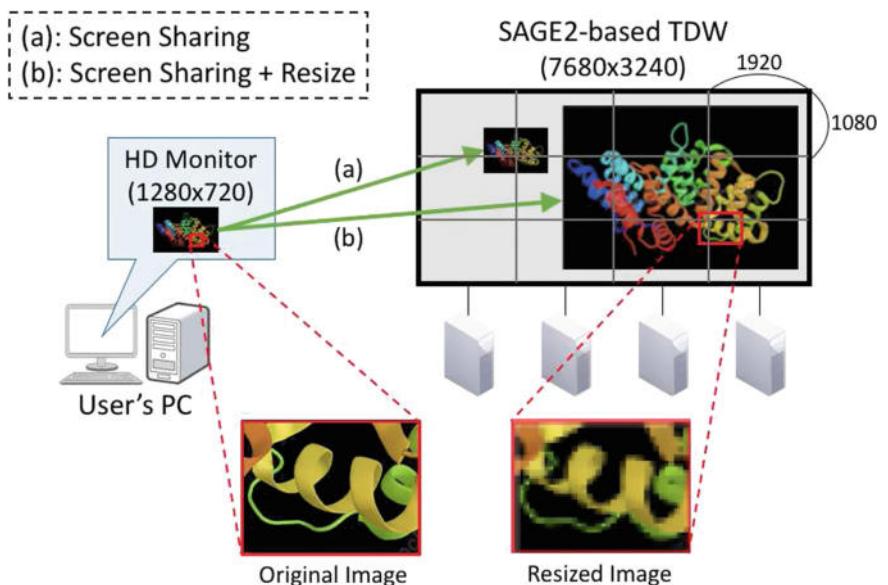


Fig. 3. Example of the deterioration of visibility caused by the *Resolution Constraint* problem

3.2 Conversion Delay

In the current Screen Sharing, video frames captured on the user's PC are streamed to a TDW by iterating four steps: (1) *Extraction*, (2) *Conversion*, (3) *Transmission* and (4) *Synchronization*. In advance of this research, we measured the processing time to refresh one frame in the current Screen Sharing by varying its resolution. This precursor experiment was conducted on the environment detailed in Sect. 6.1.

The result of the precursor experiment is shown in Fig. 4. This graph suggests that each step requires more processing time as the resolution of video frames become higher. In particular, the processing time for step (2) (*Conversion*) sharply rises related to the increase of the resolution. In other words, the processing for JPEG compression and Base64 encoding causes a serious delay in Screen Sharing. This *Conversion Delay* problem leads to a significant degradation of the frame rate when Screen Sharing displays a high-resolution application window on a TDW. To suppress the degradation of the frame rate in Screen Sharing, another technical issue arises: namely, how the delay time by the step (2) can be reduced.

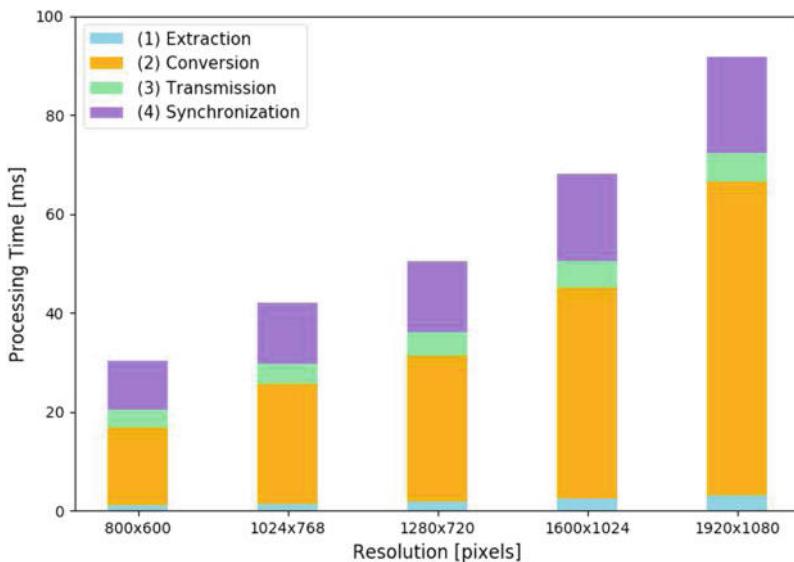


Fig. 4. Breakdown of the processing time to refresh one frame in the current screen sharing

4 Related Works

We have focused on displaying an existing application on a middleware-based TDW without source code modification. Some other researchers have also worked on this topic.

Tada et al. proposed an adapter solution for SAGE [8]. SAGE (Scalable Adaptive Graphics Environment) [9] is a middleware that is a predecessor of SAGE2. Unlike SAGE2, SAGE does not provide Screen Sharing. To utilize an existing application on a SAGE-based TDW, developers have to add many APIs provided by SAGE like the other middleware. Their adaptor solution allows

the users to display a window of the existing application on a TDW without such troublesome work. As with our method, their solution adopts a virtual framebuffer to capture image data drawn by applications. However, our research also pursues a method to improve the frame rate in a streaming high-resolution window.

Kimball et al. developed a framework to stream a user's desktop contents to a CGLX-based TDW [10]. CGLX (Cross Platform Cluster Graphics Library) [5] is a middleware for building a TDW, which has an OpenGL-based distributed rendering architecture. In their framework, captured frames are compressed with H.264 video encoding and delivered to a TDW using UDP multicast. Their research differs from our research in that they aim to achieve low bandwidth and low latency in the desktop streaming for CGLX. Our research aims to incorporate a high-resolution streaming functionality into the Screen Sharing of SAGE2.

Neal et al. implemented ClusterGL, which is a system to make existing OpenGL applications available on a TDW [11]. ClusterGL captures OpenGL commands from OpenGL applications and distributes them with their arguments to the renderers on display nodes. To enhance its performance, ClusterGL also performs several optimizations such as frame differencing and data compression. ClusterGL has a disadvantage in that it can be used only for limited applications because it supports only OpenGL 2.1 or earlier. In contrast, SAGE2 Screen Sharing with our method can be applied to a wider range of applications because it adopts not command streaming but frame streaming.

5 Proposed Frame-Streaming Method

5.1 Overview of Proposed Method

We proposed the frame-streaming method for achieving the issues described in Sect. 3. Figure 5 overviews the system architecture for Screen Sharing with the frame-streaming method we designed. The red and blue squares in Fig. 5 are the implementation to incorporate the proposed method in Screen Sharing.

First, a Xvnc server is launched on the user's PC ((A) in Fig. 5). The user has to direct the X application to draw image data on the virtual framebuffer of the Xvnc server, then access it via the VNC client. Next, the WebSocket connection is established between the user's PC and the SAGE2 server in the same way as the current Screen Sharing. After that, the procedures for streaming video frames are repeated asynchronously by the three types of threads: the *extraction thread*, the *conversion threads* and the *transmission thread* ((B-1)(B-3) in Fig. 5). The extraction thread is in charge of capturing the video from the virtual framebuffer and extracting the latest frame from it. The extracted frames are given serial numbers and stored in its queue. The conversion threads undertake the parallel frame conversion. Each conversion thread gets one frame from the queue in the extraction thread and converts it to a JPEG image in Base64 format. The converted frames are collected in the queue of the transmission thread. The transmission thread repeatedly sends the converted frames to the SAGE2 server according to the following three procedures ((1)(3) in Fig. 5).

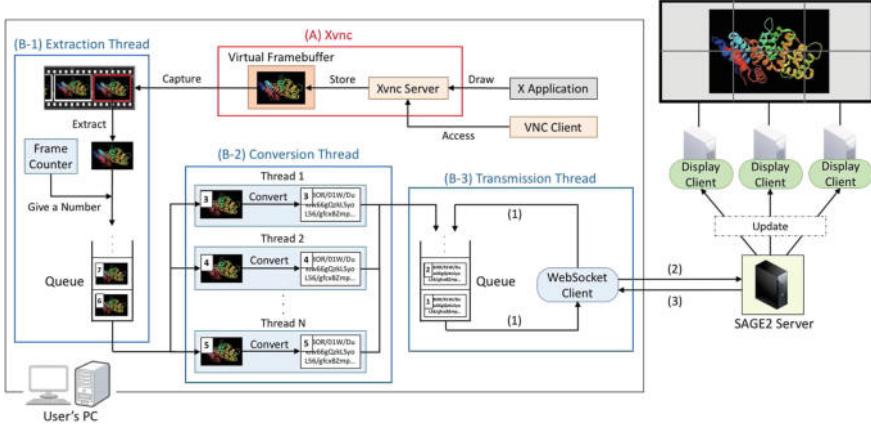


Fig. 5. System architecture for the screen sharing with the proposed method

- (1) The converted frame is retrieved from the queue and its serial number is checked. If the number is not the one which the next frame should have, the frame is returned to the queue and the procedure (1) is redone.
- (2) The Base64 string of the retrieved frame is transmitted as the *updateMediaStreamFrame* message to the SAGE2 server.
- (3) The thread waits until the *requestNextFrame* message is sent back from the SAGE2 server.

5.2 Xvnc

To achieve a solution to the *Resolution Constraint* problem, we applied Xvnc [12]. Xvnc is a VNC (Virtual Network Computing) server that can also act as an X server. The comparison between an X server and an Xvnc is shown in Fig. 6. The Xvnc is different from the usual X server in that the Xvnc uses a virtual framebuffer instead of a normal framebuffer. A virtual framebuffer is alternative memory space allocated on the shared memory for off-screen rendering. Xvnc stores image data drawn by X applications on its own virtual framebuffer, and does not output the frames to the connected monitor devices. To see and access X applications displayed on the virtual framebuffer, the user has to access them via a VNC client.

As opposed to a normal framebuffer, the size of virtual framebuffer can be optionally changed independent of the specification of the monitor devices. Therefore, Xvnc allows the user's PC to prepare the application window drawn at a larger resolution than the connected monitor devices can handle. The available resolution in the current Xvnc ranges from 32×32 to 32768×32768 , which is the sufficient support for displaying an application window on a TDW.

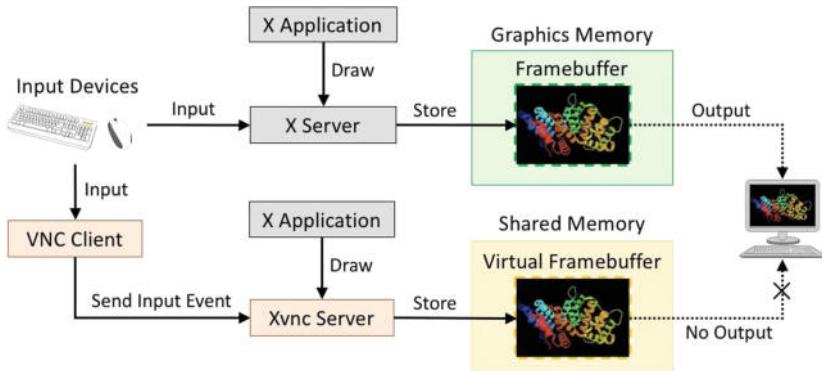


Fig. 6. Comparison between X server and Xvnc

5.3 Pipeline Streaming

To solve the issue for the *Conversion Delay* problem, we designed *pipeline streaming*: a concept for reducing wait time that occurs due to the processing for *Conversion*. Figure 7 illustrates how pipeline streaming works. The current Screen Sharing performs *sequential streaming*: the four steps for streaming video frames are iterated sequentially as shown in (a) in Fig. 7. In sequential streaming, refreshment of the TDW slows down greatly with respect to the increase of the resolution because of the processing for *Conversion*. This delay can be pre-

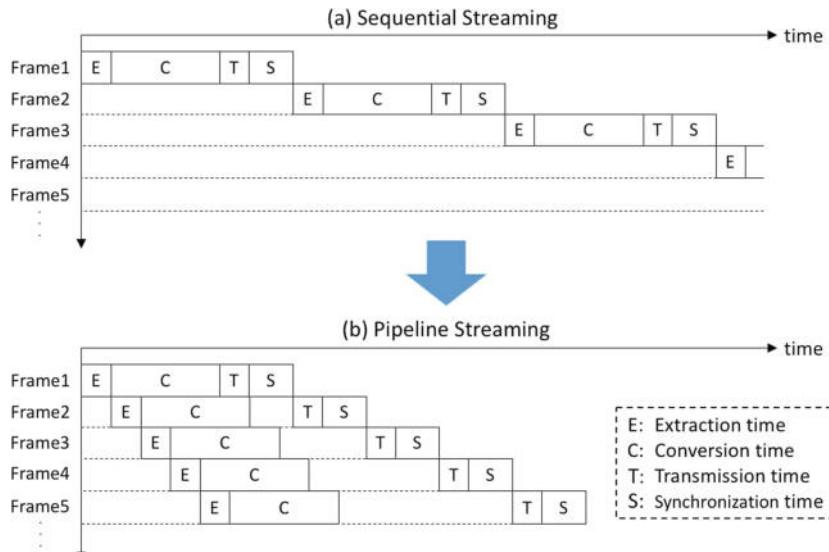


Fig. 7. Concept of pipeline streaming

vented by pipelining streaming in Screen Sharing as in (b) in Fig. 7. In pipeline streaming, *Extraction* and *Conversion* for the succeeding frames are executed antecedently during the process for the current frame. Moreover, each processing for *Conversion* is executed simultaneously with the other ones by using thread-level parallelism for further improvement of the efficiency.

6 Performance Evaluation

We conducted two experiments to evaluate the validity and the scalability of Screen Sharing with the proposed method.

6.1 Evaluation Environment

For the evaluation, we used the 24-screen Flat Stereo Visualization System in the Cybermedia Center, Osaka University [13]. This TDW is composed of seven nodes (one head node and six display nodes), each of which has the specification as shown in Table 1. All nodes composing the TDW are connected to a dedicated 10 Gigabit Ethernet switch. The head node has a Full HD (1920×1080) monitor and input devices (a mouse and a keyboard) and each display node is connected to four Full HD monitors arranged in tandem.

Figure 8 shows the evaluation environment used in the experiments. One display node was used as both the SAGE2 server and the display client. The other display nodes were used only for the display clients. All the display clients were accessed via Chromium browsers [14]. The head node became a client node to launch Screen Sharing with the proposed method. To use Xvnc, a TurboVNC server and viewer [15] were installed on the client node. The JPEG compression quality of the video frames was configured as 90. We used Cytoscape [16] as the application displayed on a TDW. Cytoscape is a representative software to visualize complex networks such as molecular interaction networks and biological pathways. The versions of these software are presented in Table 2.

6.2 Observation of Visibility

To confirm the improvement of the visibility of the application, we displayed the window with Cytoscape at the maximum resolution of the TDW (11520×4320)

Table 1. Node specifications of the TDW

Name	Specification
CPU	Intel Xeon E5-2640 (2.5 GHz) ×2
GPU	NVIDIA Quadro K5000 ×1
Memory	64 GB
OS	CentOS 7.2

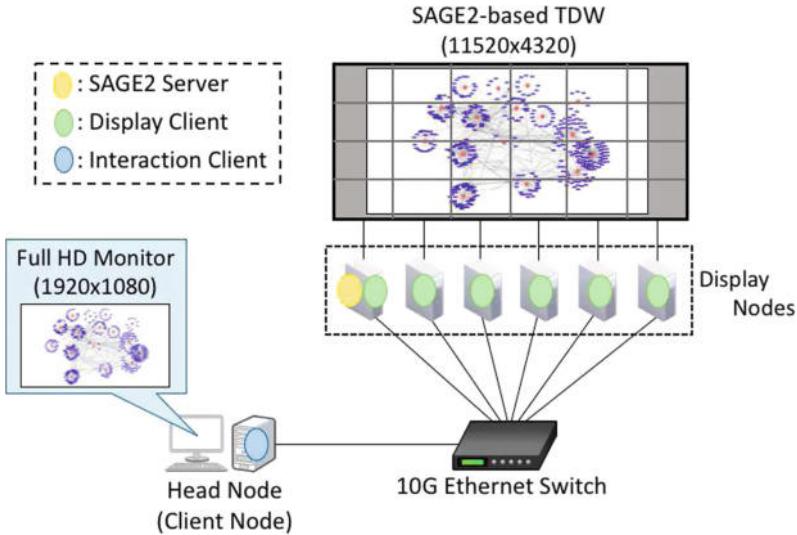


Fig. 8. Evaluation environment

Table 2. Software versions used in the evaluation

Name	Version
SAGE2	3.0.0
Chromium browser	65.0.3325.181
TurboVNC server	2.1.2
TurboVNC viewer	2.1.2
Cytoscape	3.6.1

using the Screen Sharing with the proposed method. Figure 9 is a snapshot of the Cytoscape on the TDW. We also displayed the same window using the current Screen Sharing and the resize operation, and compared them.

From (a) and (b) in Fig. 10, it can be seen that the proposed method enables Screen Sharing to display the application window more precisely. The key difference between (a) and (b) is the visibility of the label strings. In (b), the label strings became blurry as a result of the resize operation. By contrast, all the label strings in (a) are clearly displayed. This result shows that the proposed method is effective in improving the visibility of the application window displayed by Screen Sharing.

6.3 Frame Rate Measurement

To investigate the scalability of Screen Sharing with the proposed method, we measured the frame rate by varying the resolution of the window with Cytoscape.

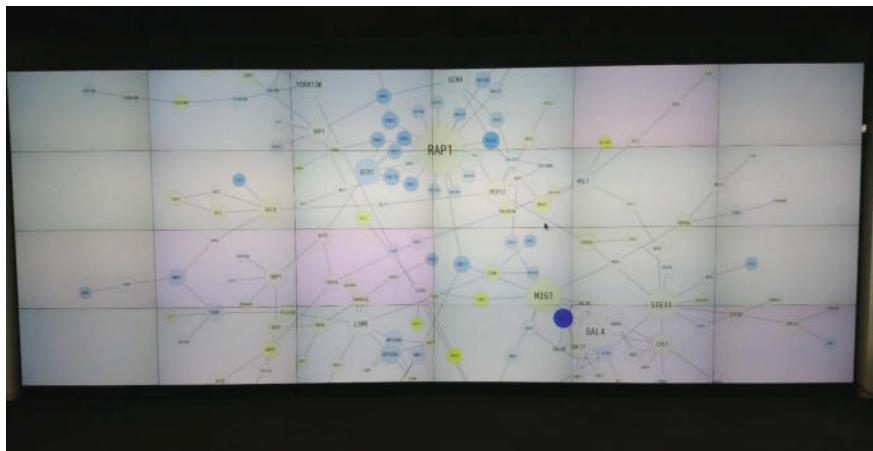


Fig. 9. Cytoscape on the TDW

The number of the conversion threads was also changed to 14. In addition, the frame rate achieved by Screen Sharing with the sequential streaming (i.e. the current Screen Sharing) was also surveyed.

The result of the measurement is depicted in Fig. 11. This graph indicates that the Screen Sharing with the proposed method outperforms the Screen Sharing with the sequential streaming at all the resolution which we measured. Moreover, it is implied that the frame rate is improved by increasing the number of the conversion threads. The proposed method enabled Screen Sharing to achieve 19.2 fps (frames per second) when displaying the window at 4 K resolution (3840×2160).

6.4 Discussion

Sections 6.2 and 6.3 demonstrated that our proposed method enabled Screen Sharing to stream a high-resolution application window at a better frame rate. However, there is still room for improvement in the proposed method for more practicality. The points to be improved are discussed below.

First, it is necessary to make the proposed method available on the platforms other than X Window System. The current proposed method can be applied only to X applications because of Xvnc. To display a wider range of applications on a TDW, optional kinds of VNC servers have to be leveraged in the proposed method. For example, the Xvnc server should be replaced with a TightVNC server [17] to display Windows applications.

Second, further extension of pipeline streaming is needed to enhance the frame rate when streaming a high-resolution window. In Fig. 11, the degree of improvement of the frame rate becomes smaller as the resolution increases. There are two reasons for this result. Firstly, the processing time for *Conversion* becomes too large to be suppressed by pipelining. To realize maximum efficiency,

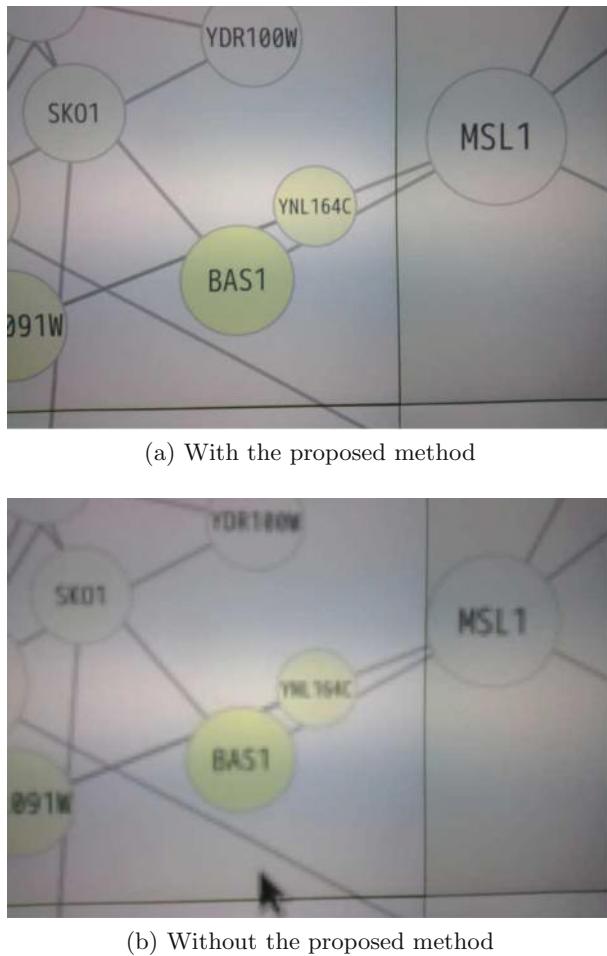


Fig. 10. Comparison of visibility on the TDW

Conversion for each frame should be completed by the end of *Synchronization* for the previous frame. Nevertheless, it becomes impossible to achieve this requirement when streaming the high-resolution window because of the rapid increase of the processing time for *Conversion*. To deal with this inconvenience, accelerating *Conversion* in itself is necessary. For example, it is effective to re-implement the processing for *Conversion* by using the primitive programming languages such as C/C++ or Fortran. Secondly, the processing time for *Transmission* and *Synchronization* becomes so large that it cannot be ignored. When the resolution of the video frames is high, a serious delay is likely to be caused by the network congestion in *Transmission* and *Synchronization*. This situation slows

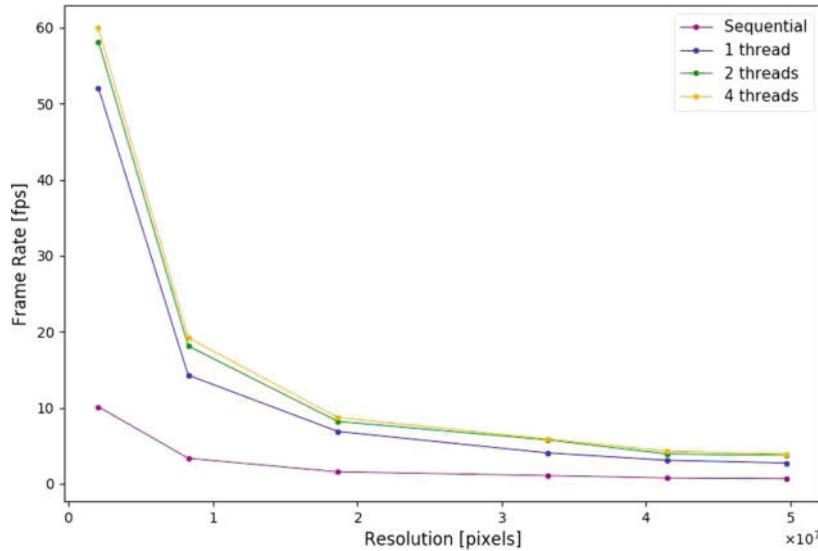


Fig. 11. Frame rate of screen sharing

down the entire streaming because *Transmission* and *Synchronization* cannot be pipelined. To cope with this situation, it is essential to reduce the network traffic of the streaming with some techniques such as data compression.

7 Conclusion

In this paper, we have developed a frame-streaming method to realize the high-resolution streaming functionality in Screen Sharing. The evaluation in this paper demonstrated that our proposed method enables Screen Sharing to stream a high-resolution application window regardless of the specifications of the connected monitor devices. The evaluation also showed that our proposed method improves the frame rate with pipeline streaming.

For our future work, we will tackle the new problems associated with the proposed method described in Sect. 6.4. Through this effort, we aim to make SAGE2 more applicable to various research areas.

Acknowledgements. This work was supported by the JSPS KAKENHI Grant Number JP18K11355 and the “Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures” in Japan (Project ID: jh160056-ISH, jh170056-ISJ, jh180077-ISJ).

References

1. Brown, J.B., Nakatsui, M., Okuno, Y.: Constructing a foundational platform driven by Japan's K supercomputer for next-generation drug design. *Mol. Inf.* **33**, 732–741 (2014)
2. Lau, C.D., Levesque, M.J., Chien, S., Date, S., Haga, J.H.: ViewDock TDW: High-throughput visualization of virtual screening results. *Bioinformatics* **26**(15), 1915–1917 (2010)
3. Humphreys, G., Buck, I., Eldridge, M., Hanrahan, P.: Distributed rendering for scalable displays. In: Proceedings of the 2000 ACM/IEEE Conference on Super-Computing, vol. 30 (2000)
4. Marrinan, T., Aurisano, J., Nishimoto, A., Bharadwaj, K., Mateevitsi, V., Renambot, L., Long, L., Johnson, A., Leigh, J.: SAGE2: a new approach for data intensive collaboration using scalable resolution shared displays. In: Proceedings of the 2014 IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 177–186 (2014)
5. Doerr, K.U., Kuester, F.: CGLX: A scalable, high-performance visualization framework for network displays. *IEEE Trans. Vis. Comput. Graph.* **17**(3), 320–332 (2011)
6. Johnson, G.P., Abram, G.D., Westing, B., Navr'til, P., Gaither, K.: DisplayCluster: an interactive visualization environment for tiled displays. In: Proceedings of the 2012 IEEE International Conference on Cluster Computing, pp. 239–247 (2012)
7. Community—SAGE2. <http://sage2.sagecommons.org/community-2>
8. Tada, T., Date, S., Shimojo, S., Ichikawa, K., Abe, H.: A visualization adapter for SAGE-enabled tiled display wall. In: Proceedings of the 2011 IEEE International Conference on Granular Computing, pp. 613–618 (2011)
9. Renambot, L., Jeong, B., Jagodic, R., Johnson, A., Leigh, J.: Collaborative visualization using high-resolution tiled displays. In: Proceedings of the 2006 ACM CHI Workshop on Information Visualization Interaction Techniques for Collaboration Across Multiple Displays, pp. 1–4 (2006)
10. Kimball, J., Wypych, T., Kuester, F.: Low bandwidth desktop and video streaming for collaborative tiled display environments. *Future Generation Computer Systems* **54**, 336–343 (2016)
11. Neal, B., Hunkin, P., McGregor, A.: Distributed OpenGL rendering in network bandwidth constrained environments. In: Proceedings of the 2011 Eurographics Symposium on Parallel Graphics and Visualization, pp. 21–29 (2011)
12. Xvnc—the X-based VNC server. https://www.hep.phy.cam.ac.uk/vnc_docs/xvnc.html
13. Cybermedia Center: Osaka University large-scale visualization system. <http://viscmc.osaka-u.ac.jp>
14. Reis, C., Gribble, S.D.: Isolating web programs in modern browser architectures. In: Proceedings of the 2009 ACM European Conference on Computer Systems, pp. 219–232 (2009)
15. TurboVNC—Main/turboVNC. <https://www.turbovnc.org>
16. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**(11), 2498–2504 (2003)
17. TightVNC: VNC-compatible free remote control/remote desktop software. <https://www.tightvnc.com>



A Secure Scalable Life-Long Learning Based on Multiagent Framework Using Cloud Computing

Ghalib Ahmad Tahir, Sundus Abrar, and Loo Chu Kiong^(✉)

University of Malaya, Kuala Lumpur 50603, Malaysia
 {ghalib,sundus.abrar}@siswa.um.edu.my,
 ckloo.um@um.edu.my

Abstract. The major problem on the road to artificial intelligence is the development of lifelong learning systems. They have the ability to learn new concepts incrementally overtime. They are also able to allocate required resources dynamically without human intervention and are able to store data securely. In this work we have extended the incremental classifier and representation learning method known as iCaRL to meet this criterion. The proposed solution is able to learn strong classifiers and a data representation simultaneously. It is able to allocate an optimal scaling plan to meet its resource requirements without human intervention. It securely stores propriety image data by using state of the art interplanetary file system and block chain technology. Finally, it is able to focus on object of interests in an image using attention network. We have shown by experiments on CIFAR-100 and Image net 2012 that it performs better in terms of accuracy than the existing iCaRL system while satisfying criteria of lifelong learning.

Keywords: Machine learning · Auto scaling · Life long learning · Incremental learning

1 Introduction

The past few decades have seen significant advancement in the area of machine learning theory and algorithms. However, there has been relatively little effort to build systems that use these algorithms to learn new tasks by retaining the knowledge of the task while maintaining the scalability and security of the system. We observe in our daily life that natural vision systems like human eyes are incremental in nature. They continuously learn new information while preserving the existing knowledge at the same time. Another practical example is of our own self. Throughout the day we continuously learn new things like meeting new people, trying new food items, looking at various design of clothes, different working environments etc. All this new information is retained in our memory without forgetting our previous knowledge and preferences. In addition, some important features of natural vision systems are (1) security: we cannot pass information about the things which we have seen without our consensus and (2) scalability: we continuously learn new things and are able to determine objects of interest in an unsupervised way. On the other-hand, when we look at most of the existing artificial

intelligence-based recognition systems, they can only be trained in the batch settings. They require the training data, class information in advance. They are not scalable, and they do not have enough data security which makes them breach-able. Most of them are also unable to identify the important objects in an image which leads to poor recognition performance. Hence, there is an immense need to introduce such features in machine learning algorithms such that they become capable of learning, retaining and using that knowledge over a life time and are free from risk of security breaches.

Recently researches address this issue partially; one of the major objectives is to learn new classes incrementally without retraining the whole network. This is known as class incremental learning. There are three properties [1] of a class incremental system,

1. The classifier must be trainable from data stream in which examples of different classes arrives at different times.
2. It should be able to give competitive classification accuracy of the observed classes.
3. Its memory footprint and computational requirement should remain bounded or grow very slowly with respect to the number of classes observed so far.

A system satisfying these properties is said to be a class incremental system. These also serve as the road map to achieve the goal of lifelong learning. With the advancement in multi-agent systems, cloud computing [2] and cryptography [3] the goal of performing lifelong learning seems to be achievable in the near future. For a lifelong learning system involving image recognition, we believe the system must also:

1. Be able to dynamically allocate required computational resources.
2. Be able to detect objects of interest in an image in a human way.
3. Be able to store image and related data securely by using state of the art cryptography techniques.

Despite the vast progress in the area of the image classification [references] there is no single solution which satisfies all the criterions for lifelong learning. In this work we have extended iCaRL [1] which is a recent training strategy that allows learning in a class incremental way. Their classifier satisfies the criteria for class incremental learning but does not qualify for lifelong learning. We purpose three components in combination with iCaRL that are capable of lifelong learning:

1. Using machine learning to predict the computational resources required for training as number of classes increases continuously and choosing the optimal scaling plan.
2. Detecting the object of interest in an unsupervised manner by combining fast attention network (AttentionNet) with convolution neural network (CNN).
3. Storing the image data especially exemplars securely using interplanetary file system and block chain.

We explain the details of these steps in Sect. 2 and subsequently put them into the context of previous work in Sect. 3. In Sect. 4 we have compared the results of our purposed solution with the existing iCaRL method. We have also presented the results of SLA (service level agreement violations) by comparing the optimal scaling plan with fixed resources.

Algorithm 1: Classification Using iCaRL

```

Input image to be classified:  $x$  ;
Require class exemplar sets:  $\rho = (\rho_1, \dots, \rho_t)$  ;
Require feature map:  $\varphi: X \rightarrow R^d$  ;
FOR  $y$  in 1 to  $t$  DO:
     $\mu_y = \frac{1}{|\rho_y|} \sum_{p \in \rho_y} \varphi(p)$  // mean of exemplar images
ENDFOR
 $y^* \leftarrow \operatorname{argmin}_{y=1, \dots, t} \|\varphi(x) - \mu_y\|$  // nearest prototype
output class label  $y^*$ 

```

2 Method

This section describes the proposed methodology. We discuss the main components of our solution and explain how their combination helps us to achieve lifelong learning up to some extent. It also the criteria for lifelong learning that we proposed earlier. We adopt a multiagent-based approach where the system activities are carried out by different agents that are in constant communication with one another. Each agent is responsible for its own tasks and informs the next agent after each dependent task is completed. The agent-based approach is discussed below.

2.1 Class Incremental Learning Agent

This agent learns classifiers and feature representation from incoming data stream simultaneously. For classification, it relies on a set of exemplar images which it selects dynamically from the data stream. For each observed class there is one exemplar set. Algorithm 1 shows the mean of exemplar classes which is used to classify images into a set of observed classes. The nearest mean of exemplar is used for predicting a label y of new incoming image x .

At first it computes prototype vectors $u_1, u_2, u_3, \dots, u_n$ of all observed classes which is the average feature vector of all exemplars for each class by using the following equation,

$$\mu_y = \frac{1}{|\rho_y|} \sum_{p \in \rho_y} \varphi(p) \quad (1)$$

Then it computes the feature vector of a new incoming image x and assigns it the label with the most similar prototype.

$$x^* = \operatorname{argmin}_{y=1, \dots, t} \|\varphi(x) - \mu_y\| \quad (2)$$

For training, it processes batches of classes using an incremental strategy. Whenever data of new class is available it updates the network parameters and exemplars. Under the hood it uses CNN which was used only for representational learning and not for actual classification of incoming data. For classification it uses nearest-mean-exemplar strategy as mentioned above which computes the feature vector of incoming data and assigns it with nearest prototype.

iCaRL: Architecture and its Drawbacks: Our class incremental agent uses incremental classifier iCaRL. It uses CNN [4–6] such as ResNet or Desnet. They illustrate the network as a trainable feature extractor and use it to extract feature followed by a single layer of classification with as many sigmoid as total numbers of classes observed by the network. They normalize the feature vector and also its result. The resultant network output for any class is given by the following equation.

$$g_y(x) = \frac{1}{1 + \exp(-a_y(x))} \text{ with } a_y(x) = w_y^T \varphi(x) \quad (3)$$

The network architecture of ResNet, Desnet fails to perform like human perception process. As we know from our daily life that our visual attention process has the ability to perceive objects. There are many types of perception; the most common being recognizing, organizing and interpreting. One of the most common type of perception is visual perception. It interprets the world around us and forms a mental representation of the environment. Human mind attention and focus play a vital role in the mental representation formed by us. This is because neural patterns for those things which we see and pay attention to are stronger as compared to those which we see without paying attention to. This is also one of the major reasons that we retain those things longer on which we focus.

On the other hand, when we look at the general convolutional neural networks they do not focus on the important object in an image. Some of the networks which have this capability are very deep, which makes them less preferable for the class incremental learning. We also believe that this is an important aspect for lifelong learning systems as it must have some intelligence to pay more attention to the important things. This makes them similar up to some extent with the human mind which has good capacity of remembering important things instead of everything. It will also result in the improvement of overall accuracy.

Proposed Architecture Using Attention Network: AttentionNet is a deep convolutional neural network purposed by Yoo et al. [7]. It casts an object as an iterative classification problem. It provides quantized weak directions and ensemble of iterative predictions helping in convergence to accurate object boundary. In the following paragraphs we have explained its working, followed by its network architecture and finally its integration with the existing architecture.

Initially the input image is wrapped in a fixed CNN and fed into an attention network. It outputs two directional images comparable to the Top Left corner and Bottom Right corner of the input image. For example, the predictions which is possible in case of TL are the following: right, right-down, down, stop and no instance in this image F. Consider that the prediction gives us translating left or bottom right. After

that, image is cropped to a corresponding directional vector and then fed again into the network until the following criteria is met. F or ■ in both directions. When the network returns F at both corners the test ends with no target image found. On the other hand, when the network returns ■ at both of the end the present image is the finalized result.

In Fig. 1, we have illustrated AttentionNet architecture. At first there is pooling and normalization layer. After that the next seven layers are fully connected layer as described in VGG-M network by Chatfield et al. [8]. The stride and filter size are smaller at the first layer as compared to the stride at the second convolutional layer. The detection framework requires two predictions, but we prefer a CNN for classification to be trained by soft-max. Because the regression using a mean square force the model into a difficult bounding box. On the other hand, classification with soft-max forces the model to be activated at true directions rather than exact direction. The resultant output layer is separated to use softmax with each branch of prediction. These output layers are denoted by Conv8-TL and Conv8-BR. Finally, ReLU layer is used to get the detection.

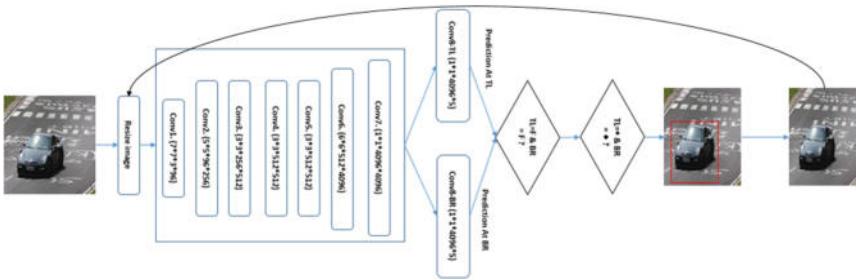


Fig. 1. Attention network

Figure 2 portrays the working of convolutional neural network with Attention Network. In our architecture, the Attention Net is applied after the first convolutional layer. At first the output of first convolutional layer is fed into an Attention Network which finds an object of interest in an image in an unsupervised way unlike other methods which require some information for drawing the bounding box. The output of the attention network is then given to next layer of the convolutional neural network. In this way network is more focused on object of interest and extract more meaning full features resulting in a better representation of classes.

2.2 Data Security Agent

This agent is responsible for storing and retrieving data securely. We believe that data security is the one of the vital issue for lifelong learning as large organizations especially in the health care area are concerned about it [9, 10]. For this reason, symmetric and asymmetric key cryptography have been widely used in the past. Recent advances in the area of block chain technology [11] and the inviolable level of security provided

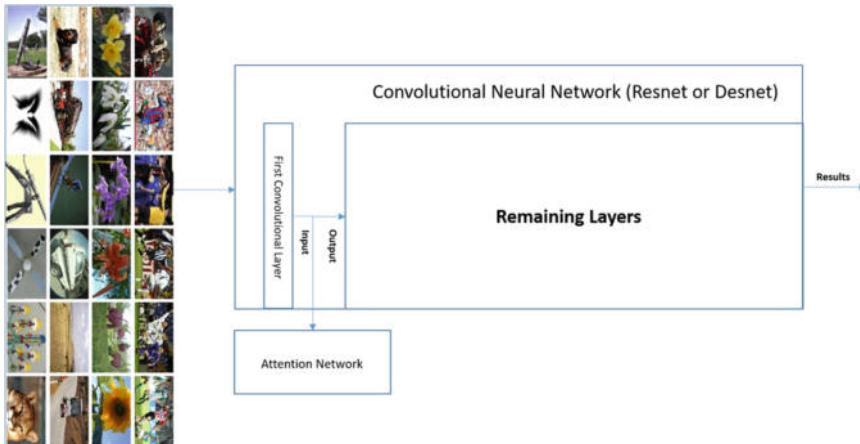


Fig. 2. Convolutional network with attention network

by it makes them the popular choice for most of the problems. However, from technological perspective, block chain is not without its wart.

Current proof of work has slowed the speed up to a crippling level especially for storing large data files like images. As we discussed above that in iCaRL, exemplar images need to be stored for each class. We have proposed a solution using interplanetary file system [12] and block chain technology for enhanced security. The encryption process is demonstrated in Fig. 3. At first, we have used interplanetary file system and asymmetric encryption [13] to store all the exemplar images. In this way we can store third party images securely, the third party will take the public key of service provider and encrypt the exemplar images with it. Then the encrypted exemplar images are added to the interplanetary file system. Now the service provider can decrypt the exemplar data using its private key whenever the new class data is observed. In this way malicious party cannot decrypt the file because they do not have the private key.



Fig. 3. Encrypting exemplar images using public key and IPFS

Encrypting the exemplar images increases the security of data, but it still does not meet the state-of-the-art criteria. In order to achieve state of the art criteria, we have coupled the block chain with the interplanetary file system. Figure 4 demonstrates our proposed scheme. Instead of encrypting the whole class images, only the resultant hash of each class provided by interplanetary file system is encrypted. Whenever the data of new class is received, at first it gets encrypted with the public key of the receiver and then the new block gets created which encrypts the resultant hash from interplanetary file system. In this way the best solution is created as it can maintain the simplicity of data required for the block chain and at the same time provides the benefit of peer to peer decentralized file system.

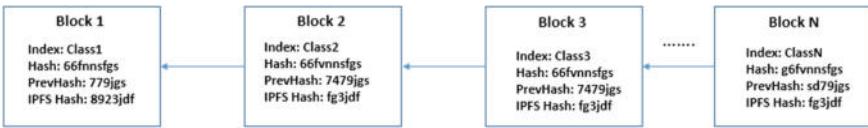


Fig. 4. Encryption of IPFS hash using block chain

2.3 Monitoring Agent

This agent runs on each cloud instance. It starts at boot time when new instance is launched and stores its apache access log and monitoring data in small local database. The agent monitors CPU utilization, memory utilization, network utilization. The stored data of each instance is then further used to classify VM instances. It sends this data to monitoring agent every minute or based upon the time interval provided by monitoring agent using jade message API. For better efficiency, the Instance profiling agent automatically deletes data from database which is older than 1 week.

In order to compute instance memory throughput as the number of bytes required to process a request (bytes/request); and instance CPU throughput as the number of clocks required to process specific request (clocks/request); an instance profiling agent smoothens out and cleans data to remove those tuples which are generated by hotspots and sudden spikes. These sudden changes in workload affect the accuracy of profiling process, therefore we have extracted smoothed 75th and 25th percentile below the SLO in correspondence with SLO threshold and 75th percentile of memory, CPU usage, network in and network out data.

The behavior of the resources required depends upon the type of dataset and classification problem. As a result, it affects the accuracy of CPU instances and memory throughput. For this reason, we re-compute instance CPU and memory throughput after every 15 min because VM churn time of most of cloud vendors is approximately 15 min.

In (4), ‘MemoryUsageInst’ represents the average percent of memory usage, whereas ‘CPUUsageInst’ represents the average percentage of CPU usage during the last time interval. Initially, we have assumed that all incoming requests require the same processing time, because we take into account heterogeneity of requests in planning agent when we compute the capacity of instance.

$$MemoryUsageInst = \frac{\sum_{i=1}^{i=n} (memoryusagedata)}{N} \quad (4)$$

$$CPUUsageInst = \frac{\sum_{i=1}^{i=n} (cpuusagedata)}{N} \quad (5)$$

$$InstanceMemoryUsageThroughput = \frac{\sum_{i=1}^{i=n} \frac{MemoryUsageInst}{100}}{TotalClasses} \quad (6)$$

$$InstanceCPUUsageThroughput = \frac{\sum_{i=1}^{i=n} \frac{CPUUsageInst}{100}}{TotalClasses} \quad (7)$$

In the above equations ‘TotalClasses’ is the sum of the classes observed by the algorithm so far. The computational resource of a class is defined by length of feature extractor parameter, exemplar images and weight vector when a class is observed. However, these equations can only compute results when the system is stable whereas most of cloud-based applications are heterogeneous in nature. For this reason, we gathered total number of CPU and Memory usage during the time interval. Then, we compute the expected CPU and Memory usage by multiplying current classes with their respective ideal throughputs. Finally, we divide the real resource usage with their expected usage.

$$CPUUsageExpected = TotalClasses_{inst} * InstanceCPUUsageThroughput \quad (8)$$

$$CPUWorkloadComplex = \frac{\sum_{i=1}^{i=n} \frac{PercentCPUUsage}{100}}{CPUUsageExpected} \quad (9)$$

$$MemoryUsageExpected = TotalClasses_{inst} * InstanceMemoryUsageThroughput \quad (10)$$

$$MemoryWorkloadComplex = \frac{\sum_{i=1}^{i=n} \frac{PercentMemoryUsage}{100}}{MemoryUsageExpected} \quad (11)$$

2.4 Prediction Agent

Prediction Agent forecasts the incoming classes for training of next forecasted window. The length of the forecasted window is 15 min because the VM churn time of most of the cloud environments is approximately within this range. We have used recurrent kernel extreme learning machine (RKELM) for this purpose. Our agent uses a ticker behavior which repeats itself every 15 min and predicts future resource requirement of next forecasted window. It predicts the incoming class of next window. This agent also stores the forecasted data of next monitoring window in database. It then sends forecasted data to planning agent.

Prediction is vital for determining accurate resource requirement in case of lifelong learning. For this reason we have performed various experiments to determine the best method from state of the art techniques. We have selected extreme learning machine based methods for prediction as they are faster and more accurate as compared to previous methods such as SVM [14, 15]. Experimental results shows that RKELM learning technique performs better as compared to ELM and RELM. In the following paragraphs we have described the working of ELM, RELM and RKELM and their performances.

Extreme Learning Machine: An Extreme Learning Machine (ELM) is a single layer feedforward network (SLFN) which was encouraged by biological learning and solved the issues faced by back propagation based learning algorithms. The hidden nodes of an ELM computes random combination of input values. Thus the input to the hidden weights need not to be computed by computationally extensive algorithms and can be randomly generated independently from the input data. Figure 5 demonstrates the architecture of ELM.

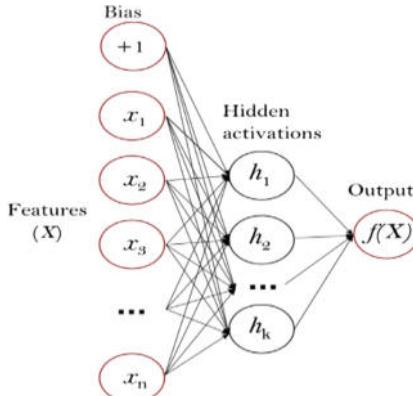


Fig. 5. Learning of ELM

Researchers have proven that universal approximation holds under mild assumption, provided that the hidden layer has enough hidden nodes. The final output is finding the optimal hidden to output weights which can be done in the analytical way without the need of computationally expensive iterative processes, such as back propagation [16]. More formally, let x be a d -dimensional row feature vector belonging to one out of m possible classes. Such a feature vector corresponds a class label y represented by an m -dimensional unit row vector. Its single positive c^{th} component, denoted as y_c , indicates that x belongs to class $c \in C = \{1, \dots, m\}$. The vector x is the input for ELM. Thus, the value of the j^{th} input neuron corresponds to the j^{th} component in a data sample x . Let the hidden layer be composed of L hidden neurons and let its connection with the input neurons is denoted by the random matrix $R \in R_{d \times L}$. The output of the j^{th} neuron in the hidden layer is computed as,

$$h_{j(x)} = \sigma(x, R_j, b_j) \quad (12)$$

where R_j is the j th column of R , b_j is a random bias parameter associated with the j th hidden neuron and $\sigma(\cdot)$ is a differentiable activation function, such as the Sigmoid. Let $\beta \in R_{L \times m}$ be the matrix of weights connecting the hidden layer composed of L nodes with the m output nodes and the output row vector of the hidden layer

$$h(x) = [h_{1(x)}, \dots, h_{L(x)}] \quad (13)$$

with respect to the input x . The output of the network is then defined as:

$$\hat{y} = h(x)\beta \quad (14)$$

The function $h(x)$ maps the data from the d -dimensional input space to the L -dimensional hidden layer random feature space where the input-to-hidden node weights R are randomly generated according to any continuous sampling probability distribution.

Let $\{(x_{(i)}, y_{(i)})\} i = 1, \dots, n$ be the set of n training samples pairs. Since the input-to-hidden weights are randomly generated, they do not require tuning, and the learning algorithm consists in finding a proper hidden-to-output weight matrix β such that

$$H\beta = Y \quad (15)$$

In general, H is not a square matrix this is the reason an exact solution may not exist. An approximate solution however can be found by solving the following minimization problem:

$$\arg \min \hat{\beta} ||H\hat{\beta} - Y|| \quad (16)$$

Which is a standard least-squares problem whose solution can be found by using the orthogonal projection method.

Recurrent Extreme Learning Machine: The recurrent extreme learning machine is a feedback network. It takes the predicted output of previous step and pass as an input for prediction of the next step. The recurrent extreme learning machine approach for training is defined as follows [17].

$$\hat{y}_p = H_p \hat{\beta}_p \quad (17)$$

where \hat{y}_p denotes the predicted value and output matrix is calculated by using the following equation.

$$H_p = \begin{bmatrix} g(w_{1,1}S_{1,1} + b_1) & \cdots & g(w_{L,z+p}y^{-p+1} + b_{z+p-1}) \\ \vdots & \ddots & \vdots \\ g(w_{L,1}S_{L,1} + b_1) & \cdots & g(w_{L,z+p}y^{-p+1} + b_{z+p-1}) \end{bmatrix} \quad (18)$$

where $g(\cdot)$ is the activation function, input weight (w) and biases (b) are randomly assigned, z is the number of hidden nodes, p is the number of prediction step $p = (1, 2, \dots, P)$, y^{-p} shows the previous p -th value of the output. As the first prediction value does not have any prediction result yet. For this reason, the above equation can be defined by a Moore-Penrose generalized method as follows.

$$\hat{\beta}_p = H_p^\dagger Y_p \quad (19)$$

where, H_p^\dagger is generalized inverse matrix of H_p . Therefore, the prediction can be calculated by Eq. (19) in each step. Figure 6 shows the architecture of recurrent extreme learning machine.

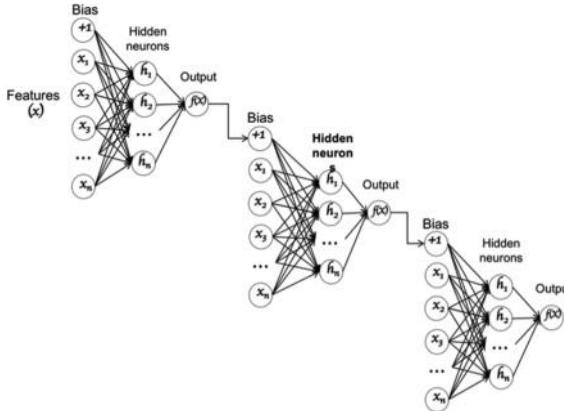


Fig. 6. Architecture of recurrent ELM

Recurrent Kernel Extreme Learning Machine: Recurrent Extreme Learning Machine has several limitations. The inputs and biases need to be assigned randomly. Furthermore, the number of hidden nodes plays an important role in prediction accuracy. In order to resolve these limitations kernel matrix is created by applying kernel method on the training input data and then replacing the hidden layer output matrix of RELM [17].

$$K = k(S, S) \quad (20)$$

In the above equation K represents the kernel matrix of input data. The output weight of Recurrent Kernel Extreme Learning Machine is computed by the following equation.

$$W_{out,p} = \left(K + \frac{1}{C^r} \right)^{-1} Y_p \quad (21)$$

In this $W_{out,p}$ represents an output weight in the p th step and Y_p denotes the target data. As compared to the RELM in RKELM number of hidden nodes is not specified and its output weights are not selected randomly. The predicted value is computed by using kernel matrix and output weight of the model. It is represented by the following equation.

$$\hat{y}_p = K^T W_{out,p} \quad (22)$$

For keeping the training window size same and reducing its computation, the training input data is generated in $p + 1$ step using the following equation.

$$S_{p+1} = \begin{bmatrix} X_{1,1+p} & \cdots & y_1^{-p} \\ \vdots & \ddots & \vdots \\ X_{L,1+p} & \cdots & y_L^{-p} \end{bmatrix} \quad (23)$$

We can use any kernel function as far as it satisfies the Mercer's condition. Most common function is RBF. In our case we have used the RBF kernel. Figure 7 represents the working of recurrent kernel elm.

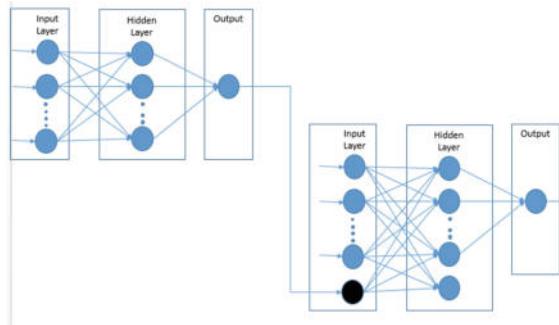


Fig. 7. Architecture of recurrent kernel ELM

2.5 Data Access Agent

Data access agent manages the database. It receives all the data from other agents and stores it to MYSQL database after processing. Any agent that wants some data for its processing sends a request to data access agent through jade API and it provides data after fetching the database. It maintains separate database tables for multiple types of information such as access logs, CPU utilization, future predictions, scaling plan details etc. It maintains and updates the database tables when new data is received. It sends request logs of previous time interval to prediction agent which after training forecasts number of requests for next time interval and sends this information back to data access agent. Data integration agent communicates with data access agent to get current

request rate and predicted request rate per minute and compares both to represent current picture of the system state. Data access agent maintains profile for every instance type which stores the maximum capacity of instance in terms of maximum CPU utilization and maximum number of requests it can process without exhausting. This information is used by planning agent to make appropriate scaling plans.

2.6 Planning Agent

The planning agent is one of the major component and is responsible for up-scaling of computational resources, choosing the best scaling plan and ranking VM instances. It computes the current capacity of the system based on the workload complex, and general ideal throughputs of CPU utilization, memory utilization, predicted workloads and cost of scaling plan.

Initially our framework monitors a default scaling plan. After 15 min, the prediction agent predicts new incoming classes and sends the prediction data to planning agent. Then, the planning agent processes the request and determines the total capacity by taking the sum of capacity of each running instance. Then, Algorithm 2 is used for predicting the maximum resources required for handling new incoming classes in next time window.

Algorithm 2: Scaling Cloud Resources

Input

- List of available instance types containing cost and capacity
- Cost array *costSum*

Output

- Scaling plan instance list *scalPlanInst*

```

FOR i in 1 to totalCapacity DO:
  FOR j in j to SizeOfAvailableInstances DO:
    Initialize maxCost to zero
    InstInfo instanceInfo = availableInstances.get(j);
    IF instanceInfo.capacity > i THEN
      continue
    ENDIF
    cost = instanceInfo.cost + costSum[i - instanceInfo.capacity]
    IF maxCost == 0 Then
      costSum[i] = cost
      instanceIndex[i] = j
      previousSumIndex = i - instanceInfo.capacity
      maxCost = cost
    ENDIF
  ENDFOR
ENDFOR
WHILE capacity > 0
  scalPlanInst.add (availableInstanceList.get (instIndex[i]))
  capacity = previousSumIndex[i]
ENDWHILE

```

2.7 VM Management Agent

The VM management agent executes the scaling plan and can do both horizontal and vertical scaling. In horizontal scaling, it adds new VM instances to the scaling plan and in vertical scaling it increases the capacity of currently running instances. This agent processes incoming data in its cyclic behaviour and based upon that, performs various operations such as shutting down of VM instances, increasing their capacity or launching new VM instances based on the information received by planning agent. We have used Amazon API [18] which hides most of the underlying details, making the implementation easier for developers and researchers.

3 Related Work

The idea to learn classifiers in an incremental way dates back to classical neural network literature. There are also several studies in the recent era related to incremental learning [19–22]. In this section we discuss some of the most relevant studies. One of the backbones of lifelong learning is an incremental classifier. We have divided the learning techniques with fixed representation and the techniques which can also learn data representation. We have discussed these techniques both from classical point of view, which includes leveraging linear classifiers, ensemble of weak classifiers, nearest neighbor classifier etc. and deep learning point of view. Deep learning can be further classified into two categories depending on whether they require old data or not.

When the data representation is fixed, the main problem is to design such an architecture which can learn new classes during the training of model without requiring the access to all the training data seen so far. One of the approaches which has this property is nearest class mean (NCM). In this method, class prototypes are represented by taking the mean of vectors and as a result the observed training data is not needed every time. It classifies new incoming training data by assigning it the nearest prototype [19, 23, 24]. The major problem of NCM is that it cannot be extended to non-linear data representation. There are some other approaches which can be used for class incremental learning. Kuzborskij et al. [20] has showed that when we train the linear multi-class classifier from a small amount of data, we can avoid the loss of accuracy. Royer and Lampert [25] proposed a system which can adapt classifiers to time varying data, but it cannot handle new incoming classes. Several researchers [26–28] aimed at distinct problem of open set image recognition in which test examples do not fall under the training classes. An ensemble-based approach is proposed by Polikar et al. It can handle the increasing number of classes, but it needs the data of all the old classes every time [29, 30]. Lampert et al. [31] has proposed zero short learning. Their solution can classify previously unseen classes, but it does not have the ability to perform training for those classes.

Deep learning networks have achieved a considerable success not only for their classifying ability but also for their ability to find suitable data representation [32–35]. As mentioned above, they can be categorized in two major categories. In the first category, the algorithm does not require old data. Jung et al. [36] has presented a method for domain transfer learning. The performance on old task is maintained by

freezing the final layer of previous tasks and preventing the shared parameters of weights in feature extraction layer from changing. Kirkpatrick et al. [37] has purposed an approach in which important parameters of old tasks are constrained to remain close to their original values when looking for a solution of a new incoming task. The drawback of this approach is that shared weights between old and new tasks will have conflicting constraint. Several authors [33, 38] have used knowledge distillation [38] for object detection and classification for maintaining the performance of new classifier over the old tasks. Some of the methods [39] are using data generated by GANS. Also, the data generated by GANS are not as good as actual data. The second category of data require part or whole data. The method purposed by Rebuffi [1] uses an exemplar images to select data when adding new classes. [40] has purposed a network which grows network hierarchically when new training data are added hierarchically.

Although several solutions are proposed but none of them satisfies the criteria of lifelong learning. From security concerns to scalability of server are not addressed by these mechanisms. There is lot of research gap which researchers need to address to achieve the ultimate goal of lifelong learning.

4 Experiments and Results

In this section we have discussed the protocol of experiments for lifelong learning. We have done several experiments in order to prove that our purposed solution satisfies the criteria of lifelong learning.

4.1 Benchmark Protocol

There is no agreed protocol for evaluation of lifelong learning and even there is no standard protocol for evaluating class incremental algorithms. So, we have purposed the following protocol for evaluating the lifelong learning. At first, we have evaluated prediction methods. After that we have evaluated our novel scaling plan algorithm and compared its performance in terms of SLA violations. Finally, we have evaluated our purposed version of iCaRL which uses Attention Network with the convolutional neural network for feature extraction.

4.2 Prediction Agent Evaluation

We have evaluated ELM, RELM and RKELM. The process of working of prediction agent are as follows.

1. Monitoring agent monitors continuously the resources of each instance, in the cloud environment.
2. The monitoring agent then sends the data to data access agent which is responsible for handing data operations.
3. Data Access Agent on the request of prediction agent sends the data of current time window.

4. Prediction agent predicts the incoming classes and computational resources of next time window by using the above-mentioned algorithms.

Data Collection in Cloud: We have gathered the data by running the incremental classifier. It is deployed on the amazon cloud with the objective that the incoming class data must be served instantly. Table 1 further describes the detail of the incoming class.

Table 1. Incoming class data

Variable	Description
Total number of samples	140
Benchmarking instances duration	140 min
Sampling interval	1 min
Sample resource	Monitoring agent monitors across all ec2 instances
Maximum number of instances	5
Minimum number of instances	1

In our experiment we have evaluated our prediction results using extreme learning machine, recurrent extreme learning machine and recurrent kernel extreme learning machine. The evaluation is done on these metrics Mean Absolute Error (MAE); Mean Absolute Percentage Error (MAPE); Root Mean Square Error (RMSE); Mean Square Error (MSE); Relative Absolute Error (RAE). The mean absolute error tells us how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by the following equation [41].

$$MAE = \frac{\sum_{i=1}^{i=n} |y_i - x_i|}{n} \quad (24)$$

The mean absolute percentage error for the prediction model is given by the following formula.

$$MAPE = \frac{100}{n} \sum_{t=1}^n \frac{|A_t - F_t|}{A_t} \quad (25)$$

A lower value of it indicate better fit for the prediction model thus indicating superior prediction accuracy [42]. Root mean square error or root mean square deviation [43, 44] is represented by the following equation.

$$RMSD = \sqrt{\frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{n}} \quad (26)$$

Mean square error is the average of the square error which is denoted by the following formula [45].

$$MSE = \frac{1}{n} \sum_{i=1}^{i=n} (y_i - y_{1i})^2 \quad (27)$$

Finally we have also computed relative absolute error which is denoted by the following formula.

$$E_i = \frac{\sum_{j=1}^{j=n} |p_{ij} - T_j|}{\sum_{j=1}^{j=n} |T_j - \bar{T}|} \quad (28)$$

Table 2 presents the comparison of prediction methods based on above mentioned metrics.

Table 2. Error comparison for new incoming classes

	ELM	RELM	RKELM
MAE	1.40	1.39	1.385
MAPE	0.188	0.1856	0.1851
RMSE	8.37	8.28	8.17
MSE	70.0569	68.55	66.74
RAE	0.186	0.1852	0.184

The prediction results in Table 2 show that recurrent kernel extreme learning machine performs better followed by recurrent extreme learning machine and extreme learning machine.

4.3 Evaluation of Scaling Plan

Scalability of resources is a very important factor in terms of lifelong learning systems. This is because new classes are always being observed which demands computational resources to be increased dynamically. Furthermore, if the exemplar images are continuously decreased then the classification accuracy will decrease so there is a need to maintain exemplar data as explained by Rebuffi et al. [1]. In our case we are taking the size of exemplar images from user. We have proposed a novel algorithm which finds the best optimal scaling plan in terms of both price and performance. For our purposes, we have used three different types of cloud instances,

- (1) T2.micro
- (2) T2.small
- (3) T2.medium.

For our scenario, we prefer general purpose instances instead of memory-oriented instances or CPU-oriented instances because general purpose instances have a balanced combination of memory and CPU. Specifications of these instances are given in Table 3. In future, we intend to evaluate our framework for other instance types such as memory-oriented instances and CPU oriented instances as well.

Table 3. Specification of virtual instances

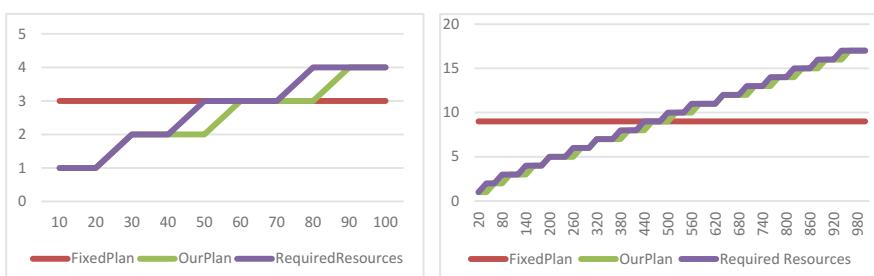
Instance type	vCPU	Memory (GB)	Price/hour
T2.micro	1	1	0.013\$
T2.small	1	2	0.026\$
T2.medium	2	4	0.052\$

In order to evaluate scaling plan, we have compared the service level agreement violations. In our case it is described as the total number of times our scaling plan failed to provide enough computational resources for training model of new classes and total number of times when computational resources are 20 percent more than the required resources. In our case for the iCIFAR-100 with 10 classes per batch we have given 1000 prototypes exemplar images and for ImageNet ILSVRC 2012 [46] 20 classes per batch. Now the objective of the scaling plan algorithm is to provide best scaling plan with enough computational resources and minimum cost. We have compared the SLA violations with the general scenario when fixed resources are defined in advance without the knowledge of incoming classes and exact number of classes. For fixed resources we are evaluating for two scenarios.

In the first case we are evaluating for the scenario when we allocate less fixed resources. When small fixed computational resources are allocated we have more “High Resources SLA violations” at the start for “FixedPlan” as compared to end. In the end of experiment “FixedPlan” has more “Low Resources SLA violations” because computational resources are not enough to handle increasing number of classes as compare to “OurPlan”. Table 4 and Fig. 8 presents the comparison of SLA violations with small fixed resources.

Table 4. Comparison of SLA violation with small fixed resources

	LR SLA VIO (CIFAR-100)	HR SLA VIO (CIFAR-100)	LR SLA VIO (ImageNet ILSVRC 2012)	HR SLA VIO (ImageNet ILSVRC 2012)
OurPlan	1	1	8	0
FixedPlan	3	4	24	23

**Fig. 8.** Comparison of instances required VS a FixedPlan (LR) versus OurPlan (CIFAR-100), b FixedPlan (LR) versus OurPlan (ImageNet ILSVRC 2012)

In the second case we are evaluating for the scenario when we are allocating large number of fixed resources. We have more “High Resources SLA violations” at the start for “FixedPlan”. The reason behind is that large number of computational resources will not be used at the start when the smaller number of classes is observed by the model. At the end of experiment “High Resources SLA violations” decreases because the computational resources required by model are more. As compared to that “FixedPlan” has a smaller number of SLA violations both at the start and end of the experiment. Table 5 and Fig. 9 summarize these results.

Table 5. Comparison of SLA violation with large fixed resources

	LR SLA VIO (CIFAR-100)	HR SLA VIO (CIFAR-100)	LR SLA VIO (ImageNet ILSVRC 2012)	HR SLA VIO (ImageNet ILSVRC 2012)
OurPlan	1	1	12	0
FixedPlan	0	10	0	50

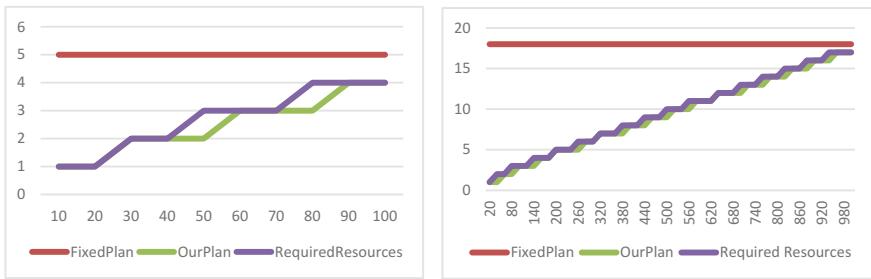


Fig. 9. Comparison of instances required VS a FixedPlan (HR) versus OurPlan (CIFAR-100), b FixedPlan (HR) versus OurPlan (ImageNet ILSVRC 2012)

4.4 Performance Evaluation with Attention Network

We have described above that class incremental classifier is using convolutional neural network. But the convolutional neural network extracts general features from images which also included unnecessary objects in the background. For this purpose, we have combined it with attention network [47] which has the ability to focus on important object in the foreground in an un-supervised way. This results in improved accuracy. Till now there is no agreed protocol which exists for evaluation of class incremental learning. For this reason, we have used the following evaluation process of a given multiclass classification dataset. We have arranged the classes in the random order. At first, we have trained the method in the class incremental way. After that evaluation is performed for those classes on which we have trained the classifier. We have made sure that no overfitting can occur by not revealing the testing results to the model. Currently we have reported the average of these accuracies called average incremental accuracy.

For CIFAR-100 data we have trained classes in the batches of 5, 10, 20 and 50 classes at a time. The evaluation measure is standard multiclass accuracy on the test set. Table 6 summarizes the results.

Table 6. Comparison of classification results

Batch size	iCaRL (%)	iCaRL with attention network (%)
5 classes	61.2	62.8
10 classes	64.1	66
20 classes	67.2	69.2
50 classes	68.6	70.65

5 Conclusion

Life Long learning is the major problem on the road for the development of these systems. There are certain criterions which needs to be satisfied for the development of it. These include criterions for class incremental learning, data security, dynamic allocation of resources and focusing on the object of interest. For this reason, we have proposed a secure scalable lifelong learning system based on multi-agent framework using cloud computing. Our proposed framework uses state of the art method for data security using block chain, prediction of future data training requests, dynamically allocating cloud resources for training of the model and finally focusing on the object of interest. Our experiments show that the proposed system has better accuracy and it also satisfies the criteria for lifelong learning. In future we will perform extensive experiments on several other datasets. We will also dynamically allocate resources when large number of new classes is observed by a model creating a surge.

Acknowledgements. We would like to extend our acknowledgements to the *UM Grand Challenge Project ICT* Project No *GC003A-14HTM* for funding this project.

References

1. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: incremental classifier and representation learning. In: Computer Vision and Pattern Recognition (2017)
2. Amazon: Cloud computing. [Online]. Available: <https://aws.amazon.com/what-is-cloud-computing/>
3. Wikipedia: Cryptography. [Online]. Available: <https://en.wikipedia.org/wiki/Cryptography>
4. Rawat, W., Wang, Z.: Deep convolutional neural networks for image classification: a comprehensive review. MIT Press Journals (2017)
5. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroud, A., Shuai, B., Liu, T., Wang, X., Wang, L., Wang, G., Cai, J., Chen, T.: Recent advances in convolutional neural networks. In: Computer Vision and Pattern Recognition (2015)
6. Convolutional neural networks for visual recognition. [Online]. Available: <http://cs231n.github.io/convolutional-networks/>

7. Yoo, D., Park, S., Lee, J.Y., Paek, A.S., So Kweon, I.: AttentionNet: aggregating weak directions for accurate object detection. In: Computer Vision and Pattern Recognition (2015)
8. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In Proceedings of British Machine Vision Conference (BMVC) (2014)
9. Kester, Q.A., Nana, L., Pascu, A.C., Gire, S., Eghan, J.M., Quaynor, N.N.: A cryptographic technique for security of medical images in health information systems. *Procedia Comput. Sci.* **58**, 538–543 (2015)
10. Taitsman, J.K., Grimm, C.M., Agrawal, S.: Protecting patient privacy and data security. In: Perspective, pp. 977–979 (2013)
11. The great chain of being sure about things, 31 Oct 2015. [Online]. Available: <https://www.economist.com/briefing/2015/10/31/the-great-chain-of-being-sure-about-things>. Accessed 2017
12. Benet, J.: IPFS—content addressed, versioned, P2P file system
13. Rouse, M.: Asymmetric cryptography. [Online]. Available: <https://searchsecurity.techtarget.com/definition/asymmetric-cryptography>
14. Liu, X., Gao, C., Li, P.: A comparative analysis of support vector machines and extreme learning machines. *Neural Netw.* **33**, 58–66 (2012)
15. Bucurica, M., Dogaru, R., Dogaru, I.: A comparison of extreme learning machine and support vector machine classifiers. In IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania (2015)
16. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. *Neurocomputing* **70**(1), 489–501 (2006)
17. Liu, Z., Loo, C.K., Masuyama, N., Pasupa, K.: Multiple steps time series prediction by a novel recurrent kernel extreme learning machine approach. In International Conference on Information Technology and Electrical Engineering (2017)
18. Ans, B., Rousset, S.: Avoiding catastrophic forgetting by coupling two reverberating neural networks. *C. R. Acad. Sci.* **320**(12), 989–997 (1997)
19. French, R.M.: Catastrophic interference in connectionist networks: can it be predicted, can it be prevented? In: Conference on Neural Information Processing Systems (NIPS) (1993)
20. French, R.M.: Catastrophic forgetting in connectionist networks. *Trends in Cogn. Sci.* **3**(4), 128–135 (1999)
21. Robins, A.V.: Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.* **7**(2), 123–146 (1995)
22. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Distance-based image classification: generalizing to new classes at near-zero cost. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(11), 2624–2637 (2013)
23. Kuzborskij, I., Orabona, F., Caputo, B.: From N to N+1: multiclass transfer incremental learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3358–3365 (2013)
24. Polikar, R., Upda, L., Upda, S.S., Honavar, V.: Learn++: an incremental learning algorithm for supervised neural networks. *IEEE Trans. Syst.* **31**(4), 497–508 (2001)
25. Cauwenberghs, G., Poggio, T.: Incremental and decremental support vector machine learning. In: Proceedings of the 13th International Conference on Neural Information Processing Systems, Denver, CO (2000)
26. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Metric learning for large scale image classification: generalizing to new classes at near-zero cost. In: European Conference on Computer Vision (ECCV) (2012)
27. Ristin, M., Guillaumin, M., Gall, J., Van Gool, L.: Incremental learning of NCM forests for large-scale image classification. In: Conference on Computer Vision and Pattern (2014)

28. Royer, A., Lampert, C.H.: Classifier adaptation at prediction time. In: Conference on Computer Vision and Pattern (2015)
29. Li, F., Wechsler, H.: Open set face recognition using transduction. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)* (2005)
30. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Towards open set recognition. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)* **36** (2013)
31. Bendale, A., Boult, T.: Towards open world recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
32. Muhlbaier, M.D., Topalis, A., Polikar, R.: Learn++.NC: combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes. *IEEE Trans. Neural Netw.* **20**(1) (2009)
33. Polikar, R., Upda, L., Upda, S.S., Honavar, V.: Learn++: an incremental learning algorithm for supervised neural networks. *IEEE Trans. Syst. Man Cybern.* **31**(4) (2001)
34. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell. (T-PAMI)* (2013)
35. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* (2013)
36. Li, Z., Hoiem, D.: Learning without forgetting. In: Computer Vision and Pattern Recognition (2017)
37. Misra, I., Shrivastava, A., Gupta, A., Hebert, M.: Crossstitch networks for multi-task learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
38. Saxena, S., Verbeek, J.: Convolutional neural fabrics. In: Conference on Neural Information Processing Systems (2016)
39. Jung, H., Ju, J., Jung, M., Kim, J.: Less-forgetting learning in deep neural networks. Learning (2016)
40. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Learning (2016)
41. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. Machine Learning (2015)
42. Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Zhang, Z., Fu, Y.: Incremental classifier learning with generative adversarial networks. In: Computer Vision and Pattern Recognition (2018)
43. Xiao, T., Zhang, J., Yang, K., Peng, Y., Zhang, Z.: Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In: ACM Multimedia (2014)
44. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient based learning applied to document recognition. In: Proceedings of the IEEE (1998)
45. Krizhevsky, A.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
46. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis. (IJCV)* **115** (2015)
47. Mnih, V., Heess, N., Graves, A.: Recurrent models of visual attention. In: NIPS (2014)



An Autonomic Model-Driven Architecture to Support Runtime Adaptation in Swarm Behavior

Mark Allison¹(✉), Melvin Robinson², and Grant Rusin¹

¹ University of Michigan - Flint, Flint, MI 48502, USA
{markalli,grusin}@umich.edu

² University of Texas - Tyler, Tyler, TX 75799, USA
mrobinson@uttyler.edu

Abstract. The use of unmanned vehicles in swarms requires significant runtime adaptation within the managing software due to the unpredictability of the environment it operates. This is compounded by rapid context changes occurring within the software elevating its operational complexity to a magnitude that renders them infeasible for humans to effectively manage. Our approach to addressing this challenge is model-driven self-adaptation using autonomic methods. This work extends and refines ongoing work on an unmanned vehicle swarm platform based on probabilistic finite state machines as behavioral runtime models and the formation of subswarms in the context of communication constrained search. We present the architecture of our work in progress as a reflection mechanism controlling short and long-term adaptive behavior. We realize short-term behavior change by the continuous transformation of structural models at runtime. To validate the architecture's autonomic properties, we provide a walkthrough of an indicative scenario pertaining to swarm resilience as proof of principle of the architecture's ability to dynamically replan under element failure.

Keywords: Autonomic computing · Swarm technology · Model-driven architecture

1 Introduction

Social animals in swarms exhibit superior capabilities to accomplish tasks which may be impractical for those working in isolation [14]. Researchers have referenced these naturally occurring models to build systems with the potential of decomposing complex tasks and executing them in a massively parallel manner. The collaboration and communication required of these systems are nontrivial as the complexity involved in their macro-level controls becomes more problematic as swarm size increases.

The challenge of each element in a swarm is threefold; they are required to self-adapt their behavior in response to uncontrollable environmental conditions,

maintain awareness and control of internal states, and possess some aspect of the global swarm knowledge [4]. This work is a part of our ongoing investigation into a model-centric autonomous swarm system. In Allison et al. [1], we presented a swarm framework based on a probabilistic finite state machine model (PFSM) in the context of communication constrained search. This platform made novel use of swarm subsets, subswarms in the performance of tasks. The PFSM determined short-term (in mission) element behavior adaptation while longer term (mission to mission) adaptation was accomplished by a genetic algorithm which propagates desired behavior.

Implementing metaheuristic solutions, as in this case, has traditionally been avoided by software engineering research and practice [9]. The stochastic nature of the approach comes with a set of complexity challenges stemming from dynamically changing constraints and requirements. Required adaptations cannot be anticipated *a priori* [6].

In simulating our framework, several challenges emerged. Predominantly, although the swarm elements could adapt their behavior based on context, the macro swarm behavior was relatively immutable during missions. A case in point, there were no allowance for element failure. While we utilized an approach to long-term adaptation based on parametric tuning of the PFSM after each mission using a genetic algorithm to evolve desired behavior, this did not directly impact the resilience and performance at runtime. This paper extends the previous work to introduce some robustness and increased short-term self-adaptation through autonomic means. For the purposes of our discourse we adopt the definition of a system as being self adaptive if it is capable of changing its control data at runtime, as put forward in Bruni et al. [2].

Autonomic computing is a strategy to address uncertainty by using technology to manage technology [5]. Incidentally, like other strategies applied in this work, Autonomic properties are inspired by the autonomic nervous system which allows biological organisms to function without active or conscious management of intrinsic systems as in the case of our internal temperature self regulation. Autonomic systems propose properties of self configuration, optimization, healing, and protection. To achieve an autonomic overlay for a distributed cyber-physical system, such as in the case of Autonomous Vehicle (AV) in swarms, there is a prerequisite for high level policy constructs and feedback control loops using *touchpoints* in critical points in the architecture. Our approach utilizes a model-driven approach to autonomic management relying on causally connected structural runtime models to represent swarm elements.

Model-Driven Engineering utilizes software models as first class artifacts [15]. It allows for complex systems to be conceptualized, communicated and operated at a higher level of abstraction than with high level language code [7]. In our approach, we use a model-driven architecture to specify complex behavior through the automated generation and transformation of models. More succinctly, our approach utilizes runtime models to realize autonomic adaptation.

Runtime models are a rapidly emerging subfield of Model-Driven Engineering used to support dynamic adaptation of complex software systems [7].

The behavior of each swarm element is specified by state machines models, while their interaction and larger concerns of the swarm is captured and managed through a causally connected refection and adaptation models. We propose that these models used in the architecture provide a relatively uncomplicated means to address short-term adaptation during a mission at the macro-level. Some examples of identified swarm concerns are dynamically reassigning tasks and re-planning based on element loss or malfunction, selecting a search strategy based on context, and securing and assuring intra-swarm communication. Element concern examples are collision avoidance, target detection/recognition, and power consumption rates. We propose the architecture as flexible enough for both contexts. Our exploration presents a distributed architecture whereby each UV's behavior is dictated by the execution semantics of a runtime model. Each UV communicates via a pub/sub broker based communications subsystem to an orchestrating autonomic manager.

The underlying thesis of this paper queries the feasibility and utility of an autonomic approach to models at runtime for swarm behavior in the context of communication constrained search. This work replaces the previous hardcoded framework with a more modular approach based on high level abstractions. Specifically, the contribution of this paper is a *Monitor, Analyze, Plan, Execute* computations over a *Knowledgebase* (MAPE-K) control loop [10] based model-centric architecture for orchestrating change at runtime.

The remainder of this paper is organized as follows: In the next section, we provide a review of the concepts and previous work which supports our approach as background. Section 3 provides the architecture in terms of its syntactic and semantic elements. We subsequently provide a walkthrough of the architecture by an indicative failure scenario as proof of principle in Sect. 4. Section 5 concludes and provides the next steps in this research path.

2 Background

In this section, we provide the backdrop of this approach with a discussion of autonomic principles wrt. model-driven architectures then offer a cursory review of our prior work to situate the approach.

2.1 Autonomic Self Adaptation

Increase complexity in software systems drives the need for autonomous adaptability [15]. Self-adaptation comes with not only a heavy computational overhead. One approach to realizing self-adaptive behavior put forward by IBM researchers is autonomic computing [12]. The architectural blueprint for autonomic computing [5] proposes that we address complexity growth by using technology to manage technology. A core concept of the autonomic principle is the MAPE-K control loop. The managed system assumes operation within a non-deterministic environment and is controlled via *effectors* and monitored via *sensors*. The system that manages uses four macro-computations: Monitor, Analyze,

Plan and Execute to effect changes utilizing a representation of knowledge about the controlled system.

2.2 Runtime Models

The manner in which computations are reified and knowledge is represented becomes the challenge. Wätzoldt and Giese [17] proposed a model classification for distributed self* systems and views a MAPE-K adaptation engine semantics as transformations on a set of models which represent the different concerns of a managed system. We utilize this taxonomy to describe the artifacts of the architecture. While we are in line with this taxonomy we have distilled the differentiations to five key model types for our particular syntactical representation of knowledge, namely:

- *Reflection Models* - A directly causally connected representation of the distributed system wrt. its elements and context.
- *Runtime Adaptation Models* - A container abstraction which constitutes the solution domain of architecture and the scope of the variation. These models are necessarily either *Change Models* - a representation of how the system's variations are managed, or *Evaluation Models* - which should capture the functional and nonfunctional properties of the managed system.
- *Monitoring Models* - represents how the reflection model is observed, essentially describing the anomalies of the system. These models contain the sensory touchpoints. Touchpoints are interfaces which expose access to a systems state information and control of operations [5].
- *Execution Models* - Represents how adaptations are effected within the controlled system. These models contain the effector touchpoints.

Figure 1 provides an overview of how these models interact to derive operational semantics. In our approach, all these models are constrained by metamodels (model of models) and are generated at runtime. These models are structural and syntactic in nature. The swarm architecture derives its operational semantics using a combination of the element's state machines and MAPE model transformations.

2.3 Swarm Elements Behavioral Models Overview

We now summarize our prior work and the swarm concept to contextualize the contribution of this work increment. Our swarm consists of two primary elements, **Autonomous Vehicles** (AV) which may be heterogeneous in performance and sensory capacity, and **Bases** which support them in terms of power and the communication of global knowledge. A key concept of our swarm is the ability for AVs to create subswarms to extend communications range, verify findings, and merge sensors to form virtual elements called *superagents*. AVs in our model constantly recalculate their ability to assist in tasks (utility) and their Point of No Return (PONR); both are highly dependent on power reserves. The manner

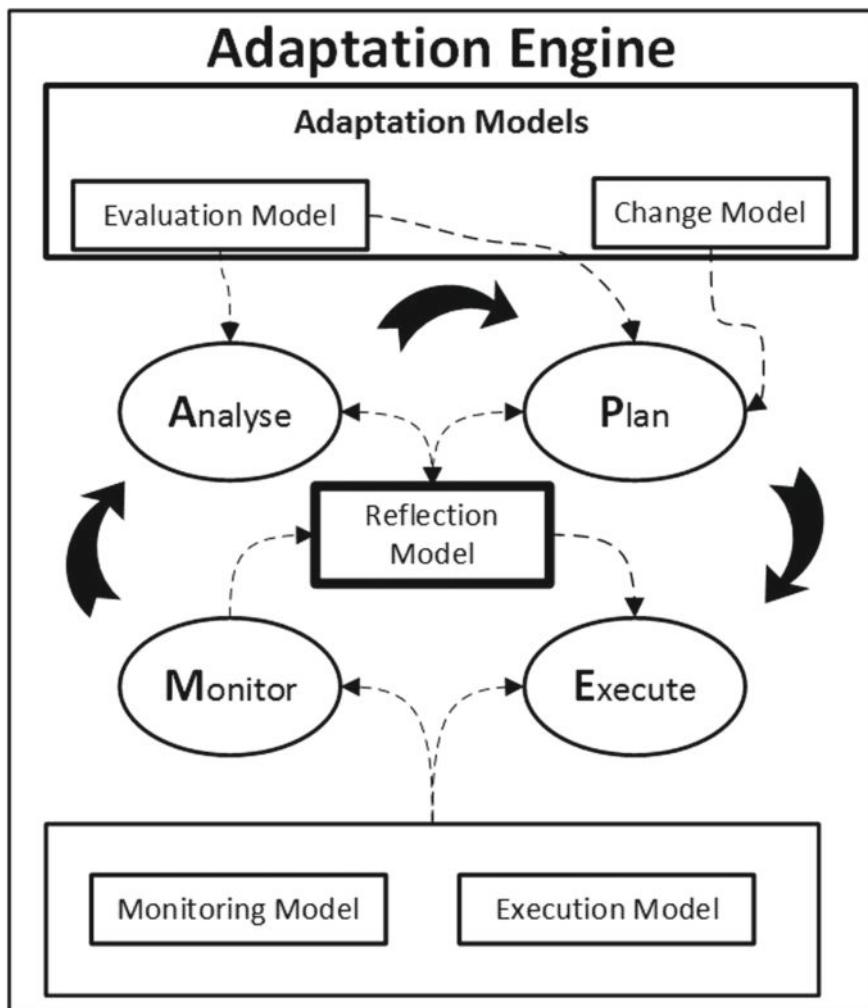


Fig. 1. MAPE runtime model taxonomy. Adapted from [17]

in which swarm elements behave individually or collaborate is derived from state machines models.

Figure 2 shows the state machine for a UV. A UV has five superstates, `At_Base`, `Returning`, `Searching`, `Tracking` and `Assisting`. Unique to this approach to runtime behavioral models are three probabilistic transitions based on a sigmoid threshold. They are: (1) *Transition* ($2.a \rightarrow 4$) - a UV transitioning from its own task to assist another as in the case of a subswarm; (2) *Transition* ($2.a \rightarrow 2.b$) - a UV transitioning from `detecting` to `recognition`; and (3) *Transition* ($3.b \rightarrow 3.c|2$) - the verification of a target by the assistance of another UV. It is the parameters of these sigmoid transitions that are tuned

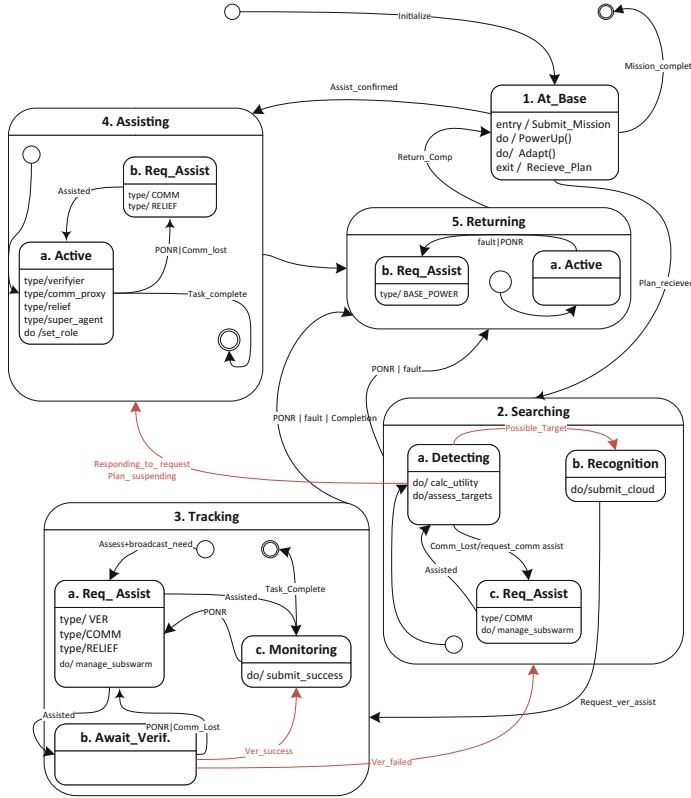


Fig. 2. State machine for UV with probabilistic threshold transitions. Adapted from [1]

between missions by our genetic algorithm for the behavior of a swarm to evolve in the long-term.

Our Bases' finite state machine's transitions are wholly discrete at this juncture of the work. Figure 3 shows the machine as having only two superstates, **Initializing** and **In_Mission**. The base stations will house our autonomic architecture in a distributed manner and will be responsible for establishing the communications framework.

3 Architectural Overview

Within the realm of swarm robotics, group architecture is identified as a research axes for cooperative behavior. Our approach is consistent with that of Cao et al. [3], viewing the architecture as that infrastructure which contains the behavior and ultimately determines the limitations and capabilities of the swarm.

To refresh the assumptions of our swarm concept - the swarm allows for heterogeneous elements. With heterogeneous elements there is a tendency towards

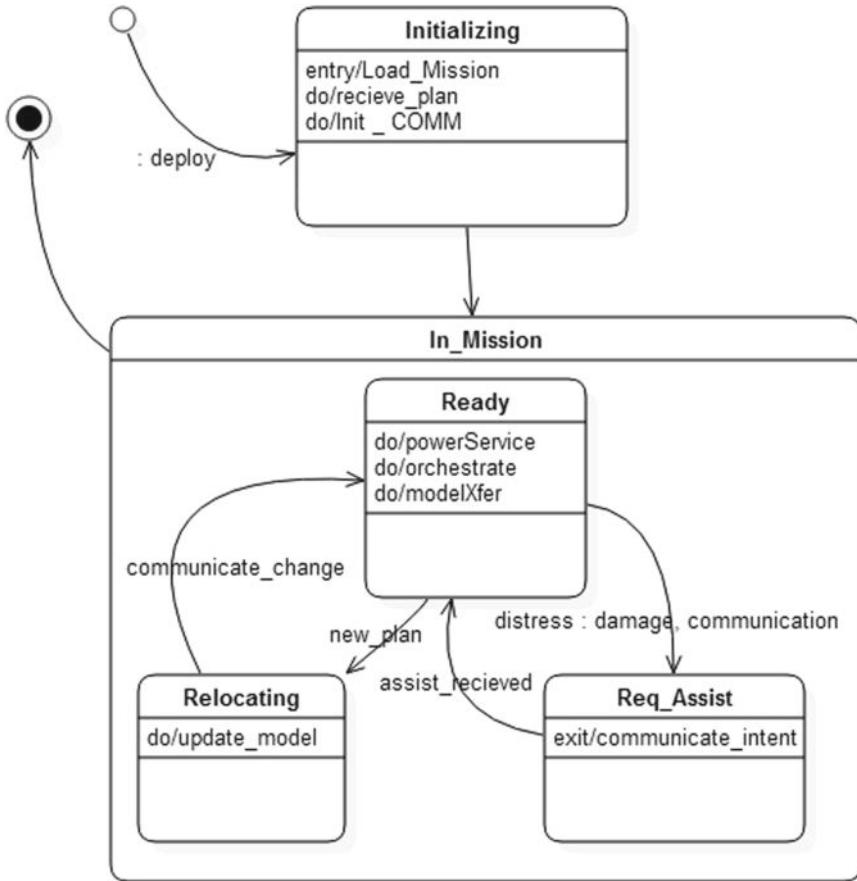


Fig. 3. State machine for swarm base station

low task coverage; the ability for an element to fully accomplish a task in isolation. Geo swarm addresses this challenge by the formation of subswarms.

Consistent with the Object Management Group's (OMG) recommendations for model-driven architectures, Our model centric approach relies on models and transformations based on a clear definition of its metamodel. In this section, we introduce the Swarm metamodel in terms of its constituent abstract syntax (Reflective) and operational semantics (Orchestrator).

3.1 The Swarm MetaModel

Figure 4 represents the abstract syntax for swarm organization. The autonomic operational semantics housed within the **Orchestrator** (separately treated in Sect. 3.2), relies on a causally connected reflection model [17]. Beginning at the top-right, the core reflection model consists of the **Swarm** and its **Mission**.

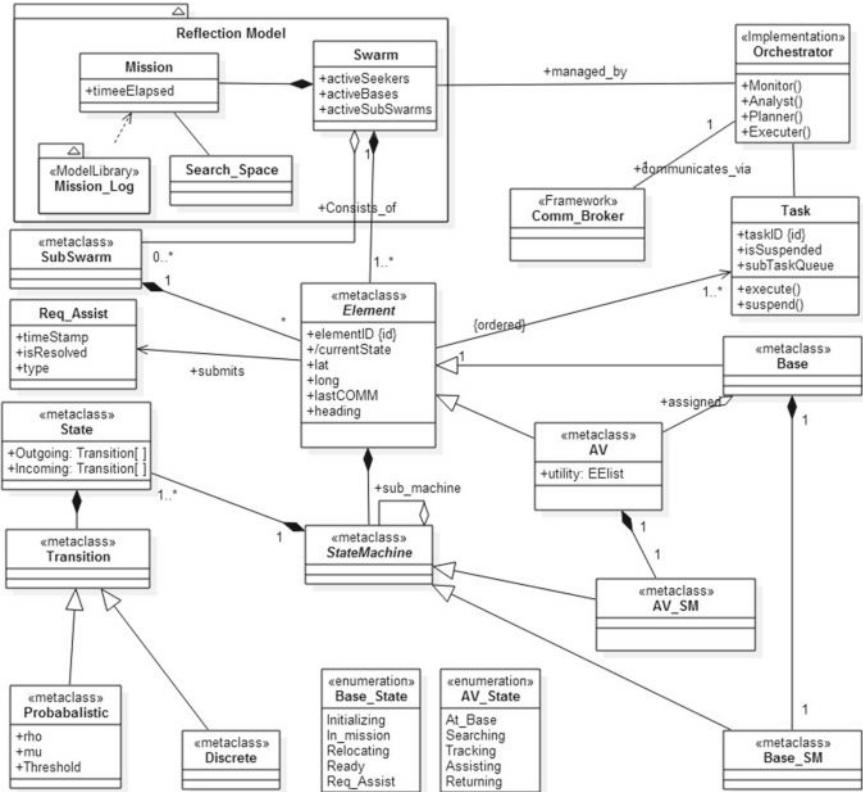


Fig. 4. MetaModel for swarm elements

A **Mission** has a **Search_Space** representation and a **Mission_Log**. The latter reflects the temporal dimension of the mission and is critical for meta-analysis and the rollback to stable states.

A **Swarm** is a collection of **Elements** which in turn may be organized into zero or more **SubSwarms**. Each **Element** is assigned a set of one or more **Tasks** to accomplish the **Mission** of the **Swarm**. **Elements** are either a **Base** or an **Autonomous Vehicle (AV)**. These **Elements**' behaviors are determined by a **StateMachine** (see Figs. 2 and 3), which may have **Discrete** or **Probabilistic** state **Transitions**. We next discuss the behavioral component of this structure - the MAPE Orchestrator (Upper-Right). Note that **Orchestrator** is UML stereotyped as an `<<implementation>>` component and is outside the Meta-model.

3.2 Operational Semantics

Figure 5 presents the deeper look into the operational semantics. The **Orchestrator** utilizes the **Reflection Model** to derive apropos direc-

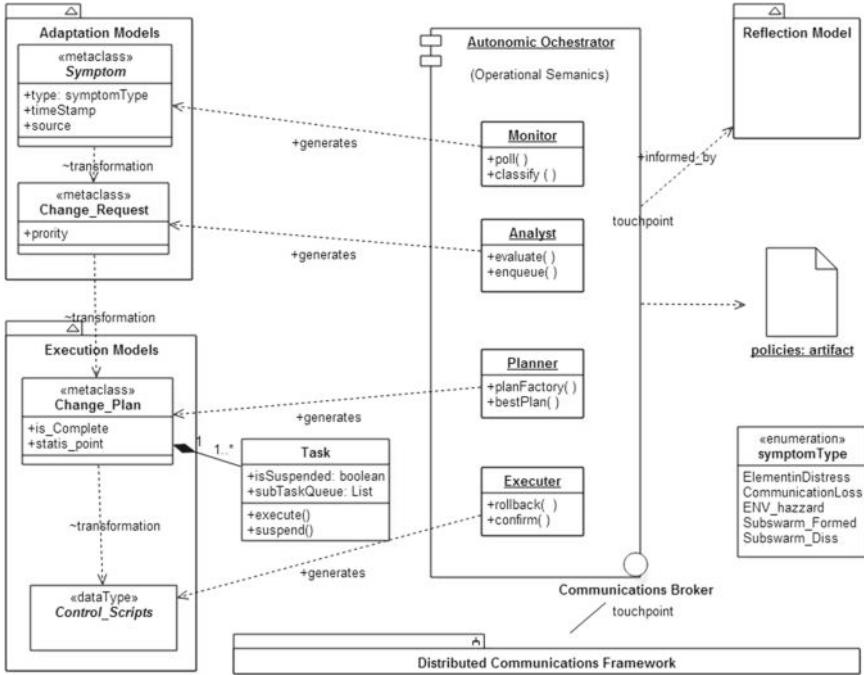


Fig. 5. Autonomic model transformation architecture

tives as **Control_Scripts** which communicates through the **Distributed Communications Framework** via a **Communications Broker** as a touchpoint. The **Orchestrator** can be viewed as a Non-Human Modeler [8] whose reliance on humans would be its **Mission and Policies**. The **Orchestrator**, in keeping with the Autonomic MAPE strategy, contains a **Monitor**, an **Analyst**, a **Planner** and an **Executor**.

Policies are the set of considerations that are designed to inform adaptive decisions. They are manually crafted inputs to the **Orchestrator**. Policies are checked for conflicts and resolved based on priority if necessary. ECA policies are well suited for systems of a reactionary nature [18]. Each policy is a finite set of Event-Condition-Action rules of the form: *iff Cond_i : Evt_i → (Act₁, Act₂, ..., Act_n)* The above rule denotes that upon the occurrence of some event *Evt_i*, the system shall perform an ordered sequence of actions *<Act_{1...n}>* given a specified condition *Cond_i* is met. Policies are categorized according to their mapping to the MAPE subcomponents.

The **Monitor** relies upon events being raised and Polls the **Reflection Model** for Symptoms and anomalies specified by **Policies**. Since much of the changes such as collision avoidance in the swarm are reactionary and handled at a lower level, the **Monitor** filters for **Element** loss or distress and tracks global behavior.

The **Analyst** acts upon **Symptoms** generated by the **Monitor**. **Symptoms** may indicate the need for a change in global behavior but will not necessarily translate to a **Change_Request** to the **Planner**. The **Analyst** looks for constraint violations and determines the priority or importance of the **Symptom** based on the policy rules. An example of an analysis policy rule would be for a base to relocate due to excessively large task reassignment queues or a large percentage of the swarm is assisting in communication.

The **Planner** utilizes a lookup of planning **Policies** to generate a **Change_Plan**. **Change_Requests** are dequeued based on priority and resources available and mapped to a sequence of directed **Tasks** to the **Swarm**.

The **Executor** encapsulates **Tasks** to **Control_Scripts** and ensures execution to completion or rolls back the entire **Change_Plan** and informs the **Planner**. It decouples the commands from their actual execution. Its purpose is to ensure that the **Control_Scripts** are enacted within a reasonable time else a plan failure is raised.

The Autonomic Orchestrator *AO* as an adaptation engine is defined more formally as the tuple:

$$AO = \langle S, \Sigma, s, \Theta, \delta \rangle \quad (1)$$

Whereby:

S - is the set of swarm behaviors that may be represented by the Swarm **Reflective** metamodel.

Σ - is the alphabet of anomalies describable by **Symptom** metamodel MM_{sym} .
 $s \in S$ - the startup configuration for the swarm. The initial reflective runtime model R_0 is the representation of s .

Θ - is the set of tasks that may be represented by the **Task** metamodel.

δ - represents the transition function encompassing the MAPE transformations.

A directly connected causal reflective runtime model R_i and an anomaly $a \in \Sigma$ under δ derives the next reflective model R_{i+1} it follows that:

$$\delta : S \times \Sigma \rightarrow S \times \Theta \quad (2)$$

4 A Self Healing Scenario

Towards reproducibility, we clarify the architecture by providing a walkthrough of a simplified indicative scenario. The ability for the swarm to self-heal is a forefront concern. Small AV's tend to be more vulnerable to failure and are susceptible to less accurate sensor readings [13]. This matter is exacerbated when this effect becomes aggregated for the functionality of an entire swarm. This begs the need for a robust mitigation strategy for resilience to element failure. The **Orchestrator** should react to failure within one or more model elements. In our application context, this may represent the AV exceeding its PONR or a mechanical or communication malfunction that requires a mitigation strategy. For example, in the event of a permanent malfunction, tasks which were assigned

to a particular AV will need to be reassigned for mission continuity. The scenario to be discussed involves an AV in distress due to a violation of its PONR.

We will thread the walkthrough with the required and generated model artifacts. To implement these transformations we rely on the ATLAS Transformation Language (ATL) [11]. ATL is a hybrid model transformation language, allowing for both declarative and imperative constructs to be utilized in transformation definitions. We have chosen this tool due to this expressiveness and its ease of integration within the Eclipse Modeling Framework (EMF) [16]. We next present how a transformation in accomplished using ATL.

Figure 6 represents a transformation instance used by our **Analyst**. Three of the metamodels developed **Symptom** represented (MM_{sym}), **Swarm** (MM_{swarm}), **Change_Request** (MM_{chg}) and ATL must conform to the Meta Object Facility (MOF) metamodel. In turn, the models M_{sym} , M_{swarm} and M_{chg} conform to their respective metamodel. An ATL transformation, `analyst.atl`, takes as input, models M_{sym} and M_{swarm} , then generates a **change_request** Model M_{chg} .

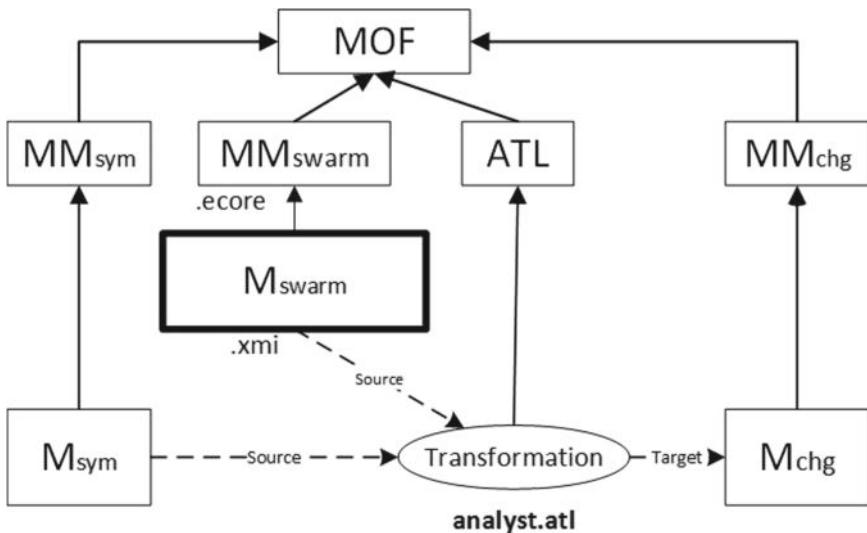


Fig. 6. Analyst transformation using ATL

4.1 Scenario Description

Our scenario begins with a small swarm of Aerial Unmanned Vehicles (AUV)s deployed to find a missing hiker. Each AUV is given a sector of the planned search space and supported by base units. During one of the sorties, one of the AUVs has exceeded its calculated PONR due to an extended wind gust. This

AUV has used much of its reserve power to compensate for this unplanned event. It will not be able to replenish its power via its assigned base, so it submits a `Req_Assist` to the swarm.

4.2 Architecture Walkthrough

Figure 7 is a simplified instance of the causally connected reflective instance of the swarm. Note it conforms to the .ecore metamodel in Fig. 8. The instance shows three AVs AUV001, AUV002, AUV003 and two Bases B001 and B002. Our AUV in trouble is AUV002. Its `currentState` is “returning” and it has already issued an assist request of type “`ponr_exceeded`”.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xmi:XMI xmi:version="2.0"
  xmlns:xmi="http://www.omg.org/XMI" xmlns="Swarm">
  <AV elementID="AUV001" currentState="Tracking">
    <currentTask taskID="Seek001"/>
    <assigned_base elementID="B001"/>
    <req_assist type="COMM" isResolved="true"
      timeStamp="7.1"/>
  </AV>
  <AV elementID="AUV002" currentState="Returning" >
    <currentTask taskID="Seek002"/>
    <assigned_base elementID="B001"/>
    <req_assist type="ponr_exceeded"
      isResolved="false" timeStamp="12.2"/>
  </AV>
  <AV elementID="AUV003" currentState="Searching" >
    <currentTask taskID="Seek003"/>
    <assigned_base elementID="B002"/>
    <req_assist type="COMM" isResolved="false"
      timeStamp="2.3"/>
  </AV>
  <Base elementID="B001" currentState="In_Mission" >
    <currentTask taskID="Service001"/>
  </Base>
  <Base elementID="B002" currentState="In_Mission" >
    <currentTask taskID="Service002"/>
  </Base>
</xmi:XMI>
```

Fig. 7. Swarm instance

The `Monitor` looks at the reflective Model for anomalies. Figure 9 shows an ATL transformation snippet. This snippet queries the model for unresolved requests and finds two. Our AUV in distress, AUV002, has an unresolved `ponr_exceeded` request and AUV003 has an unresolved `COMM` request. The `COMM` request indicates that the AUV requires communication assistance to continue its task. This would be typical of the initial stage of a SubSwarm formation to facilitate communication. `SubSwarm` formations are handled at the element level and would not be considered a swarm (macro-level) fault unless the request persists for an extended period outlined by swarm `Policies`. The `Monitor` translation results in the generation of the `Symptom` model in Fig. 10.



Fig. 8. Partial representation of swarm metamodel as EMF .ecore file

```

module Monitor;
create OUT : Symptoms from IN : Swarm;
helper context Swarm!AV def: unresolvedRequest(): Boolean =
    if not self.req_assist.isResolved then
        true
    else
        false
    endif;

rule Swarm1 {
    from
        s : Swarm!AV (s.unresolvedRequest())
    to
        t : Symptoms!Symptom (
            source <- s.elementID,
            type <- s.req_assist.type,
            timestamp <- s.req_assist.timestamp.toString()
        )
}

```

Fig. 9. Monitor's ATL transformation snippet

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xmi:XMI xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI" xmlns="Symptoms">
  <Symptom source="AUV002" type="ponr_exceeded" timeStamp="12.2"/>
  <Symptom source="AUV003" type="COMM" timeStamp="2.3"/>
</xmi:XMI>
```

Fig. 10. Scenario Symptom model generated by the **Monitor** transformation

The Symptom model along with the Reflective model serves as input to the **Analyst**. The **Analyst** transformation rules looks at the severity of the symptoms. The unresolved COMM request did not meet the need for a Change_Request and is filtered out as seen in Fig. 11.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xmi:XMI xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI" xmlns="Requests">
  <Request target="AUV002" problemType="ponr_exceeded" priority="1"/>
</xmi:XMI>
```

Fig. 11. Scenario Change_Request model generated by the **Analyst** transformation .eps

The Change_Request along with the Reflection Model servers as input to the **Planner**. The **Planner**'s transformation rules dictate that in the event of a ponr_exceeded request, a rescue plan should be initiated that involves two tasks: (1) relocate a base in close proximity that will incur the least amount of disruption to the swarm (i.e least AUVs assigned or returning); and (2) the AUV in distress will be reassigned this new base. The resulting plan is seen in Fig. 12. To ensure that the plan is carried out, the **Executor** encapsulates the tasks into two EXECUTE commands, each with a fixed duration or TTL (Time to Live), and a request for task confirmation. The TTL is important and will inform the **Executor** if they were completed on time.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xmi:XMI xmi:version="2.0" xmlns:xmi="http://www.omg.org/XMI" xmlns="Plan">
  <Task target="B002" type="relocate" lat="000034.5" long="00045.6" isSuspended="false"
timeStamp="19.2"/>
  <Task target="AUV002" type="reassignedBase" lat="000034.5" long="00045.6" timeStamp="19.2"/>
</xmi:XMI>
```

Fig. 12. Scenario Change_Plan model generated by the **Planner** transformation

The tasks now become a part of the reflexive model and are now monitored by the **Monitor**; completing the cycle.

Once the AUV is reassigned to a new base, it will converge upon the relocating base.

5 Conclusion and Future Work

In this paper, we presented a MAPE-K model-driven architecture to produce autonomic behavior in a distributed system. We utilized the Atlas transformational language to reify the operational semantics for proof of principle. This work has extended our testbed to support macro-level behaviors for the swarm. We are encouraged by the high usage of models to tame the inherent complexity which indicates that model-driven approaches and the toolsets are increasingly becoming a viable alternative to the existing object-oriented paradigm. The challenge we now face is to investigate a manner in which this architecture may be made redundantly distributed for mission assurance.

Acknowledgments. This work is supported by the University of Michigan -Flint Office of Sponsored Research.

References

- Allison, M., Spradling, M., Knock, N.: Uav collaborative search using probabilistic finite state machines. In: International Command and Control Research and Technology Symposium—Knowledge Systems for Coalition Operations (2017)
- Bruni, R., Corradini, A., Gadducci, F., Lafuente, A.L., Vandin, A.: A conceptual framework for adaptation. In: International Conference on Fundamental Approaches to Software Engineering, pp. 240–254. Springer (2012)
- Cao, Y.U., Fukunaga, A.S., Kahng, A.: Cooperative mobile robotics: antecedents and directions. *Autonomous robots* **4**(1), 7–27 (1997)
- Cheng, B.H., De Lemos, R., Giese, H., Inverardi, P., Magee, J., Andersson, J., Becker, B., Bencomo, N., Brun, Y., Cukic, B., et al.: Software engineering for self-adaptive systems: a research roadmap. In: Software Engineering for Self-adaptive Systems, pp. 1–26. Springer (2009)
- Computing, A., et al.: An architectural blueprint for autonomic computing. In: IBM White Paper, vol. 31 (2006)
- Fickas, S., Feather, M.S.: Requirements monitoring in dynamic environments. In: Proceedings of the Second IEEE International Symposium on Requirements Engineering, pp. 140–147. IEEE (1995)
- France, R., Rumpe, B.: Model-driven development of complex software: a research roadmap. In: 2007 Future of Software Engineering, pp. 37–54. IEEE Computer Society (2007)
- Garcia-Dominguez, A., Bencomo, N.: Non-human modelers: challenges and roadmap for reusable self-explanation. In: Federation of International Conferences on Software Technologies: Applications and Foundations, pp. 161–171. Springer (2017)
- Harman, M., Jones, B.F.: Search-based software engineering. *Inf. softw. Technol.* **43**(14), 833–839 (2001)
- Horn, P.: Autonomic Computing: IBM’s Perspective on the State of Information Technology (2001)
- Jouault, F., Kurtev, I.: Transforming models with ATL. In: International Conference on Model Driven Engineering Languages and Systems, pp. 128–138. Springer (2005)

12. Kephart, J.O., Chess, D.M.: The vision of autonomic computing. *Computer* **36**(1), 41–50 (2003)
13. McCune, R.R., Madey, G.R.: Swarm control of uavs for cooperative hunting with DDDAS. *Proc. Comput. Sci.* **18**, 2537–2544 (2013)
14. Şahin, E.: Swarm robotics: from sources of inspiration to domains of application. In: International Workshop on Swarm Robotics, pp. 10–20. Springer (2004)
15. Schmidt, D.C.: Model-driven engineering. *Comput.-Comput. Soc.* **39**(2), 25 (2006)
16. Steinberg, D., Budinsky, F., Merks, E., Paternostro, M.: EMF: Eclipse Modeling Framework. Pearson Education (2008)
17. Wätzoldt, S., Giese, H.: Classifying distributed self-* systems based on runtime models and their coupling. In: Models@ run. time, pp. 11–20. Citeseer (2014)
18. Wirtschaftswissenschaftlichen, D., Kradolfer, M., Dittrich, D., Alonso, D.: A workflow metamodel supporting dynamic, reuse-based model evolution (2000)



Review of Paradigm Shift in Patent Within Digital Environment and Possible Implications for Economic Development in Africa

Stephen Odirachukwu Mwim^(✉) and Tana Pistorius

University of South Africa (UNISA), Pretoria, South Africa
mwimattorneys@gmail.com, pistot@unisa.ac.za

Abstract. This study examines the impact of BMP protection on development by focusing on the challenges confronting economic growth in African communities as a result of the new paradigm in patent law. [Africa is used as a single unit in this study but this should not be construed as African homogeneity. Rather the views advanced in this study are used to could be applicable to many communities in Africa.] There are very few study on the impact of BMPs perspectives on economic development particularly in Africa. The purpose of this paper is therefore to review the extent of debates and discourses that has taken place among researchers and policy makers on the impact of BMPs perspectives on economic development in Africa. The paper deems it important to ignite or accelerate debate in this area. As a starting point the paper reviews (from the point of views of legal philosophers, policy makers and decisions of competent courts) the relevant literature, patent legislation particularly the International Treaty, policies and legal judgments. Findings from this study suggest that over and above the various criticisms levelled against the extreme liberal approach to the recognition of business methods as patentable subject matter, there are other specific implications that are associated with such approach. The most critical implication of extending patent protection to business methods is the locking-up of knowledge which may hamper human development in general and economic development in particular. Locking up knowledge that is otherwise necessary for economic advancement and competitiveness may have a negative effect on economic growth by promoting economic exclusion, particularly in African communities. This study suggests that advancing a system of BMP within the African context and the extent of protection linked to business methods is crucial in achieving a sustainable economic growth in Africa. It also suggests that a balance should be struck between the two diametrically opposing views on the protection of business methods.

Keywords: Africa · Business Method Patenting · Digital economic growth · Patent protection

1 Introduction

The perception of the relation between patent protection and development, particularly economic development, has evolved significantly in the past few years. Debate on patent protection in the international arena has been significantly influenced by the perception that there is a strong link between patent protection and economic development. Recently there has been a paradigm shift with a lot of emphasis on extending patent protection to software inventions and method of doing business within the digital environment, generally referred to as Business Method Patenting (BMP). The general perception among international organizations and the private sectors also indicates that there is a strong correlation between BMP protection and economic growth. There are two diametrically opposing views as regards the relation between patent protection particularly BMPs and development and innovation. One school of thought promotes the view that the protection of software inventions and business methods improves economic development through stimulation of innovation and creativity. The other school advances the view that such protection is unnecessary for stimulation of innovation and creativity and is in fact a hindrance to open access to resources and information required for innovative and creative modalities. Therefore, various theories and policies attach different levels importance to BMP as a consequence of its effect economic growth and innovation.

This paper focuses on how paradigm shift in patent influences the perspectives on the protection of business methods. The paradigm shift in patent considered in this paper relates to the changes in perspectives on the protection of business methods. This paper reviews how the protection of business methods within the digital environment evolved as result of pressure on policy makers and subsequent judicial decisions.

2 Literature Review

In recent years there has been increasing pressure to liberalise the availability of patent protection for software inventions and business methods [1]. In Europe, for example, a computer programme was only denied patentability if it was shown that the mathematical algorithm and program subject of the application were wholly abstract, and as such could not constitute a useful, concrete and tangible result [1, 2]. However, the practice in USA changed significantly in 1995 to allow for carrier claims in respect of software patents [3]. The acceptance of patents for computer programs assisted in closing the door on the courts' ability to distinguish between concepts and machines, but opened the door for the patenting of business methods [2]. This background marked the turning point in the patent system, resulting in a shift in the patent paradigm, especially in the USA and fairly soon afterwards in Europe, Japan and other regions. This paper reviews the important judicial decisions in USA to the extent that such decisions brought about changes in the perspectives of the protection of patent. The paper makes reference to the perspectives in Europe, Australia and Japan. [These jurisdictions are merely used as samples to show how the issue of BMPs is viewed in some other jurisdictions.]

BMPs within the digital environment is one of the latest developments in the area of IP law that deserves continues scrutiny. The emergence of BMPs has provoked debate among IP specialists. In the same vein the economic implications of extending patent protection to methods of doing business and the impact of such extension on businesses have sparked extensive discussion among academics, professional entities, jurists, and other sectors of the society [4]. The question, in light of the purpose and basic elements of the patent system, is whether or not it is desirable to include methods of doing business as a patentable subject matter, and to extend patent protection to business methods.

There is little emphasis on the implications of the various perspectives of BMPs for economic growth in African communities. There are very few study on the impact of BMPs perspectives on economic development particularly in Africa. The purpose of this paper, under the backdrop of the BMPs perspectives in general, is therefore to review the extent of debates and discourses that take place among researchers and policy makers on the possible impact of BMP on economic development in Africa communities. This paper aims at sparking debate among researchers and policy makers on the area of patent. The paper therefore examines the implications of the shift in paradigm for economic development especially in Africa with the intention to ignite or intensify debate in this area.

In this paper the evolution in BMPs and the legal philosophical principles underpinning this paradigm is considered against the backdrop of the objectives of patent protection. In order to appreciate the nature and implications of BMP protection, it is vital to understand the fundamental principles underlying the patent system in general. The reader may then be in position to comprehend the extent to which the introduction of BMP influences the perceptions and objective of patent protection which resulted in a fundamental shift in patent perspectives.

3 Research Method

The research method adopted in this paper is the examination of the different approaches adopted by the various courts and schools of thought in addressing the challenges facing patent particularly the protection of business methods in digital environment. It examines the different legal philosophical approaches to paradigm shift in patent protection with particular reference to BMP. It considers and reviews several court decisions in different jurisdictions where issues regarding BMP and software protection were considered. The paper further considers the possible implications of the shift in patent protection within the digital environment for economic growth within the developing communities in Africa. It reviews the literature in this area with the intention of igniting more debate among researchers and policy makers.

4 Fundamental Principles Underpinning Patent System

A patent denotes a document which states the scope of patent rights to exclude others from making, using, or selling an invention that is the subject of the patent [5]. In the same vein a patent also grants the patentee a statutory monopoly, among others, to exclude others from disposing or offering to dispose of, or importing the patented invention [6, 7]. A patent represents a bargain with the extended society. The inventor must, in return for the exclusive rights associated with patent, disclose the details of the invention to the public, instead of keeping them secret [2].

The patentee is, therefore, granted a limited monopoly in relation to his or her patented invention. The ultimate purpose of a patent system is to promote science and the useful arts [8]. In exchange for the promotion of science through disclosure, owners of the patented invention are allowed a limited period in which to exploit their invention, and may also exclude others from performing, in relation to the invention, those acts that only the patentee is permitted to perform. One of the reasons for allowing owners of patents the exclusive right to exploit their invention is that they may at least recoup their research and development costs [2].

Historically, patents seek to promote general welfare by protecting the fruits of intellectual creativity from actions that would frustrate the inventor's chances of reaping rewards from his or her investment of time, money, or talent [9]. Patent rights, in themselves, are not intrinsic or inherent rights [1]. As such, an optimal balancing of monopolistic disturbances to the market, on the one hand, and incentives for the creation and diffusion of new knowledge, on the other, is most desirable [1]. In the same vein some patent commentators argue that it is essential to maintain a proper balance between the interests of the public and the interest of the inventors [10]. It is necessary that inventors are legitimately and fairly rewarded for their innovations [11] but at the same time the public must be protected from having to pay exorbitant monopoly prices for goods and services.

The goal of patent law—apart from providing incentives for innovation—is to realise the ultimate contribution of the innovation to the public domain thereby enhancing consumer welfare [12, 13]. Patent protection is necessary to inspire innovation, since research and development costs can be extremely high. Patent law assumes that quality of life would suffer greatly from technological stagnation that would result from the absence of patent protection [12].

In order to realize its objectives of innovation, dissemination, and resultant consumer welfare, the patent law contains two mechanisms. Firstly, it contains an enforceable property right that protects creation and avoids exploitation and free riding of imitators. Thus, patent law provides a limited term monopoly offering both an incentive and a reward for innovation. The second mechanism is that the patentee, in exchange for the monopoly, must fully disclose the specifics of the invention to the public. Full disclosure of the invention in question adds to the knowledge base of society and promotes innovation. There is a notion that patents were traditionally intended to protect advancements in technology, which clearly encompassed machines [2, 14]. This view necessitates the importance of turning to the specific exclusions in the area of patent that need consideration.

5 Specific Exclusions

There are specific categories of inventions that, though they may meet all the basic requirements for patentability, are specifically excluded from classification as patentable subject matters. A program for a computer and a method for doing business, among other things, were clearly excluded from patentable subject matters.

It may also be argued that methods of doing business are not protected under the TRIPS Agreement. The underlying purpose of the TRIPS Agreement is to promote free trade [15]. Some writers therefore maintain that the patentability of methods of doing business cannot be sustained. From the Australian perspective for example, just like in few other jurisdictions around the globe, one of the traditional exceptions to patentability was the business method exception. This form of exception has been considered in several Australian cases. For example in *Re Cooper's Application for a Patent* it was held: "You cannot have a Patent for a mere scheme or a plan for the efficient conduct of business. The subject with reference to which you must apply for a Patent must be one which results in a material produce of some substantial character" [16]. This view was further confirmed in *Commissioner of Patents v Lee* [17] as well as in *Roger v Commissioner of Patent* [18].

On the contrary however, in *IBM's Application*, it was held that it is not logical to refuse claims to a program or carrier if the running of the program involves a technical contribution, provided that the claims are properly defined and delimited in the program in question [19]. Consequently, the exclusion of a computer program should only mean the exclusion of the program in itself. This entails the exclusion of a program which lacks a technical character [20]. Before the *IBM* case, the practice in Europe and the UK was based on the notion which claims specifically that a computer program, even if it was a program which delivered a patentable technical effect, were not granted patents [21, 22].

In terms of article 52(2), for example, mathematical methods, computer programs, and methods of doing business, *inter alia*, are specifically excluded from patentable subject matter [23]. Applying the provision to article 52(3), however, the mere presence of hardware in the computer, if a business method is implemented by running a program on a general-purpose computer, does not render the method patentable if no technical contribution to the "art" has been made [23].

The Court of Appeal in *Aerotels* stated that business-method exclusion was not limited to abstract matters [24]. The Appeal Court also stipulated that there was no need for an activity to be completed for it to fall within the exclusion [24]. The Court in the *Raytheon* [25] confirmed the decision in *Aerotels* case. In the same vein, in *Autonomy v The Comptroller General of Patents* it held that the idea of presenting information to be used in undertaking inventories in a pictorial form was a method of doing business [26]. Then in the case of *Quest International* it was held that the mere fact that an invention provides financial benefits is not enough for it to be classified as a method of doing business otherwise there is danger that nearly all patents would fall within the exclusion [27].

Although the decision in the *IBM Application* [19] brought patent protection for computer programs in the UK (and Europe) closer to that obtaining in the USA, the

latter, with the decision in the *State Street* [28] case, is of considerable interest as regards the expansion of the scope of patent subject matter which resulted in the paradigm shift in the patent system.

6 Shift in Paradigm

Methods of doing business have changed from non-patentable subject matter to patentable subject matter. The birth of BMPs has introduced a new dimension to the patent system. The paper therefore examines the various aspects of business methods, judicial approach to the BMP Justification, as well as the impact of courts' decisions and provocative patent controversies.

6.1 Aspects of Business Methods and the Contrasting Views

Although business methods include both Internet-based and non-Internet based ones, this paper mainly makes reference to the former though not exclusively. The majority of criticism relating to BMPs focuses on business processes and techniques implemented on the Internet. Internet-related patents fall into several categories. However, when one hears about BMPs, what immediately comes to mind is the model for doing business on the Internet. This model, which is exemplified by the Priceline.com reverse auction for purchasing airline tickets, is just one aspect of Internet-related patents.

The second type of Internet-related patents is where a patented invention is not intended as a business model, but rather as a means for solving specific business problems. The patents that fall into this category are often referred to as Internet business technique patents. The Amazon.com one-click patent illustrates a situation in which an invention serves the purpose of solving specific business problems. This type of Internet-related patent, when compared with the first type, is narrower in scope.

The third type of Internet-related patent involves techniques that purport to make the Internet more efficient and effective for conducting electronic commerce by solving a technical software problem [7]. The patents that fall into these categories are often referred to as Internet software technique patents. Sometimes, it is difficult to make a distinction between the second and third types of Internet-related patents.

Following the developments in the USA, the common perspective in Europe is that business methods, like computer programs, may be patentable if incorporated as part of an invention producing a technical effect [2]. Although the EPC originally denied patent protection for computer programs, the national courts in Europe and the EPO Board of Appeal expanded protection for such inventions, in response to their acceptance in countries such as the USA. Therefore, in order to determine whether or not business methods involving the use of a computer are patentable, we must consider the inventiveness of the computer program used to implement the method, rather than the method per se [2]. The European Technical Board of Appeal of the EPO held in *Sohei* that a computer-implemented business method is patentable if it is capable of having a technical effect [29]. It was stated that innovations are not always considered worthy of protection—especially if they are purely abstract business methods and where there is no means of implementing such ideas. If, however, the implementation

of a method of doing business involves a solution to a technical problem, that means of implementation is likely to be patentable [30].

Japanese patent law also experienced a turning point as a result of the business method perspectives in the USA and Europe. It has been amended to accommodate evolution in business methods. The JPO deems it proper to accept a business method when claimed as a part of an invention involving a computer program as patentable subject matter [31]. The JPO's view is that most business-related inventions can be considered as another form of software-related inventions [31]. As such, a business method implemented by way of a computer program may, if the other basic requirements for patentability have been met, be patented [2, 32].

Also it should be noted that, unlike any other BMPs, those ones tailored for Internet usage are the most controversial, and are the most likely to cause economic harm if granted when they should not be, because of their potential for impeding electronic commerce while it is still maturing [7].

Whereas Japan and Europe follow a more restricted approach, the USA follows a more liberal approach regarding the extent to which patent protection extends to business methods. For the time being, Europe and Japan have refused to follow the USA in fully extending protection to business methods [2]. It could be argued that the industrial application requirements in Japan and Europe, coupled with their limitation of patentable inventions to purely technological innovations, limit the extent to which protection may be sought for inventions of an economic nature.

Although Europe and Japan, among others, have joined the USA in projecting the view that patent protection extends to methods of doing business, the standards of patentability in Europe and Japan reflect the more traditional views of patentable subject matter [2]. This paper argues that while the USA, Europe and Japan have all proven amenable to protecting innovations in biotechnology and computer programs, Japan and Europe have been more hesitant to extend patent protection to innovations in business-related inventions.

While jurisdictions such as European in general, Japan and Australia advance the USA's approach of extending patent protection to methods of doing business—albeit cautiously—some other jurisdictions such as Mexico, China, Singapore and Malaysia, deny patents to business methods [33]. As certain jurisdictions have indeed extended patent protection to methods of doing business this papers eventually considers the possible implications for the justification of BMPs.

In an attempt to justify the need for BMPs, some commentators argue that all human endeavours involving the application of time, money and mental labour in a business setting deserve equal protection [34, 4]. Furthermore, BMPs can be seen as an important way to safeguard such application as well as any other invention of innovative financial techniques [35]. Other commentators who support the new paradigm for patents argue that the protection of methods of doing business results in the patent system keeping pace with the technologies of the digital and information age, including e-commerce and data processing [4].

It is noted that the courts play a critical role in shaping the direction of patent system. There are some important court decisions that influenced the perspectives of BMP and how the protection of business methods is viewed.

6.2 Judicial Approach to the BMP Justification

Shortly before the decision in *State Street* [28], the courts attempted to limit the business method exception by arguing that a patentable claim could be business-related on condition that it was directed to otherwise patentable subject matter [30]. As a result of the new economy of high technology where the disparity between methods and means was starting to blur, the business method exception began to lose its usefulness [36]. Due to recent advances in technology, what was ordinarily regarded as an idea could now evolve into a useful system through the application of a new technology, and as such become patentable [30].

Coupled with the influence of emerging technology, the broad provision of section 101 of the USA Patents Act, which states that whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter or any new and useful improvement thereof, may obtain a patent [37]. What stands out in the above provision is the absence of the technicality requirement. This opened the door for a more liberal approach in the patent system.

In keeping with the above view, the USA Supreme Court held in *Diamond v Chakrabarty* [38] that a patent could be granted to anything made by man which is consistent with the use of the term “any” in section 101 of the USA Patents Act [39].

The decision of the Federal Circuit in *State Street* may be considered a turning point in the patent system. This decision eliminated the restriction on the patentability of methods of doing business in the USA [33]. This case concerned a computerised business method that pooled mutual fund assets into an investment portfolio organised as a partnership for tax benefits. The subject of the application was essentially a computer program for the management of investments owned by various investment funds, which had merged their funds for that objective. The program was organised advantageously for tax purposes.

The patent in this case involved a data processing system for managing a partnership of pooled funds in accordance with certain provisions of the Internal Revenue Code (IRC) and implementing regulations. It concerned, inter alia, a combination of a general purpose digital computer with a CPU, a data storage memory, and five means for performing various functions. The listed functions corresponded to the requirements of the Internal Revenue Services (IRS) regulations and underlying statutory provisions when carried out by means of instructions of a computer program [28]. In essence, there are statutes and regulations to the effect that, “in order to obtain a single level of taxation rather than suffer double taxation, *a, b, c, d and e* must be done. The patent in question claims the combination of means for performing *a*; means for performing *b*; ... and means for performing *e*” [28, par. 1378]. The patent claims, therefore, included any computer programme used to implement the accounting operations for the business method in question rather than just the particular programming code used in the process [28]. The district court found that the effect of claim 1 was to foreclose any computer-implemented accounting method necessary to manage this kind of financial structure [28]. The district court held that patenting an accounting system necessary to carry out a certain type of business is tantamount to a patent on the business itself. It further held that, since “such abstract ideas are not patentable, either when regarded as methods of doing business or as mathematical algorithms, the 056 patent must fail” [28, par. 516].

On appeal, the Federal Circuit Court reversed on both grounds upon which the district court had supported its conclusion [28]. With regard to the algorithm ground, the Federal Circuit held that Claim No. 1 in the application did not claim an abstract idea. It was argued that the calculations produced a useful, concrete and tangible result, viz a price figure accepted and relied upon for regulatory and other business purposes [28]. Turning to the business method perspective, the Federal Circuit maintained that the issue of the business method exception is ill-conceived and must be laid to rest. In its own words, the court stated: “We take this opportunity to lay this ill-conceived exception to rest” [28, par. 1375] by ignoring commentary and dismissing case law, by finding all of it to be either *obiter dicta* or decided on grounds other than the business method rule, for example, on obvious grounds [40]. In fact, the court held that there was no business exception, and never had been [40]. It was further stated that a financial business method that transforms data to produce a useful, concrete and tangible result is eligible for patent protection, and that the “business method” and “mathematical algorithm” statutory subject matter exception categories had little, if any, applicability. On a more serious note, the Federal Circuit abolished the business method exception as an unwarranted limitation to statutory subject matter [28]. The court officially announced the end of the *per se* un-patentable rule for business methods [28]. The consequence of the *State Street* decision is that business methods as a patentable subject matter are no longer tested or judged on the basis of the exception principle, but rather against the basic requirements for patentability.

The views developed in *State Street* are not only controversial, but also set a precedent for later court decisions in the USA [2]. In keeping with the decision in *State Street*, other cases followed suit. A crucial decision in *AT&T v Excel* [41], followed the judgment in *State Street*. In *AT&T* the court reaffirmed and strengthened the decision in *State Street* by stating that any computer-implemented invention, apparatus, or method that is new and useful is a patentable subject matter [41]. The court held that a telecommunication service provider’s method for modifying message records used by local and long-distance telephone service providers to monitor and eventually bill long-distance calls, in order to allow them to easily identify a caller’s long-distance service provider, is a statutory subject matter and as such, eligible for patent protection. The court considered the District Court’s contrary conclusion that the invention could be categorised as reciting a “mathematical algorithm” to be improper [41].

The decision in *AT&T Corp v Excel Communications Inc.* [41] went further by lowering the evidentiary threshold to allow mathematical algorithms to be patented. AT&T was granted a patent for a modification of the telephone-charging method in the USA. Excel Communication, a competitor of AT&T, took action against the patentability of the method. The Court of Appeal only examined the ground for exclusion of a mathematical algorithm, and held that the process was in principle capable of patent protection. The court was of the view that the test to determine patentability was whether the algorithm-containing invention as a whole, produces a tangible and useful result [41]. The determination should not be based on whether there is a mathematical algorithm at work. The court held that the method used by AT&T Corp provided concrete results [41]. As such, the notion of physical transformation advanced in *In re Alappat* [42] cited in *State Street* [28] and *AT&T Corp* [41] where the court ruled that the issue was not whether or not a process or method possessed some

arbitrary degree of physicality, but whether or not it produced a useful, concrete, and tangible result, was upheld.

Note that, in *AT&T Corp v Excel Communications Inc.* [41] the question of a business method as a ground for exclusion was not considered, since the court in the *State Street* case had already rejected this ground.

Another critical case in which issues concerning BMPs were addressed is *Amazon.com* [43]. Here Amazon.com filed suit for infringement against its online competitor, Barnesandnoble.com. The US Patent Office granted Amazon.com a patent over the method and system for placing a purchase order via a communication network. Amazon.com claimed that the one-click method was a major innovation in e-commerce, which allowed customers to order, pay and arrange for delivery of any item the company sold with a single click of a mouse [44]. The district court held that Amazon.com had clearly shown that there was indeed infringement, and that it was likely to suffer irreparable harm if no injunction was issued, and furthermore that it was in the public interest to grant the injunction because it would encourage innovation. Therefore, the court issued Amazon.com with a preliminary injunction to prevent Barnesandnoble.com from further use of its technique for purchasing products [43]. On appeal, however, the defendant argued that Amazon.com had merely patented the final click in what was in fact a lengthy process of steps that buyers must necessarily follow in order to make an Internet purchase. In response, Amazon.com argued that the invention solved the problem of net shoppers abandoning their shopping carts before completing the transaction [43]. The Court of Appeals set aside the decision of the District Court based on doubts as to the validity of the patent [43].

The decision in the *State Street* [28] formally resulted in the birth of BMPs in the USA. Following this decision, there was no longer a patentable subject matter exception for methods of doing business. Therefore, business methods qualified as patentable subject matter subject to the general requirements for patents.

This decision unjustifiably expands the scope of patentable subject matter to such a point that it reduces considerations of patentability in the USA to a minimal determination of utility and novelty. This extension has blurred acceptable notions of the kinds of innovation that should be protected [2].

The decision in *State Street* also makes the application of e-commerce, especially in this era of digitisation, more critical. In this regard, some commentators maintain that the *State Street* decision has such an enormous impact on business-related inventions on the Internet because it is there that methods of doing business have rapidly combined with emerging computer technology to fuel the emergence of e-commerce. Accordingly, Internet companies, more than any others, have received wide publicity and sparked persistent controversy for the skilful exploitation of their patents [12].

6.3 The Impact of Courts' Decisions and Provocative Patent Controversies

The impact of the court's decisions regarding business-related inventions on the "Internet-based business sector" is enormous [12]. Accepting a method of doing business as patentable subject matter has triggered a lot of controversy and uncertainty in the Internet-based business sector [12, 45].

Here are some of the provocative patent controversies relating to the Internet-based method of doing business. The first example is the patent controversy involving Amazon.com and BarnesandNoble.com. In this instance, Amazon, the nation's largest online bookseller, holds a patent for its “one-click” check-out feature, which is the Web analogue of the items arranged near the supermarket register designed to trigger impulse buying. Amazon sued Barnes and Noble, alleging that the latter's single-click “Express Lane” Web purchasing technique infringed on the former's “one-click” check-out feature. The District Court held, on the one hand, that encouraging Amazon to continue to innovate, and on the other, that forcing competitors to come up with their own ideas, unquestionably best served the public interest. Accordingly, the court prevented Barnes and Noble.com from continuing to offer an “Express Lane” feature that infringed on the claims of Amazon's patent. The ruling stirred serious criticism and resulted in calls to boycott Amazon because of its attempts to tax e-commerce through patents [46, 47].

Another example is the patent controversy involving Priceline.com who received a patent for their “reverse auction” service. This e-commerce system enables consumers to name their own price for a variety of goods and services. The Priceline.com patent was harshly criticised neither as novel nor non-obvious [46, 48].

7 Possible Implications for Economic Development Particularly in Africa

The concept of the BMP is very broad since it connotes different types of methods, conduct, and processes linked to a computer, including economic undertakings, marketing strategies, financial services, and other strategies such as sports, games and legal [49]. Unrestricted extension of patent protection to BMPs is a threat to the public domain and Internet-based knowledge since such could limit availability of some basic information that could have been readily accessible for socio-economic development. [The term “Internet-based knowledge” is used in this dissertation to refer to the kind of knowledge that is made available on the Internet.] Members of communities that are already economically disadvantaged could be exposed to deeper socio-economic crisis associated with locking of Internet-based knowledge.

The implications of insisting on a broad patent protection for business methods are twofold. In the first place, it can lead to “lock-in” and monopolisation of information that should under normal circumstances fall within the public domain. Secondly, it may contribute to perpetuating the divide between the “haves and have-nots” in relation to the Internet usage and access—the privileged and the less privileged—in relation to Internet facilities and accessibility of the Internet-based information that could have been readily available. [The term “privileged”, in the context used here, refers to economic powers, on the one hand, and knowledge of and access to digital technology in general and Internet tools in particular, on the other hand.]

Supporting the idea that BMPs may result in the “lock-in effect”, some patent commentators maintain that BMPs have the effect not only of creating the potential to secure vast monopoly powers, but also of generating a reward greatly in excess of the

cost of the claimed invention [50]. The critics of unrestrictive BMPs are apparently critical of the monopolistic power that may arise from the developments in BMPs.

It could be argued that the so-called superhighway of electronic commerce could be partially converted into a toll road if the boom in BMPs continues at its accelerated pace. This may in turn result in taxing the Internet through patents [51]. In the same vein, patented products may become too costly as a result of their condoned monopoly status [12]. Another concern is that BMPs will turn the Internet into a big business-controlled institution. If such developments in patent law are not carefully monitored, BMPs will disable the independent, unaffiliated, critical, questioning creativity that the Internet of the last ten years has produced [52]. All the concerns raised by the critics have specific socio-economic implications specifically for communities in Africa that are already disadvantaged economically.

Bearing in mind its danger, some commentators are very clear in their position when they state that a software-embodied business method is not and should not be treated as patentable subject matter [53, 54]. Some are of the view that BMPs have crossed the boundary from a substantial and tangible world into the realm of thought and abstraction [55].

It is necessary to note once again that the “lock-in effect” of extending patent protection to business methods is interwoven with the view that BMPs perpetuate the existing barrier between the privileged and the less privileged. Many individuals including upcoming entrepreneurs and business professionals, in the developing nations particularly in Africa fall within the category of the less privileged. Therefore, the “lock-in effect” of BMPs becomes even more problematic when considered against the backdrop of economic and social orientation in Africa.

There is an extreme view that business methods should not be patentable at all because of the enormous economic burdens it places on consumers and society in general’ [4, p. 241]. The impact of such patent system is more drastic for developing nations and especially those in Africa that already experience various kinds of digital technology divides. Strict protection over BMP may bar small businesses especially in Africa from competing, particularly in the digital environment, with their rivals in developed countries. This new patent paradigm therefore encourages both lack of access to knowledge and lack of access to economic development particularly in Africa.

8 Conclusion

Many communities in Africa are lacking in innovation while others are faced with challenges in the area of economic development. Yet innovation and creativity in Africa are not only improperly promoted by the current IP system, but also constrained by sub-optimal IP-related policies and practices. It is shown in this paper that the current paradigm in patent law renders the situation even more critical. Extending patent protection to methods of doing business and software programs could disadvantage small businesses within African communities and prevents them from participating actively in network market systems and enterprises that open door to economic advancement. Little effort is made through the engagement of researchers, IP scholars,

institutions, and civic organisations to develop workable IP systems and policies that will address the challenges confronting human and economic development in Africa as a result of digital divide and socio-economic exclusions.

In light of the above, it is necessary to examine and thoroughly analyse any policy and set of laws that aims to protect business methods. There is no doubt that it is necessary to develop policies and laws that assist in regulating access to software that embody business methods, if the basic requirements as would contain in those policies and laws have been met. However, it is highly doubtful whether patent law is the most suitable area of the law for protecting business methods. Generally speaking, it is problematic to insist that business methods should be recognised as patentable subject matter. Another possible solution is to strike a balance between the two extreme contrasting views with regards to the protection of methods of doing business. There should some stringent requirements to be met for business methods to receive protection under patent.

This paper aims at accelerating debate among researchers and policy makers on how BMP perspectives affect economic development in Africa. The paper also suggests that any proposed policies, in response to extension of protection to software and business methods, must be sensitive to the individual circumstances of various communities, nations and continents. Such policies must recognise the fact that different communities and nations are confronted by different social, political, and economic challenges that often shape their perceptions of patent issues.

References

1. Savin, A.: EU Internet Law. Edward Elgar Publishing, Incorporated, UK (2017)
2. Ouellette, L.L.: Patent experimentalism. Va. Law Rev. **101**, 65 (2015)
3. In re Beauregard (Fed Circuit 1998) 53 F 3d 1583
4. Pagán, C.O.C.: Business method patents: a controversy for companies. Revista Derecho Puertorriqueño, **50**, 239 (2011)
5. Pienkos, J.T.: The Patent Guidebook. American Bar Association (2004)
6. The Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) of 1995
7. Allison, J.R., Lemley, M.A., Schwartz, D.L.: Our divided patent system. In: The University of Chicago Law Review, vol. 82, p. 1073 (2015)
8. The United States Constitution
9. Lemley, M.A.: Software patents and return of functional claiming. In: The Robert W. Kastenmeir Lecture, University of Wisconsin Law School (2012)
10. McNamara, J., Cradduck, L.: Can we protect how we do what we do? A consideration of business method patents in Australia and Europe. Int. J. Law Inf. Technol. **16**, 96 (2007)
11. IP and Competition Review, Final Report of September (2000)
12. Marsnik, J.S., Thomas, R.E.: Drawing a line in the patent subject-matter sands: does Europe provide a solution to the software and business method patent problem. Int'l Comp. L. Rev. **34**, 227 (2011)
13. Hulse, R.: Patentability of computer software after State Street Bank & (and) Trust Co v Signature Financial Group Inc: evisceration of the subject matter requirement. UC Davis Law Rev. **33**, 491 (2000)

14. Desai, D.R., Magliocca, G.N.: Patent, meet napster: 3D printing and the digitization of things. *Georgetown Law J.* **102**, 1691 (2014)
15. Nuno Pires de Carvalho: The Trips Regime of Patent Rights, §27 at 45 (2005)
16. Re Cooper's Application for a Patent (1901) 19 RPC 53
17. Commissioner of Patents v Lee (1913) 16 CLR 138
18. Roger v Commissioner of Patent (1910) 10 CLR 701
19. IBM's Application, (IBM's Application, Re (1999) EPOR 318)
20. Singer & Singer European Patent Convention 112
21. Vicom Systems Inc's Application EPOR 74 (1987)
22. Davis, J.: *Intellectual Property Law*. Oxford University Press, UK (2012)
23. European Patent Convention 5 October 1973
24. Aerotel v Telco Holdings (2014) 1 all ER 225 67
25. Raytheon v Comptroller General of Patents, Designs and Trade Marks EWHC 1230 (2014)
26. Autonomy Corporation v The Comptroller General of Patents EWHC (2008)
27. Quest International v Odour Selection T619/02 (2015) OJ EPO 63
28. State Street Bank and Trust v Signature Financial Group Inc (Federal Circuit 1998) 141 F 3d 1368
29. Sohei (T/769 (1995) OJEPO 52)
30. Fisher, T.J., Signore, P.J.: An opposition to the recently proposed legislation related to business method patents. *J. Comput. Inf. Law* 397 (2002)
31. Japanese Examination Standards Office, Coordination Division Examination of business-related inventions (Dec 1999). Available at <http://www.jpomiti.go.jp/infoe/treatment.htm> (date of use: 19 June 2005)
32. Rai, R.K., Jagannathan, S.: Do business method patents encourage innovation? *BC Intell. Prop. Tech. F.* (2012)
33. Heines, M.H.: *Patents For Business*. Praeger Publisher (2007)
34. Masur, J.: Patent Inflation. *Yale Law J.* **121**, 470 (2011)
35. Business Methods and Madness: America's Patents System. *The Economist* (2008)
36. In re Schrader 22 F 3d (Fed Cir 1994) 298
37. USA Patents Act
38. Diamond v Chakrabarty 447 US 303 (US Supreme Court) (1980)
39. Nack, R., Nägerl, J.S.H., Walder-Hartmann, L.: The "Technical Invention" criterion. In: Haedicke, M.W., Timmann, H. (Hrsg.) *A Handbook on European and German Patent Law* (2014)
40. Stern, R.H.: Scope-of-protection problems with patents and copyrights on methods of doing business. *Fordham Intell. Prop. Media Ent. L. J.* 124 (1999)
41. AT & T Corp. v Excel Communications Inc. 172F. 3d 1352 (Fed. Cir. 1999)
42. In re Alappat 33 F 3d 1544
43. Amazon.com, Inc. v Barnesandnoble. Com. Inc. 239F. 3d 1343 Court of Appeals (Fed. Cir. 2001)
44. Bender, J.: Business Method Patents: The View from the USA. *EIPR*, p. 378 (2000)
45. Steuer, R.M.: Customer-instigated exclusive dealing. *Antitrust Law J.* **68**, 239 (2008)
46. Berkowitz, B.: Business Method Patents: Everybody Wants to Be a Millionaire. *Practising Law Institute*, vol. 36, p. 693 (2000)
47. Ostrow, S.H.: Is all this skepticism warranted? *New York Law J.* **39**, 7 (2000)
48. Dretfuss, R.C.: Are business method patents bad for business? *Santa Clara Comput. High Tech Law J.* 263 (2002)
49. Varela, S.L.: Damned if you do, doomed if you don't: patenting legal methods and its effect on lawyers' professional responsibilities. *Fla. Law Rev.* **60**, 287 (2008)
50. Posner, R.A.: The law & economics of intellectual property. *Daedalus* **131**, 5 (2002)

51. Raskind, L.J.: State Street Bank decision: the bad business of unlimited patent protection for methods of doing business. *Fordham Intell. Prop. Media Ent. L. J.* **10**, 67 (1999)
52. Lessig, L.: Death of cyberspace. *Wash Lee Law Rev.* **57**, 337 (2000)
53. Durham, A.L.: Useful arts in the information age. *BYU Law Rev.* 1419 (1999)
54. Sommer, J.H.: Against cyberlaw. *Berkeley Tech Law J.* 1145 (2000)
55. Gleick, J.: Patently Absurd. *New York Times Magazine*, p. 44 (2012)



*Thing: Improve Anything to Anything Collaboration

Giancarlo Corti^(✉), Luca Ambrosini, Roberto Guidi, and Nicola Rizzo

Institute for Information Systems and Networking,
University of Applied Sciences and Arts of Southern Switzerland,
Manno, Switzerland
giancarlo.corti@supsi.ch

Abstract. This is a work in the context of Collaborative Working Environment (CWE). In particular, in that of collaboration and productivity software tools. CWEs have seen the adoption of application software to addresses business problems as team communication and workload management. Instant messaging solutions, mobile devices and the virtual assistant paradigm have also come into the picture. Software tools in this context lack nonetheless real collaborative features. The problem that we address in this work is therefore that of a truly collaborating team collaboration and productivity software environment. Our approach leverages recent trends, like that of instant chats, virtual assistants and the Internet of Things, focuses on team members utterances and on a customizable and configurable bot framework to automate routine tasks, provide content over structure information management, and enable workflow management to improve productivity. The result is a prototypical software product which: enables the collaboration of both humans and Internet enabled things alike, provides easy context driven collaboration (i.e. entities graphs), allows the systematic processing of messages exchanged in a concurrent multi-user environment to fulfill team actions, provides a middle layer bot framework that handles the dialog flow for these actions and the interaction with any external systems. All of which distinguishes it from current software solutions. Given the features and the architecture of our original software components, we can confidently state that their adoption would enable software developers to create more effective collaboration and productivity working environment software tools.

Keywords: Collaborative working environment · Collaboration · Productivity · Instant messaging · Virtual assistant · Internet of things · Chatbot · Framework · Entithing · Story · Story manager

1 Introduction

Productivity is of paramount importance for the success of any enterprise [1]. For many of us (intuitively the majority) working means collaborating with others in some sort of organizational unit, most likely a team. Collaboration occurs

when two or more people or organizations (more generally: collaborating parties) work together to achieve a (shared) goal. It also means using a variety of tools, including software. Software tools essential to collaboration and productivity address business problems, such as tracking, task management, content and knowledge management, team communication, diversity, workload management and monitoring, connectivity and accessibility.

Recently, with the evolution of web technologies and the diffusion of mobile devices [2], we have witnessed the adoption of instant messaging solutions [3] and the advent of the virtual assistant paradigm [4], all of which contributes to modern CWE tools [5].

This complex reality of ideas, concepts and products, as a function of collaboration and productivity, is the context of this work.

This paper is organized as follows. In Sect. 2 we recap the state of the art in the vast reality of software tools and technologies in relation to the context. In Sect. 3 we outline the problem we want to address and in Sect. 4 we describe our approach to it. In Sect. 5 results are described in detail for each one of the key components making up the prototypical solution. We conclude in Sect. 6 by highlighting the limitations of our work as well as the opportunities we see it can open up.

2 State of the Art

Popular software products include collaboration and productivity tools, that: enable instant messaging like Slack or Stride; allow content management like Confluence or Dropbox Paper; allow project management like Trello. More focused tools also exist, like Google Keep, for example, for note taking, or Skype and Google Meet, for audio conferencing. All of this to list just a few.

Instant messaging tools usage often leads to information fragmentation, which makes it difficult to retrieve information and conversation regarding a specific topic [6]. To address these problems some tools introduced features like single message replies, threads and hashtags [7,8]. Instant messaging tools, in particular, can be extended by means of virtual assistants (i.e., chatbots) that assist in realizing predefined, repetitive tasks. Bots aim to improve team efficiency through task automation and effectiveness improving decision making, supporting team communication and coordination across tasks [9].

During 2016 chatbots had a huge rise in popularity which led them to be included in the MIT Technology Review as one of the ten breakthrough technologies of 2016 [10]. Major technology players released their own product. Amazon released Alexa. Apple released Siri. Facebook messenger chat bots API and many new players joined the scene [11].

Frameworks relevant software products include Botkit,¹ Rasa Core² and Dialogflow.³

¹ <https://botkit.ai/>.

² <https://core.rasa.com/>.

³ <https://dialogflow.com>.

3 The Problem

CWEs have seen the adoption of application software, which address many business problems (e.g., task management, team communication, workload management and monitoring, etc.). Instant messaging solutions, mobile devices and the virtual assistant paradigm have also come into the picture.

However, collaborative work management, workstream collaboration and content collaboration software tools or platforms are still very much isolated worlds when it comes to using them for team collaboration and to improve on productivity.

The bottom line is: if the tools that are to be used for collaboration do not work together all that well, how can they be assumed to be of any real help in a team's real work life? The problem that we address in this work is therefore that of a truly collaborative collaboration and productivity tool, whether at the level of a small three people? team or that of an entire organization.

4 Approach

Given the issues, we identified what we wanted was a truly collaborative working environment software tool to help productivity. We centered our reasoning on four aspects: *collaboration*, *communication*, *information management* and *content over structure*.

As far as *collaboration* is concerned, we included both Internet enabled things and virtual assistants, in addition to humans, as part of the collaborating parties (i.e., the team) definition. The idea behind this is that if we have to know whether or not a room is free for a meeting, the fastest way is to ask the room itself. If we have to define a task, we want the virtual assistant to help out with filling in the blanks and eventually save it in the right place for us. Let us make anything and also any virtual assistant part of the collaboration, then.

Turning to *communication*, all recent CWE software tools include messaging components to enable instant and lean communication between team members (i.e., instant chats). We therefore focused on the messages exchanged in such chats, which we interpret systematically to infer intents and to extract pieces of information that might be useful to automate work that would have to be carried out manually otherwise (e.g., scheduling an event and/or persisting it in some personal or shared calendar management application). As hinted above, the approach here is that of an original chatbot, that be perceived as being part of the team and truly collaborative. Able, for example, to suggest options, retrieve missing information or require some other member of the team to provide it.

With regard to *information management*, we want an approach that enables contextualization. That is, we want the pieces of information that make up the knowledge base of a given team to be linked and interlinked based on the context in which they are created so that they are easy to retrieve. And when they are, so too are all the others that are in the same context.

Finally, we want a tool that favors *content over structure*. Meaning that users should not waste their time finding relevant features and learn to use them on

yet another user interface arrangement. They should not waste time on routine manual operations either. *Content* should drive their work.

To put everything together, we want to develop a prototype as a proof of concept.

The next section will summarize the results we have obtained and will explain its concepts and software components.

5 Results

5.1 The Prototype's Macro Components

The prototypical collaboration and productivity solution we developed as a proof of concept is composed by the following macro components (see Fig. 1 for a structural diagram):

- **Frontend:** The frontend provides a web user interface to enable instant communication and to present content to its users. It includes an instant chat component, where team members, organized in channels, can exchange messages and a subset of commonly used widgets (e.g., calendar, task manager, etc).
- **Backend:** The backend manages the application logic and features a persistence layer to store and retrieve data. It exposes a websocket interface needed by the instant messaging chat and any external system that might be interested in them (e.g., the bot subsystem).
- **Bot subsystem:** The bot subsystem is a middle layer software framework. This component is connected to an instant messaging channel, acting like a chat client. It can interface with external systems for its needs, through configuration and customization. It currently supports an interface to a backend system (the backend manager), a list of user defined workers (that can take incoming messages and operate on them), and the context management engine (the story manager), that will be detailed later in this section.
- **Interpretation layer:** The interpretation layer currently leverages different approaches to text interpretation. It uses both Machine Learning (with Rasa NLU [12] for intent classification and entity extraction), Natural Language Understanding techniques (namely, SUTime [13] for time expression recognition and temporal tagging, NLTK [14] and spaCy [15] for tokenization and Named Entity Recognition) as well as some custom programmatic procedures (for simpler tasks, like deciding if a message is a yes or no answer). It was developed using a REST interface, so that it can easily be replaced with other implementations as long as the communication interface is preserved with the clients.

The prototype is demo ready as a self-hosted containerized web application. Each developed component is tested systematically and automatically.

In the following sections we will introduce the key concepts and the elements that characterize the bot subsystem and that are needed to understand what it does and how it works in more detail.

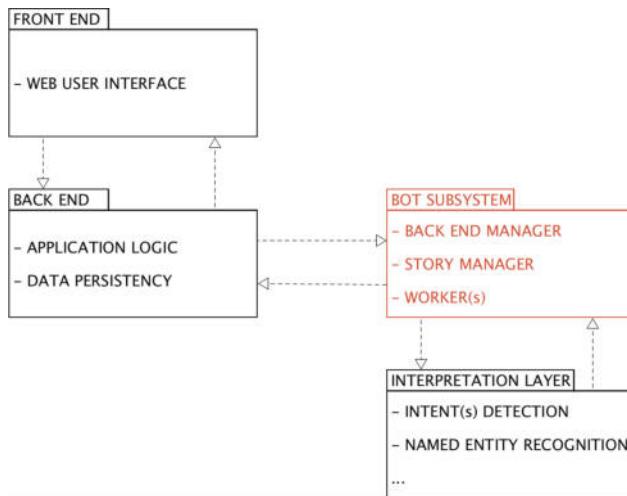


Fig. 1. Macro components

5.2 The Bot Subsystem

The bot subsystem is the core, framework oriented, component of this work. It can be controlled via an abstract controller class. Our prototype implements a concrete WebSocket interface that listens to a multi-user chat channel in the backend. In turns, to control any external backend, both to retrieve data and execute actions, an abstract backend manager class is provided, which in our case is implemented concretely as a REST interface. The logic of the bot is implemented by the story manager component, which is customizable and configurable and which will be described in detail later in Sects. 5.4 and 5.5. Finally, the bot provides an abstract worker class to implement concrete working agents, typically for text analysis and interpretation duties. The bot will use a list of workers of arbitrary length. Workers are configurable components that can be plugged and unplugged at runtime as needed. Figure 2 is an object oriented class diagram of the bot subsystem.

5.3 The “Entithing”

The first concept we want to introduce is that of the *entithing*. Wanting to provide the ability to treat all domain entities in a seamless fashion (e.g., a calendar event, a note, a message, a task, a reminder, a decision, or whatever else may be relevant for a specific domain) we were faced with the need of finding a suitable common label. We therefore coined the term: *entithing*, by contracting and compounding *entity* and *anything*. From now on in this paper we will therefore use this neologism to refer to anything from a chat message, to a filename, to an Internet enabled thing, or whatever may be part of the collaboration environment.

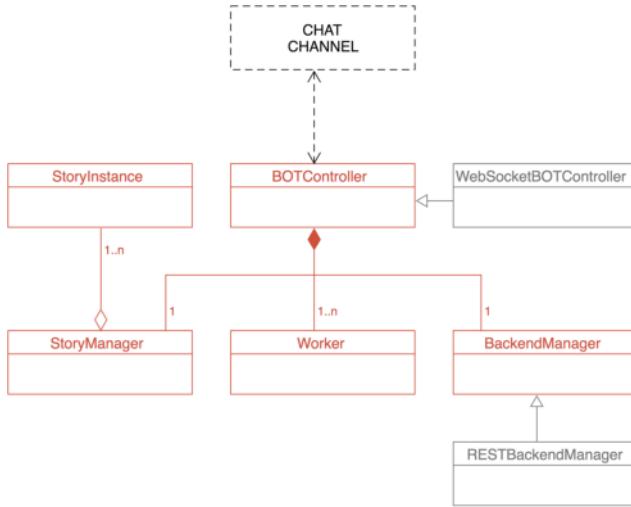


Fig. 2. Bot subsystem class diagram

5.4 The “Story”

We now introduce the concept of *collaboration story*, or more simply: *story*. A *story* represents any collaborating parties shared goals (e.g. canceling an event) and is the formalization of the information and the actions needed to realize it. A *story* can have to do with anything, from managing tasks, calendar events, decisions, etc., to managing more complex matters, like workflows or content (information in general) management. More precisely, a *story* is defined by:

- **Goal:** The eventual intent of the collaborating parties. For example: to book a room for a meeting.
- **Entities:** The pieces of information relevant to and/or needed to reach the story goal. For example: the list of participants to the meeting, its date, time and duration, the room number, etc.
- **Intents:** User intents interpreted by the interpretation layer that the story might need or want to react to. For example: the collaborating parties provide useful information for story completion (like a time indication for the meeting), or are unsure of something and a confirmation is required before proceeding, etc.
- **Completion actions:** What needs to be carried out especially in order to complete the story, when all the needed pieces of information cannot be gathered automatically. For example: retrieve the list of participation from a remote system, ask the collaborating parties the date, the time or the duration of the meeting, by posting a question in the team’s chat channel, etc.
- **Completed actions:** What needs to be done once the story is complete. For example: create an entry in each team members calendar, book a room through the company’s information system, publish a feedback by sending a message in the team’s chat channel or by sending out e-mails, etc.

From a technical point of view *stories* are defined and customized as templates using JSON configuration files (see Fig. 3 for an example of an event deletion). Our idea of *story* is inspired by that of Rasa Core.

```
{
  "uuid": null,
  "domain": "event",
  "name": "event_delete",
  "timeout": 600,
  "entities": [
    {
      "name": "title",
      "type": "text",
      "value": "",
      "filled": false,
      "action": "actions.utter:Utter",
      "after_action": "actions.notify_story_update:NotifyStoryUpdate",
      "args": [
        {
          "name": "utterance",
          "value": "What is this event's title?"
        }
      ]
    }
  ],
  "intents": {
    "entity_filled": {
      "action": "fill",
      "after_action": "actions.notify_story_update:NotifyStoryUpdate",
      "args": [
        {
          "name": "entity_name",
          "value": "title"
        }
      ]
    }
  },
  "after_class": [
    {
      "action": "actions.event_delete:EventDelete",
      "after_action": "actions.notify_story_update:NotifyStoryUpdate",
      "args": []
    },
    {
      "action": "actions.utter_event_deleted:UtterEventDeleted",
      "after_action": "actions.notify_story_update:NotifyStoryUpdate",
      "args": []
    }
  ]
}
```

Fig. 3. Event deletion story definition template

5.5 The Story Manager

The *story manager* is the heart of the bot subsystem framework and is the component that manages the *stories*. It loads the *story* templates, instantiates *stories*, tracks their status, acts according to their templates to retrieve needed entities (i.e., pieces of information), asks for missing ones if needed, executes all

the defined actions, including the final actions that realize the *story* goal (see Fig. 4 for an overview of its flow).

Every chat message is received by the bot subsystem and handed over to any workers that are enabled for message analysis and information extraction. The interpreted objects (i.e. the extracted pieces of information) are fed to the *story manager*, which either creates a new story instance or updates an already existing one. If it can, it uses the interpreted objects to fill as many information gaps as possible up front. Whenever a *story* is updated the *story manager* checks if it is complete (i.e., if all the required pieces of information are available). If it is, then the defined story completed actions (i.e. after-class) are executed. After-class actions may include user feedback actions, like posting a summary message back in the chat or sending out emails. Eventually, the *story* is removed from the list of the pending ones (although it is kept in the *story manager*'s history for any later needs). On the contrary, if it is not deemed as completed, and one or more entities are still missing, then the predefined completion actions are executed. These may include posting a message in the channel to ask users to provide the missing pieces of information.

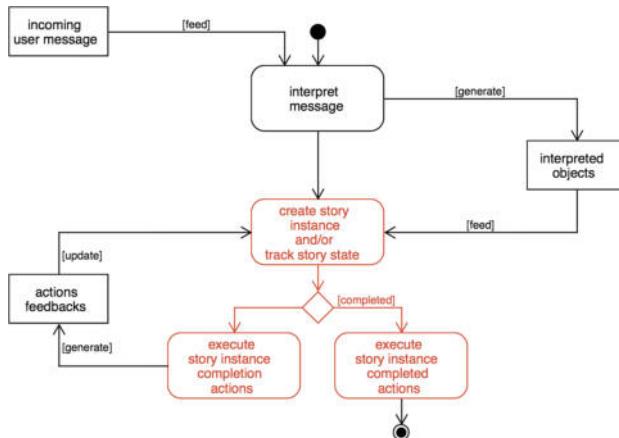


Fig. 4. Story manager flow overview

Our *story manager* work and configuration logic are inspired by that of programming languages test frameworks, such as Java's JUnit⁴ or JavaScript's Jasmine,⁵ in which callbacks are configurable before and after each test method as well as before and after each test class. In our case, user custom procedures are loaded dynamically (see Fig. 5), depending on the manager's execution step.

This component is currently written in Python and requires user custom procedures to be written in Python, too. The available *story* templates are

⁴ <https://junit.org/junit5>.

⁵ <https://jasmine.github.io/>.

loaded dynamically and can be hot swapped, meaning that they can be added or removed without stopping the service, thus zeroing downtime, in case of reconfiguration.

5.6 The “Focus View”

Building on the *entithing* and on the *story manager* and wanting to address the limitations of information management in current CWE software tools, we also introduce the idea of the *focus view*. This refers to the ability to persist the association of any *entithing* to any other and build a graph of linked *entithing* nodes. Graphs can then be retrieved by focusing on any one of their nodes.

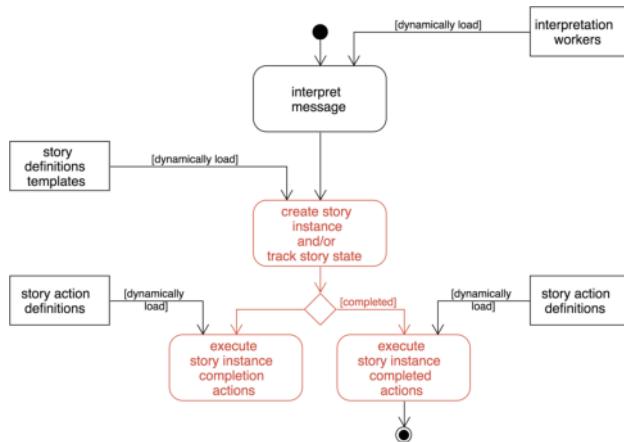


Fig. 5. Dynamically loaded story manager’s components

With this operation, we enter the so called *focus view*. This makes it very simple to focus for example on a domain *entithing* (let us say a task, like preparing the final scientific report for fooproject) and immediately retrieve the messages that were exchanged about it, the files that were attached to it, the calendar event that marks its deadline, and so on so forth. On top of this, while in this mode, any new content that is created is then automatically linked to the same context (i.e., graph). Figure 6 shows a context graph for the scheduling of a conference call, linked to the messages exchanged by the collaborating parties (two human users and a bot) to define and eventually save it into the team’s calendar application.

To achieve this, given an object oriented view of our data structure, we define a many-to-many reflexive association on the *entithing* class (see Fig. 7). Being an abstraction of any domain entity, such an association very simply, but also very efficiently, allows us to build the graph structures that are needed to support the *focus view*.

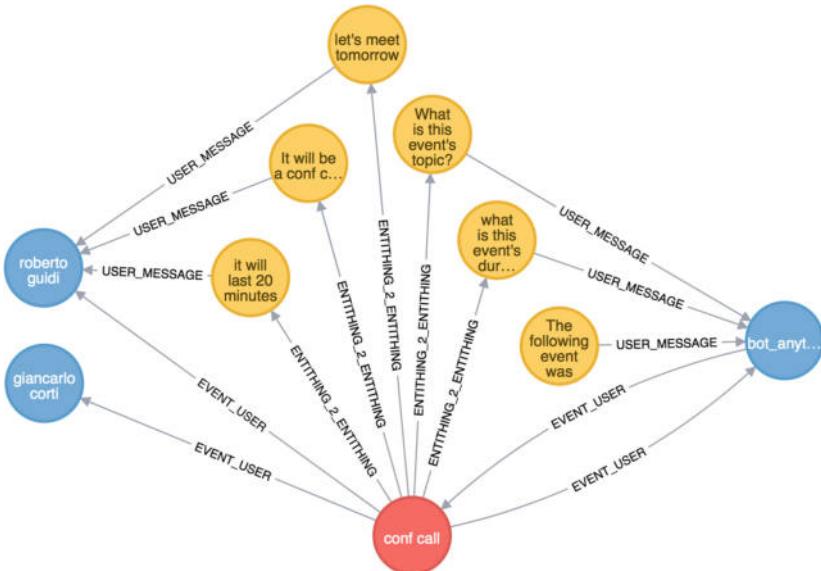


Fig. 6. Entithing context graph example

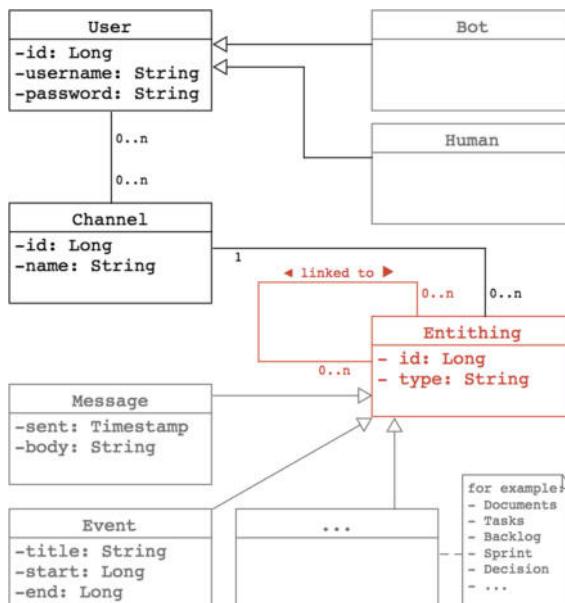


Fig. 7. Class diagram with entithing-2-entithing reflexive association

The strategies that allow identifying the need for a link between any two *entithings*, for now, use pragmatic or speculative approaches. For example, to link a document with the chat message used to upload it, with the user who sent the message, is straightforward. Message interpretation and some degree of inference ability is however required to link it to the project that might have been casually mentioned in one of the exchanged messages. This is done by focusing on the communication and therefore by analyzing the content of the message. If no inference is possible, the virtual assistant (exploiting the capabilities of our *story manager*) can always lend a helping hand and step in to ask users to provide complementary information to allow a useful link to be established, in a very pragmatic way.

For the implementation of data persistence in our prototype we used a Neo4j⁶ graph database.

6 Conclusions

Given the results, we are confident we are able to support the idea that more collaborative CWEs, that improve collaboration and help productivity, can be developed by extending the collaboration paradigm to Internet enabled things and by including a configurable and customizable virtual assistant as a collaboration party, rather than as a simple conversational agent.

In particular, the bot framework is a lightweight, yet powerful component, that can be plugged in already existing platforms and tailored to users needs. An enterprise personal assistant could be experimented by configuring and customizing it to interface with the rest of one's organization (an exemplary use case could be one in which I need to "tell my team I'm late" and I want my assistant to know who my team is and what it should do to inform them). Given this flexibility, it is a valuable product in its own right.

There are many limitations in this work we are very aware of. The biggest refer to how messages are associated to a given *story* and to the supported communication media. Messages are currently treated and interpreted in isolation and only pragmatic and speculative approaches have been used to attribute them to a given *story* (i.e., a context). Analytic approaches should also be applied here. These relate to thread detection [16–18]. It is particularly challenging as users' discussion contexts should be created and maintained live rather then with a posteriori analysis. Ellipsis phenomena are another important area of further work, to improve the ability of the bot to adapt to users and to improve the quality of its interpretations, especially because it works in a multi-user scenario.

Given our focus on communication, work is undergoing to include an e-mail agent to feed team members' e-mail messages into relevant chat channels, so that they can, in turn, be fed to the bot and profit from its collaborative features. Some tests have also been carried out to enable our chat with spoken language, through the use of Google Assistant APIs [19]. Both of these aspects merit further work.

⁶ <https://neo4j.com/>.

Acknowledgments. The authors would like to thank Salvioni SA, especially in the persons of Rocco Salvioni and Lorenzo Erroi, for their initiative, without which this work would not have been possible, Giacomo Poretti as our institute's industry relations and enabling person, and the Swiss Innovation Promotion Agency as the main financial sponsor.

References

1. Loveman, G., Management in the 1990s (Program): An Assessment of the Productivity Impact of Information Technologies, Series 90s (Management in the 1990s (Program)). Graduate School of Business Administration, Harvard University. <https://books.google.ch/books?id=VaJfGwAACAAJ> (1990)
2. Nguyen, T.H., Megan, R.: Hype cycle for mobile device technologies. Retrieved from Gartner database. Technical Report, vol. 4103 (2017)
3. Systems, R.: Enterprise instant messaging speeds up business productivity. <https://www.revesoft.com/blog/english/enterprise-instant-messaging-solution-speeds-business-productivity/>
4. Searle, S., et al.: How virtual assistants, immersive experiences and robots will impact your organization. Retrieved from Gartner database. Technical Report, vol. G00348690 (March 2018)
5. Wilson, N., Mike, W., Mann, K.J.: Critical capabilities for enterprise agile planning tools. Retrieved from Gartner database. Technical Report, vol. G00351645 (June 2018)
6. Insignia: Measuring the pain: What is fragmented communication costing your enterprise? <http://www.ucstrategies.com/uploadedFiles/Siemens%20UC%20Report%20for%20UCStrategies%20Com.pdf>
7. Slack: Slack message threads. <https://get.slack.help/hc/en-us/articles/115000769927-Message-threads>
8. Telegram: Reinventing group chats. <https://telegram.org/blog/replies-mentions-hashtags>
9. Storey, M.-A., Zagalsky, A.: Disrupting developer productivity one bot at a time. In: Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering, pp. 928–931. ACM (2016)
10. Dale, R.: The return of the chatbots. Natural Lang. Eng. **22**(5), 811–817 (2016)
11. Data Monster: 25 chatbot platforms: A comparative table. <https://chatbotsjournal.com/25-chatbot-platforms-a-comparative-table-aeefc932eaff>
12. Rasa NLU: T. GmbH. Language understanding with Rasa NLU. <https://nlu.rasa.com/>
13. SUTime: Stanford. <https://nlp.stanford.edu/software/sutime.html>
14. NLTK: Natural language toolkit—NLTK. <https://www.nltk.org/>
15. spaCy: Explosion AI—industrial-strength natural language processing. <https://spacy.io/>
16. Shen, D., Yang, Q., Sun, J.-T., Chen, Z.: Thread detection in dynamic text message streams. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, pp. 35–42. Washington, USA, Seattle (2006)
17. Page, H.A., Craig, M.H.: Topic detection and extraction in chat. In: 2008 IEEE International Conference on Semantic Computing. Santa Clara, CA, USA (2008)

18. Traum, D.: Issues in multiparty dialogues. In: Dignum, F. (ed.) Advances in Agent Communication, pp. 201–211. Springer, Berlin, Heidelberg (2004)
19. Google: Google assistant SDK for devices. <https://developers.google.com/assistant/sdk/overview>



eSense 2.0: Modeling Multi-agent Biomimetic Predation with Multi-layered Reinforcement Learning

D. Michael Franklin^(✉) and Derek Martin

Kennesaw State University, Marietta, GA 30114, USA

mfranklin@kennesaw.edu

<http://ksuweb.kennesaw.edu/~dfrank15/>

dmart133@students.kennesaw.edu

Abstract. Learning in multi-agent systems, especially with adversarial behavior being exhibited, is difficult and challenging. The learning within these complicated environments is often muddied by the multitudinous conflicting or poorly correlated data coming from the multiple agents and their diverse goals. This should not be compared against well-known flocking-type behaviors where each agent has the same policy; rather, in our scenario each agent may have their own policy, sets of behaviors, or overall group strategy. Most learning algorithms will observe the actions of the agents and inform their algorithm which seeks to form the models. When these actions are consistent a reasonable model can be formed; however, eSense was designed to work even when observing complicated and highly-interactive multi-agent behavior. eSense provides a powerful yet simplistic reinforcement learning algorithm that employs model-based behavior across multiple learning layers. These independent layers split the learning objectives across multiple layers, avoiding the learning-confusion common in many multi-agent systems. We examine a multi-agent predator-prey biomimetic sensing environment that simulates such coordinated and adversarial behaviors across multiple goals. This work could also be applied to theater wide autonomous vehicle coordination, such as that of the hierarchical command and control of autonomous drones and ground vehicles.

Keywords: Artificial intelligence · Multi-agent systems · Strategy · Hierarchical reasoning · Predator-prey

1 Introduction

Real-world artificial intelligence and learning is often made more difficult by the various goals that each agent has and the complex interactions between agents within a system [12]. This is especially true in multi-agent systems where the desired interactions take on a strategic, intelligent meaning. The normal approach of using monolithic policies is rendered ineffective because of having

multiple behaviors for each agent, some of which frequently conflict, and exacerbated by the multiple teams of agents which are in direct conflict. In this case, the system being modeled in the simulation is the predator/prey dynamic. To do so, each agent is considered a biomimetic model of a sensing agent. This means that each agent has its own ‘personality’ - a methodology of movement, a set of goals to seek, other agents to avoid, etc. Further, each agent has a sensing grid - this grid is how the agent sees the world around them. This might be modeled as passive echolocation, active electrolocation, or any of several sensing models previously presented in the original eSense paper [6]. Expanding on that, there are additional sensing layers that are seeking and avoiding goals and other agents, depending on the model in force. Additionally, the new model includes multiple biomimetic models interacting within the environment.

eSense 2.0 takes the single-agent success and builds on it to add multiple layers of perception. This increases the performance of each agent and increases the level of biomimicry. eSense 2.0 also places these much more capable agents into multi-agent scenarios and allows interactions on multiple levels. The experiments will show the increased expressivity, higher fidelity modeling, and the multi-agent interaction capability of the system.

2 Related Works

This new extension of eSense, originally proposed here [6], adds multiple layers of reasoning for each agent and allows for multiple agents within the environment.

Working within complex, multi-agent, adversarial environments has proven difficult, as shown in [2]. Here there is ample evidence of the compounding nature of complexity escalation involved in these challenging systems. eSense was developed to mitigate, if not eliminate, this exponential growth of complexity by offering split-levels of learning, adaptation, and execution.

In [9] the authors have presented a mathematical analysis on predator/prey relationships and their interactions within shared environments. This interaction can be understood in terms of the pressure each set of animals places on each other from their presence as relates to their distinct populations and amount of shared environment. This work was helpful in modeling these interactions within eSense and understanding a mathematical modeling of these relationships.

The interaction of predators and prey, especially understanding the balance of their populations and level of their interactions within the same ecosystem, is modeled thoroughly in [7]. This work, though older, is still cited as a reference on the concept of persistence (where persistence is defined as a greater than zero population now and at the limit). This work contributed the idea of balance of populations and the concept of persistence to this research. It is important to note that the referenced paper works in a deterministic environment, so the work proposed herein expands beyond this consideration into both non-deterministic and stochastic space.

One important intuition that was confirmed in [10] was that of understanding the difference between isolating predators from prey and analyzing them

separately versus studying and understanding them in the proper real-world context. This more accurate context allows for coordinated and reactive strategic behavior from the predators while tracking the prey. There are also two additional works inside this paper that elucidate the complexities of these interactions more clearly. They note first that the multi-trophic games of habitat choice impacts this kind of real world interaction significantly. Secondly, they note that the scale of these interactions matter. While this is not surprising, it is important to have confirmed empirically. This research proposed and confirmed in this paper understands both of these ideas and seeks to model them appropriately (i.e., with the correct level of expressivity and complexity to allow this kind of large-scale strategic interaction).

In [15], there is a lengthy treatise on the complexities that arise from multiple populations of predators and prey existing in the same interactive environment. The exhaustive analysis of the Hopf bifurcation and multiple steady-state bifurcations is especially informative in understanding the pressure concepts of multiple agents occupying a competitive space. For the work contained in this research, this paper offered a thorough mathematical analysis of competitive-cycle dynamics of these interactions that are insightful and illuminating. This research builds on the premises and expands them to stochastic and non-deterministic scenarios like those found in the experiments conducted in the simulations to prove the eSense 2.0 expanded models.

There were many models used in the previous work that were now revisited for information about the expanded biomimetic modeling done in eSense 2.0. [1] offers a treatise on the modeling of electric fish for experimentation and simulation in multi-agent scenarios. These models were expanded by comparing this previous work with [3]. This additional reference provided more detail on the types of models available and insight into passive and active models for both types of target location. Additionally, [8] offers specific information for modeling passive electrolocation and understanding how it is used in the real world. This work was essential to ground the symbols for each experiment and to ensure experimental veracity. The earlier models and final experimental models were enhanced by using the information from [11]. Each of these contributed to the modeling of the fields utilized by both the active and the passive location models.

The integration of the various models and bringing them from real-world, biological models, into simulated entities within a reinforcement learning environment was aided significantly by the work in [4]. This work offers some background insight into how others have approached these types of learning environment models. In particular, this work studies exploration and exploitation in reinforcement learning, and, vitally, the balance needed for them both to be effective. This work helped verify the need for utilizing the ϵ -greedy approach utilized by this research. In the proposed work the exploration and exploitation are balanced progressively during the execution of the algorithm. Initially, with ϵ high, the algorithm leans towards utilizing more exploration to explore the state-space more thoroughly. Eventually, over time, as ϵ decreases, the algorithm utilizes more exploitation. While there were no other papers found that

have applied a similar model as the one proposed herein, this paper did at least offer insight into the validity of the approach.

The concepts of Reinforcement Learning, of which both Temporal Difference learning and SARSA- λ are examples, were described in [13, 14]. These works were utilized to confirm the models used for the learning systems for these experiments and to gain insight into common settings for both approaches. While these referenced works propose and define various aspects of reinforcement learning, they do not propose anything similar to the multi-layered dynamic approach described in this research. Further, they allude to why this particular goal, a dynamic reward, is a non-starter for reinforcement learning (that is, convergence is statistically improbable).

3 Methodology

This new eSense modeling builds on the previous work, referenced in the abstract, and expands on it significantly. In the previous work there were a number of innovations that led to the overall success of eSense, and those will be summarized here for clarity. For complete background, please review [6].

Originally, eSense innovated the idea of taking the simple-yet-powerful reinforcement learning technique SARSA- λ and utilizing it in creative ways to accomplish complicated learning. In particular, it is well known from [12] that SARSA- λ is a clean, efficient minimal information learning technique for reinforcement learning, but it does not work if the goal is moving. A moving goal essentially erases all learning in the Q-table because of the history contained within the grid. For clarity, assume, *w.l.o.g.*, that the learning in the SARSA- λ is Q-learning, and this learning uses two distinct tables for tracking the progress of the learning. Standard SARSA expands the typical Q-learning into a Q(s, a) table that stores the values of taking any move from the current state (i.e., Q holds values of each a for and given s). This Q(s, a) table is then consulted any time the agent is preparing to move to select the next action based on one of two options. In standard ϵ -decay technique, the next move is selected pseudo-randomly with ϵ probability and by max value with $(1 - \epsilon)$ probability. This helps the agent to explore more often early on and exploit the learned data more frequently as the learning progresses. The second table utilized in the SARSA- λ variant is the e-table. This second table holds a type of memory of states visited since the epoch has begun and allows for a longer history of updates to the Q(s, a) table to be made each iteration. Typically only the previous states would be updated, but a decaying reward can be effectively propagated back along the entire path of moves by utilizing the e-table data. This method of updating in a typical grid world example means that as long as the goal is stationary, the Q(s, a) table will eventually hold a policy that offers the best move from any given state, thus achieving a minimal pathing from the origin point to the goal. However, as mentioned, if the goal were moved each epoch, not only would the Q(s, a) table no longer lead to the goal, it would actively lead away from it. If, for example, the goal were moved only once, then the learning could eventually

repair the table to point towards the new goal location, but it would take much longer than it did the first time because of its bias to the old goal location.

To overcome this, the eSense technique was devised. As presented in the reference paper, this limitation was removed through layering the behaviors into multiple grids. eSense works on multiple levels of learning through the use of a master grid for obstacle detection and avoidance and another layer for sensing (i.e., examining what is around the agent and reacting to that rather than using the master grid). This means that the agent can wander around the master grid and learn obstacle avoidance without worrying about goals. The sensing layer can then be utilized for goal-seeking. The sensing layer is homeostatic, centered around the agent. As the agent searches the grid, using learned data in the master grid to avoid obstacles, the goal eventually moves within the sensing grid. This triggers the learning in the sensing grid to react to the presence of the goal. The key intuition within this technique is that the learning is identical, but the action set is reversed (this can be thought of as moving the goal towards the agent - which is not possible, but it informs the agent's direction of movement). This was a key innovation of the eSense methodology.

Once this technique was proven, eSense went even farther by allowing for a moving target. In the original formulation the goal was stationary, just placed randomly around the grid. In the final formulation, the moving goal became a prey and the agent became a predator. The sensing grids were converted to reflect the various biomimetic models (both passive and active echolocation and electrolocation). This means that the sensing grids had differing sizes and shapes. As the predator searched the master grid the prey would follow a pattern of movement dictated by the program. When the prey entered the predator's sensing grid the predator would react to move towards the prey in an effort to catch it. To be clear, this was not programmed behavior - the predator learned these behaviors from scratch using the simplistic SARSA- λ reinforcement learning without any prior knowledge. This is a significant outcome and the novel contribution of the original eSense paper.

eSense 2.0 expands significantly on this multi-layered learning methodology to include even more layers with additional agents in the system. First, the prey is now an agent. The prey has its own obstacle avoidance master grid that is learning to avoid edges, obstacles, and other obstructions. The prey also has another master grid that is marking locations where food has been found (the food is the prey's goal and can be located at any number of stochastic locations around the grid). Additionally, the prey has two sensing grids. The first sensing grid is designed to detect and react to food. When food appears on this sensing grid, the agent learns through trial and error to seek after the food. The second sensing layer is the predator avoidance layer. This sensing layer detects when predators are within range and learns to avoid them (this is the same learning technique, but with the opposite set of actions). This multimodal learning is difficult for traditional agents because trying to learn a large, complex monolithic policy is both contradictory (learning to move towards and away from goal objects) and confusing (clouding up the learning with contrary

goals and opposing actions) [5]. The new multi-layered approach presented in eSense 2.0 allows for less complex learning techniques with single-goal objectives, thus overcoming this learning complexity and confusion. Second, the predator also has the multi-layered approach. As with the prey, the predator has its own obstacle avoidance layer. This could be the same master layer as the prey, but by giving each agent their own obstacle avoidance layer each agent's size can be considered independently. For example, a smaller prey can slip through a smaller opening in the obstacle field that a predator cannot. This individualized behavior is an important part of the eSense methodology. Further, the predator also has an additional master layer to track the most likely places to find prey as well as two sensing layers. The first sensing layer is seeking prey (its food source) while the second is learning to avoid other predators. This configuration allows for an entire hierarchical ecosystem of predators and prey, as well as allowing for multiple agents within each layer.

Each layering the agent model is performing SARSA- λ , though with different ranges and setups. Each layer is learning according to the update function shown in Eq. 1. This updates the $Q(s, a)$ table by utilizing the reward r for moving to the next state, the next values provided from taking the chosen next action (a') from the next state (s') (stored as $Q(s', a')$). It is mitigated by the learning rate, α . The algorithm for the updates and the movement tracking history is shown in Fig. 1. This shows the step by step updates shown in Eq. 2. The update amount, the δ , is calculated in Eq. 3. The e-table is incremented for every space that is visited, according to Eq. 4. The decaying updates in the e-table are updated according to Eq. 4 using the discount rate γ and the decay rate λ . This results in an eligibility trace (a history of decaying rewards based on the previously visited, and, thus, eligible spaces that can receive an update/reward). These traces are similar to those shown in Fig. 2.

$$Q(s, a) = Q(s, a) + \alpha(r(s', a') + \gamma Q(s', a') - Q(s, a)) \quad (1)$$

```

Initialize  $Q(s, a)$  arbitrarily and  $e(s, a) = 0$ , for all  $s, a$ 
Repeat (for each episode):
    Initialize  $s, a$ 
    Repeat (for each step of episode):
        Take action  $a$ , observe  $r, s'$ 
        Choose  $a'$  from  $s'$  using policy derived from  $Q$  (e.g.,  $\varepsilon$ -greedy)
         $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$ 
         $e(s, a) \leftarrow e(s, a) + 1$ 
        For all  $s, a$ :
             $Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$ 
             $e(s, a) \leftarrow \gamma \lambda e(s, a)$ 
             $s \leftarrow s'; a \leftarrow a'$ 
    until  $s$  is terminal

```

Fig. 1. SARSA- λ algorithm [12]

$$Q(s, a) = Q(s, a) + \alpha \delta e(s, a) \quad (2)$$

$$\delta = r(s', a') + \gamma Q(s', a') - Q(s, a) \quad (3)$$

$$e(s, a) = e(s, a) + 1 \quad (4)$$

$$e(s, a) = \gamma \lambda e(s, a) \quad (5)$$

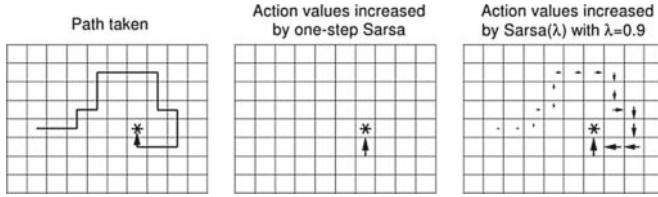


Fig. 2. SARSA- λ eligibility traces [12]

As can be seen from this formulation, each layer described above is actually composed of multiple layers (the $Q(s, a)$ table and the e -table). This means that each layer can be learning on its own independently. Each of these layers is small enough to learn quickly and is focused on only one aspect of the agent's behavior, so the monolithic policy can be avoided and replaced with smaller policies customized for each layer. This arrangement of layers means that there must be one additional layer, the agent layer, that controls the focus of the agent across these multiple layers. The layers are arranged in a hierarchy, as shown in Fig. 3. The agent layer sits at the bottom of the hierarchy and organizes the behavior of each agent by receiving the fusion of the sensor layers. For example, the prey agent layer is constantly running the baseline obstacle avoid layer (meaning that it considers all higher actions with respect to the base action of avoiding obstacles). Additionally, it is adding information to its food location layer each time it finds food. Of course, when food is sensed on the food sensing layer (generically, the goal seeking layer), it reacts to pursue that food. Finally, the highest priority layer is the predator avoid layer. This means that the agent is constantly wandering the master grid avoiding obstacles and seeking food. When it finds food, it notes that location and seeks after it. Suppose a predator is sensed - now the agent layer shifts priorities to moving away from the predator, but considers all of the lower actions. In other words, it will move away from the predator, but towards food, all while avoiding obstacles. It also notes what it learns, (e.g., where it sensed predators or food, both stored on their own history layers). Again, to reiterate, this behavior is not programmed in - the agent is given no information ahead of time other than that food receives a positive

reward and dying a negative reward. The agents are learning from scratch with no other information than the rewards given. Each layer is able to adapt and learn quickly because the layers are separated as described.

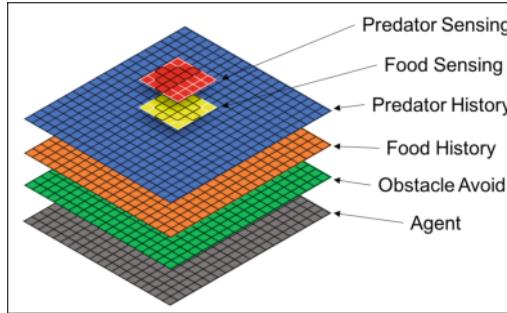


Fig. 3. eSense layers

The information from each layer is fused into a best action for the agent in a vector fashion and stored on the agent layer. As can be seen in Fig. 4, the input from each layer is laid out on the graph with both a direction and a magnitude. The magnitude is the weighting for each layer, and this can be provided as a model (in the case of biomimicry) or it can be learned over time by experience. The sum total of each layer's input is the resultant vector on the agent layer, shown in Fig. 5. The resultant vector is discretized into the action that most closely matches the intention of the resultant vector and the action is chosen. When the weighting is correct, the optimal performance of layer-fusion is obtained and the behavior maximizes the most important choices while being mindful of all choices. This can also be thought of as maintaining a primary goal (e.g., survival), while operating on the sub-goals (e.g., feeding). This complicated, modeled behavior is being achieved with several simple layers rather than with large, monolithic and unwieldy layers that would take a long time to learn and be difficult to adapt over time.

The smaller sensing layers are shaped in accordance to the biomimetic models upon which they are based (e.g., electrolocation or echolocation) as well as the modality of sensing (e.g., active or passive). Two of these shaped sensing layers are shown in Fig. 6. The sensing grids are homeostatic, meaning that they stay centered on the agent (whose location is marked within each of these grids). These grids can be shaped to model any reasonable type of sensing array, or, more generally, to resemble any type of goal-seeking apparatus. In any case, the sensing layers are fused to the agent so that all available layers can send their data to the agent layer for processing.

Once a goal condition is encountered within a sensing grid (e.g., a food source, a predator, etc.) the agent layer can then process the appropriate action to maneuver the agent towards or away from the goal. As stated previously, the

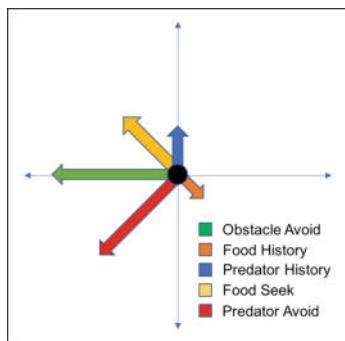


Fig. 4. Sensing layer fusion

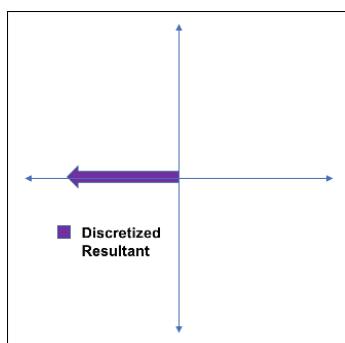


Fig. 5. Agent layer resultant

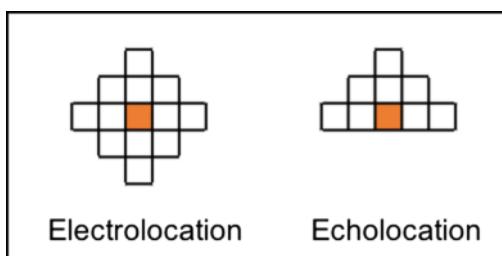


Fig. 6. Sensing grids (agent location noted)

sensing layers work in reverse because the agent cannot move the goal, so the appropriate action is considered as if the goal were movable, then the reciprocal action is taken. For example, if food were detected, the agent would want to move the food towards it, but it cannot. Instead, it takes the inverse action and effectively moves the sensing grid towards the goal. The learned action becomes to move towards food and away from predators, or, more generally, towards or away from goals.

4 Experiments

In order to test these hypotheses, there were a number of experiments conducted, and they will be noted in this section. The first was to set the prey agent in action to see how well it could learn to: (1) avoid obstacles; (2) find food; (3) learn likely food locations; (4) adjust its wandering pattern in response to likely food locations. This first experiment was successful. The prey agent learned quickly to avoid obstacles efficiently using the prescribed reinforcement learning algorithm (meaning that it started with no information other than the goal rewards, when encountered). Figure 7 shows this for both the prey and the predator. It also learned to locate food and move towards it, though this learning took a bit longer because there are multiple goals (meaning that the agent had to wander enough to discover the other food locations). Figure 8 shows the prey's success at finding food, increasing over time, versus it being caught by the predator, slightly increasing over time. Once this behavior was learned, the wandering pattern of the agent became more centralized near food sources, as was hypothesized.

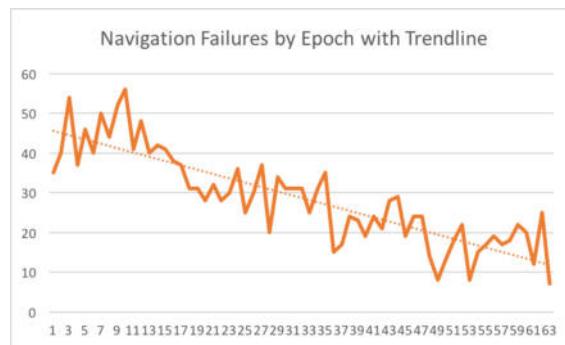


Fig. 7. Navigation failures by agents

The second experiment was akin to the first, but with the addition of a predator. The predator was simultaneously learning obstacle avoidance, food locations (i.e., the most likely locations of the prey), and predator avoidance. Of course, the introduction of a predator into the environment activated the prey's predator avoidance layer. This resulted in a successful migratory pattern for

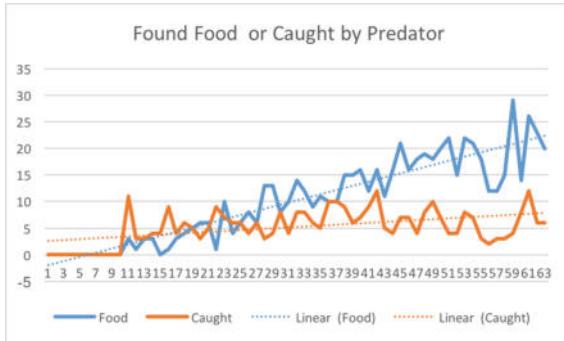


Fig. 8. Prey: found food versus caught by predator

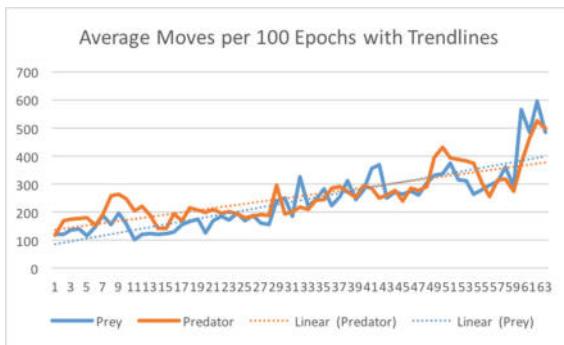


Fig. 9. Average moves without failure

the predator who learned to localize on the prey's food sources. It also slightly modified the prey's routine to learn to avoid the most likely predator locations, though this learning took longer. Figure 9 shows the average number of moves per epoch for both agents. In the end, the experiments proved successful in modeling a biomimetically accurate predator and prey relationship.

The third experiment built on the second experiment by introducing multiple prey into the environment. While this still followed the predictable results (the predator now learned a more general migratory pattern to adjust to the multiple locations where prey can be found), it was only a stepping stone to multiple predators and multiple prey. This finalized the progression of the experiments and showed that the prey can learn to distribute themselves across the food sources, predators can spread out to maximize available prey to each, and both prey and predators can avoid their own kind. Figure 10 shows that the addition of multiple agents show did not affect performance, and is thus scalable (the trend line is nearly identical to the single agent average moves graph).

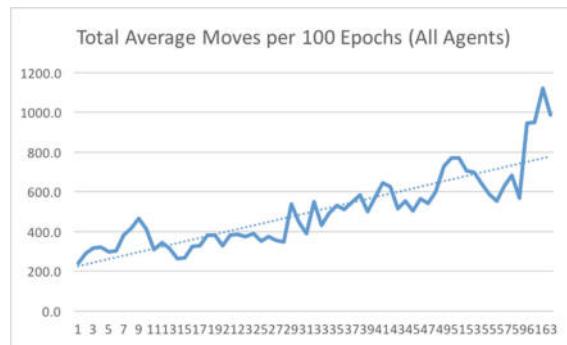


Fig. 10. Total avg. moves without failure (all agents)

5 Analysis

The experiments proved the efficacy and efficiency of the multi-layered predator/prey biomimetic modeling. Further, they verified that complex, intricate, and multi-agent behavior can be learned through even simple reinforcement learning as long as the various behavioral elements are spread across multiple coordinated layers. By keeping each layer simple and focused, the agents were able to learn multiple goals at one time (e.g., finding food sources while avoiding prey and obstacles) without a significant increase in training time. The biomimetic models were further expanded to include multiple behaviors, though this can be expanded in the future. There was a lot of experimentation with the size and shape of the sensing grids and how that impacted the learning, but it was discovered that while these helped demonstrate different models they had no significant impact on learning rates.

6 Conclusions and Future Work

This work has shown that biomimetic modeling can be realized through simple, multi-layered learning techniques. Additionally, the experiments verified that multi-agent interaction, even with teams of agents, works well without significantly slowing down the learning. It should be noted that the introduction of more prey or more predators once the learning has advanced may cause disruption and instability, but the learning can adapt. This will be tested in greater detail in future work. Finally, in future work the hypotheses will be expanded to include a larger food chain (where predators have predators). Also, there is the hope to introduce group behavior versus lone wolf behavior to see if this can be modeled effectively and, if so, what impact it has on learning.

In conclusion, this work has shown tremendous promise from its simple beginnings and has become more robust through expansion. It is the author's hope that this will continue to be true with further expanded experimentation and more complex modeling.

References

1. Ammari Habib, T.B., Garnier, J.: Modeling active electrolocation in weakly electric fish. *SIAM J. Imaging Sci.* **6**(1), 285–321 (2013)
2. Batista, G.E.A.P.A., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newslett.* **6**(1), 20–29 (2004). <https://doi.org/10.1145/1007730.1007735>
3. Boyer, F., et al.: Model for a sensor inspired by electric fish. *IEEE Trans. Robot.* **28**(2), 492–505 (2012)
4. Coggan, M.: Exploration and Exploitation in Reinforcement Learning, 3(3), p. 1448. CRA-W DMP Project at McGill University, Scholarpedia (2008)
5. Franklin, D.M.: Strategy inference in stochastic games using belief networks comprised of probabilistic graphical models. In: Proceedings of FLAIRS (2015)
6. Franklin, D.M., Martin, D.: eSense: BioMimetic modeling of echolocation and electrolocation using homeostatic dual-layered reinforcement learning. *Proc. ACM SE* **2016** (2016)
7. Freedman, H., Waltman, P.: Persistence in models of three interacting predator-prey populations. *Math. Biosci.* **68**(2), 213–231 (1984)
8. Hopkins, C.D.: Electoreception: Passive Electrolocation and the Sensory Guidance of Oriented Behavior. Springer, New York (2005)
9. Hussein, S.: Predator-prey modeling. *Undergraduate J. Math. Model.: One + Two* **3**(1), 32 (2010)
10. Lima, S.L.: Putting predators back into behavioral predator-prey interactions. *Trends Ecol. Evol.* **17**(2), 70–75 (2002)
11. Shieh, K.T., et al.: Short-range orientation in electric fish: an experimental study of passive electrolocation. *J. Exp. Biol.* **199**(11), 2383–2393 (1996)
12. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, Chaps. 4, 5, 8 (1998)
13. Taylor, M.E., Whiteson, S., Stone, P.: Comparing evolutionary and temporal difference methods in a reinforcement learning domain. In: Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation. ACM (2006)
14. Woergoetter, F., Porr, B.: Reinforcement learning. *Scholarpedia* **3**(3), 1448 (2008)
15. Yi, F., Wei, J., Shi, J.: Bifurcation and spatiotemporal patterns in a homogeneous diffusive predator-prey system. *J. Differ. Equ.* **246**(5), 1944–1977 (2009). <https://doi.org/10.1016/j.jde.2008.10.024>, <http://www.sciencedirect.com/science/article/pii/S0022039608004373>



System of Organizational Terms as a Methodological Concept in Replacing Human Managers with Robots

Olaf Flak^(✉)

Faculty of Radio and Television, University of Silesia in Katowice,
Katowice, Poland
olaf.flak@us.edu.pl

Abstract. Although IT systems fill and automate more and more areas of human life, and thus, manager's work, one can ask why, at the end of the second decade of the 21st century, one cannot hire a robot as a manager? The paper presents the reasons for such a gap in applying algorithms and robots to business life and the possible solution to this problem. The reasons are the methodological problems in management sciences such as H. Koonst's "theory jungle", large subjectivity in theories, "overproduction of the truth", chaos in definitions and scientific language, building "islands of knowledge" instead of developing a stable model of reality. The solution for these obstacles in building real knowledge on manager's behavior, which is the necessary foundation of work automation, is the system of organizational terms. This is a methodological concept of research introduced by the author together with original research tools which are on-line management tools at transistorshead.com. Both innovations in management research let conduct several research project on manager's behavior. There were also attempts of recognizing patterns in manager's behavior. The examples of results are presented in the paper.

Keywords: System of organizational terms · On-line management tools · Manager's behavior · Work automation · Team management automation

1 Introduction

The last over a dozen years have been the time of rapid development of information technology and robotics, as well as a process of replacing people's work with machines or algorithms. Also managers commonly work with electronic tools that facilitate their work, register it and indirectly register the operation of their organization [1].

There are more and more publications about the vision of replacing a manager with computer software and, as a result, creating robot managers [2]. It should be emphasized that already in 1967, P. Drucker wrote that computer systems (then—"computers"—author's note) would not only serve to collect information, but the algorithms written in them would be able to replace managers over time [3]. Despite the passage of several decades, this has not happened so far, although IT systems fill and automate more and more areas of human life, and thus, manager's work. So one can ask why, at the end of the second decade of the 21st century, one cannot hire a robot as a manager?

There seem to be several conditions that have not yet been met sufficiently to allow this to happen. The unfilled conditions are:

- predictability of behavior of people who cooperate with one another, in this case the manager-robot and participants of the organization [4],
- the ability of the robot manager to exert real influence on the participants of the organization and vice versa [5],
- the existence of a common basis for communicating knowledge about organizational reality between the manager-robot and participants of the organization (this is not just about the language used to communicate—author's note) [6].

However, the basic problem in replacing a manager with a robot seems to be the lack of a unified scientific research methodology—the basis for meeting the above conditions—in building reliable knowledge about the behavior of managers [7].

Therefore, the aim of the paper is to present an original methodological concept called the system of organizational terms which let collect information on manager's behavior and describe some patterns necessary in replacing human managers with robots. It will let in the nearest future replace human team managers with robots and develop the performance of organizations and companies.

The system of organizational terms as a methodological concept in management sciences is, in the intention of the author, a way to meet the above conditions by unifying the individual areas of the organizational reality research. Its task is to perform a similar role as the SI system in the case of the automation of physical phenomena [8].

In Sect. 2 of the paper it is described the methodological problems in building knowledge on manager's behavior and methodological and metrological solutions to these problems based on the concept of the system of organizational terms. In Sect. 3 there is a description of preliminary research on manager's behavior aimed at application of artificial managers.

The proposed managerial action representation, designed in the system of organizational terms, creates the first step to design several applications for the business practice. Basing on our experience in working with managers and companies it can be seen at least three such applications.

Firstly, the management patterns are usual practical problem in big companies where the managers, i.e. team leaders or project managers, should work with external or internal clients delivering them specified results. Especially big companies put a lot of effort to standardize employees' work and their results which can be automated and done by algorithms. As it is shown on Fig. 1, some of managerial actions could be managed by an artificial team manager. In other cases some advice could be given to human managers about their typical habits or previous actions.

Secondly, the next application concerns the recruitment process to any organization. In some cases of managerial position there is organized assessment center for candidates. It is usually a place with a conference room where a few candidates have to solve a problem together during 2–3 days. They are observed by HR specialist, psychologists, experienced managers etc. However, the results of such observations are quite qualitative and vague. If the candidate could work not only with sheets of paper but also some online management tools which would record his actions, there would be an opportunity to assess how much their style of management fit to the requirements on

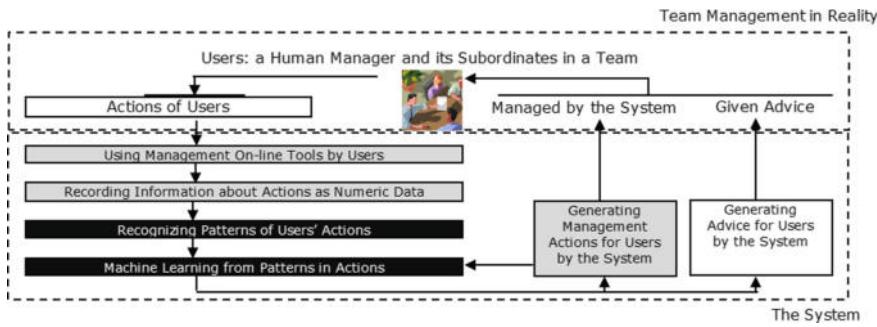


Fig. 1. The cycle of team management automation

the vacated positions. This assessment would much more precise and detailed than done by a traditional assessment method. In the end can be done automatically.

Thirdly, the system of organizational terms introduces a model of managerial actions in any situation and of any kind such setting goals, describing tasks, generating ideas, creating options, checking motivation etc. Recording managerial actions based on this representation method is essential to get data useful in concluding on the dominating habits of a certain managers and his behavior. This lets build a model of his personal managerial style and to apply this model when we want to pretend such a manager in a real life. This leads inevitably to replacement of human managers with robots in certain types of managerial work, such as setting goals and planning tasks, preparing meetings, checking motivation of a team etc.

2 Knowledge of Manager's Behavior

2.1 Methodological Problems in Building Knowledge on Managers' Behavior

The discussion about the identity of management sciences and their methodological assumptions dates back to 1961, when H. Koontz stimulated the awareness of organizational reality researchers with the concept of “the management theory jungle” [9], meaning a disordered way of practicing management sciences and the ontological and epistemological controversies growing in these sciences. Although several decades have passed since then, “we are still dealing with the jungle of management theory, and it’s more extensive and more vivid than ever” [10].

Some authors emphasize that the current situation of the management sciences shows a crisis of this science, especially their philosophical and methodological foundations. This crisis is considered to be systemic and permanent to the extent that it undermines even the *raison d'être* of management sciences as an independent scientific discipline [11]. It is even written that it does not have “theories or laws, or even a substitute for a scientific method” [12].

The problem of eclecticism in management sciences raised by the same author seems to strengthen the tendency to disperse the efforts to maintain the scientific nature

of research of the organizational reality. This is despite the fact that many researchers of organizational reality are united around the efforts to create a methodology of management sciences of a measurable and optimizing nature [13]. An example of an attempt to combine methodological approaches is combining the exchange paradigm and the paradigm of value [14] in economic sciences and by analogy combining the resource approach and the process approach in management sciences.

In management sciences, there are a few main problems, which are the obstacles in building true knowledge on managers' behavior and in the end replacing human managers with robots.

First of all, it is not entirely certain whether management sciences belong to idiographic or nomothetic sciences [15]. The solution to this dilemma affects many decisions during the scientific research, ranging from the choice of a qualitative or quantitative approach, through the selection of research objects, and finishing with the research tool and the way the results are interpreted.

Secondly, in the management sciences, the study of organizational reality dominates based on the situation at certain moments of time, which leads to a static and only temporary assessment of this reality [16]. This problem is quite complicated and has two aspects. The first aspect concerns the fact that longitudinal research is not usually carried out, but only a single registration of the studied phenomenon. One cannot infer how the researched phenomenon would change in the function of time. The second aspect is that, even if longitude studies are conducted, the time intervals between studies are too long, which in effect also amounts to a single registration of the studied phenomenon.

Thirdly, many theories in management sciences are created under the clear influence of the researcher's valuing the elements of these theories, which in the perspective of the development of science is an unfavorable phenomenon [17]. At present, it is easy to formulate "new management theories" that have not been properly verified, and the influence of subjectivity of researchers on the theories in management sciences is too large [18].

Fourthly, there are two rather critical approaches in management science to the realm of knowledge about organizational reality. On the one hand, one can find the view that the positivist striving to discover the truth about the organizational reality and to achieve the certainty of cognition is a utopia and consists of an attempt to create a better social order, because of implementing the idea of scientific management [19]. On the other hand, there phenomenon of an "overproduction of truth" about the organizational reality in the form of hundreds of millions of scientific publications exists, and the degree of its certainty cannot be accurately determined [20].

Fifth, in management science one can notice the phenomenon of the increasing diversification of the understanding of concepts and the introduction of new concepts, despite the fact that a common and coherent language is a basic element of the existence of a scientific discipline [21].

Sixth, despite the fact that in the literature of management sciences the view prevails that in these sciences there is a cumulative model of creating knowledge about organizational reality, which is related to the deepening of science as a whole [22], theories, concepts and management methods do not form one coherent research

perspective [23]. Such a large variety of approaches, paradigms and methodological concepts prevents the full use of the cumulative model of knowledge creation.

Seventh, the problem of incommensurability of the entire scientific discipline is clearly visible in management sciences, especially in the field of methods of research and interpretation of their results. This issue seems to be particularly visible, when it contrasts with the qualitative (interpretative) approach with the quantitative (neopositivist) approach. Due to the phenomenon of incommensurability of research results carried out even under the same approach, they cannot be compared with each other, which leads to the formation of “islands of knowledge” [24].

The methodological concept called the system of organizational terms is an attempt of solving these methodological problems and it consists of theoretical foundations and metrological solutions allowing to capture managerial actions in a data base.

2.2 Theoretical Foundations in Building Knowledge on Manager's Behavior

The theoretical foundation of solutions to the problems mentioned above is a view of manager's work which has been changed over one hundred years. At the beginning of scientific management, the picture of a manager in an organization was defined by his classical functions, such as a reflective planner, an organizer, a leader and a controller [25]. However, for more than 50 years a view of a nature of a manager has been dominated by two approaches.

Firstly, in 1964 Koontz and O'Donneil launched a discussion on the meaning of managerial skills [26]. In 1974 Katz proposed an approach in which managerial skills represented managerial work. The managerial skill was defined as an ability to work effectively as a team manager and to build cooperative effort within the team which the manager leads [27]. The dominating typology of managerial skills divides skills into 3 groups: technical, interpersonal and conceptual skills. Technical skills were regarded as most important for supervisors, interpersonal skills for middle managers, and conceptual skills for executives [28]. One of the latest typologies of managerial skills of managers contains such needed skills as critical thinking, problem solving, an ability to organize data, conceptual thinking, evaluating ideas, persuasive skills etc. [29].

Secondly, in 1980 Mintzberg concluded that the manager's work can be described in terms of 10 roles within interpersonal, informational and decisional areas which were common to the work of all types managers. Managerial roles are defined as areas of job activities which are undertaken by a manager [30]. Mintzberg introduced to management sciences a typology of managerial roles which contains such roles: a figurehead, a leader, a liaison, a monitor, a disseminator, a spokesman, an entrepreneur, a disturbance handler, a resource allocator, a negotiator [30]. Other researchers of team management proposed other divisions of roles, such as a leader, a peer, a conflict solver, an information sender, a decision maker, a resources allocator, an entrepreneur, a technician [31] or an explorer, an organizer, a controller, an adviser [32].

Managerial skills and managerial roles have influenced scientists and practitioners so much, that most of research on managerial work was designed as a research either on managerial skills or managerial roles. However, such approaches still do not recognize what really a team manager does [33] so that it is not possible to recognize team

managerial action patterns in (1) a time domain, (2) a content domain and (3) a human relations domain. Such patterns seem to be necessary when we think of replacing human managers with robots.

The answer to the question about what a team manager does seems to be hidden in the relation between managerial roles and managerial skills, because it is said, in order for a manager to play managerial roles, they should have some managerial skills [34]. It results in understanding playing managerial roles within their managerial skills by day-to-day activities of managers' effects in the managerial actions, which these managers make. Therefore, the managerial action can be defined as a real activity, which a manager does in order to play a managerial role when he has a certain managerial skill [35].

Then it is another question: how to describe managerial action in a universal and scalable way? The answer comes from the philosophical foundation of Wittgenstein's vision of the world which includes assumptions that the world consists of facts (the only beings in the world) and their "states of facts" [36]. In the system of organizational terms this concept was extended and it was proposed that managerial actions can be organized by events and things.

In the ontology of organizational reality, according to the system of organizational terms, it is assumed that every fact in the organizational reality can be represented by the organizational term [37]. The organizational term is a symbolic object which can be used as an element of the organizational reality model [38]. The organizational term is a close analogy to a physical quantity in the SI unit (length, mass, time etc.). It is assumed that the organizational terms are abstract objects which are used to represent the facts which appear in the organizational reality. The features of the organizational term, on the one hand, come from its definition and, on the other hand, it derives from causal relations or occurrence relations with other organizational terms [39]. When the organizational term appears, it can be changed quantitatively, qualitatively, mero logically, and substantially [40].

According to the logical division, organizational terms are divided into two classes: primal and derivative organizational terms. Facts, which are resources in the organizational reality [41], are represented by primal organizational terms. Facts, which are processes in the organizational reality, are represented by derivative organizational terms. By the same token the system of organizational terms combines the resource approach and the process approach in the management science. It combines processes which effect in resources. In pairs they create managerial actions [34]. In addition, the next logical division creates different types. The number of types is not defined.

Main principles of the ontology of the organizational reality in the system of organizational terms are as follows [42]:

- The ontology of the organizational reality consists of facts.
- The facts represent managers' work.
- There are two different classes of facts: things (factT, represented by a primal organizational term) and events (factE, represented by a derivative organizational term).
- In any moment of time only one relation between facts appears (there are two classes or relations: "creates" and "starts").

- Between facts there are such relations: factE creates fact, factT starts factE, factE starts factE.
- All relations appear in a considered period of time.
- Every relation between facts appears one by one within time (according to the assumption that “human activities are mostly serial”).

The pair of facts—factT and factE (represented by the primal organizational term and the derivative organizational term; in the common language of management by a process and a resource)—is just called a managerial action and it appears one after another, creating the managerial action. The examples of managerial actions are described in Sect. 3.

From a philosophical point of view, as shown in Fig. 2, each event (a process) and thing (a resource) have the label I.J, in which I and J represent a number and a version of a thing, respectively. Event 1.1 causes thing 1.1, which in turn releases event 2.1 that creates thing 2.1. Thing 2.1 starts event 3.1 which creates thing 3.1. Then, thing 3.1 generates a new version of the first event, i.e. event 1.2. In such a way, a new version of the first thing is created, which is called thing 1.2.

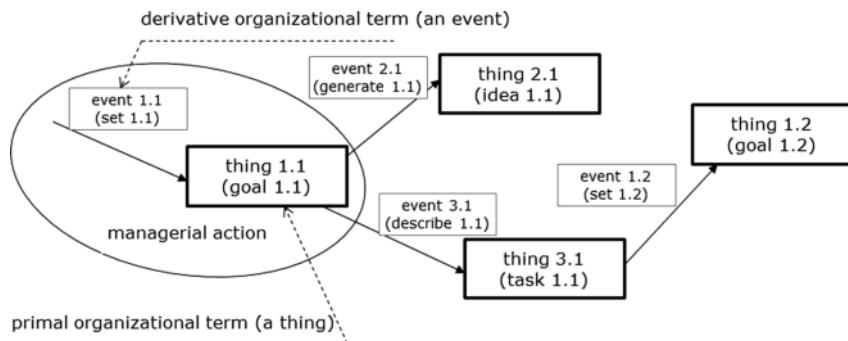


Fig. 2. Fundamental structure of managerial actions

From a practical point of view, as shown in Fig. 2, when a team manager sets a goal, in a certain moment of time a managerial action occurs represented by setting 1.1 (an event) and goal 1.1 (a thing). Specifically, as shown in Fig. 1, each event and thing have the label n.m, in which n and m represent a number and a version of a thing, respectively. What is important, goal 1.1 has features in time, content and human relations domains.

If later (e.g. after the next managerial action—describing 1.1 and task 1.1) this team manager does the next setting of the same goal, he launches the next managerial action. As the result of it the features of this goal are changed (goal 1.1 changed into goal 1.2) and represent the second version of this managerial action (described by the pair of the event and the thing: setting 1.2 and goal 1.2). The difference between managerial action features (goal 1.2 and goal 1.1.) let do reasoning on the events which happened in this period of time. Other words, what this team manager really did.

Besides two classes of organizational terms, there are two important terms in this concept. The first is a dimension of the organizational term and the second is a measured entity. The dimension of the organizational term is a general feature of a fact. A measured entity shows how much two facts differ from one another or one fact differs from itself in the function of time. In the language of management, it means how two resources or processes differ from one another or how they differ from themselves in the function of time. The dimensions of organizational terms consist of one or more measured entities. In other words, the measured entities are the measures of resources or processes in an organizational reality.

By the same token, the system of organizational terms combines the resource approach and the process approach in management sciences. It combines processes which effect in resources. In pairs they create managerial actions. As it was mentioned above, features of managerial actions are grouped in time, content and human relations domains. They show how much two managerial actions differ from one another or one managerial action differs from itself in the function of time. This enables to track a team manager by creating a map of detailed features vectors describing “who”, “what”, “when” and the “how” [43].

This approach lets solve the main methodological problems in management science generally and in building knowledge on manager’s behavior particularly. However, in order to get data on managerial actions there is a strong need of a unique data recording method on manager work. In Sect. 2.3 such a method enriched in research tools will be presented.

2.3 Metrological Solutions in Building Knowledge on Manager’s Behavior

The system of organizational terms is an original theoretical construct in which the organization performance is tracked and recorded. In order to do so, observation techniques are used along with the online management tools.

According to the theoretical background of the research tools described in Sect. 2.2, 10 online management tools have been created. They were implemented and available with the website browser. The platform with the tools was called TransistorsHead (transistorshead.com).

From the theoretical point of view online management tools have such features:

- according to the idea of an “unit of behavior” [44] every online management tool tracks and records one specific managerial action (as it was described in Sect. 2.2),
- when a manager uses any online management tool it is equal to an event occurring in organizational environment which effects in a thing, another words, equal to a process which results in a resource, respectively [45] (as it is shown in Fig. 1),
- every tool is useful for recording a certain managerial action [42].

There are also two more prepositions. Firstly, any management tool should cover all essential features which could describe the resource (represented by the primal organizational term). Secondly, any management tool should be as simple as it is possible. Users should want to use them during the research as research tools without any external motivation.

At present there are 10 different tools for different management techniques, such as: setting goals, describing tasks, generating ideas, specifying ideas, creating options, choosing options, checking motivation, solving conflicts, preparing meetings and explaining problems. The main scientific role of every management tool is recording a managerial action. The gathered data is divided into two parts: (1) a time domain and (2) a content domain. In the time domain (1) all button clicks are registered in the function of time. Therefore it is possible to conclude what a manager did.

Figures 3 and 4 show the dashboard of TransistorsHead with the example of the managerial action called SET GOALS (the name of the goal: FICC 2019). It is divided into several parts. At the top where managers can choose working with tools (TOOLS default), administer members of their teams (TEAM), hide some created items (derivative organizational terms) into archive (ARCHIVE) and read instructions how to use the tools (TELL ME ABOUT). There are also functions like login, logout and changing password, etc. The main menu consists of 10 different tools for team management, e.g. set goals, describe tasks, specify ideas, create options, etc.

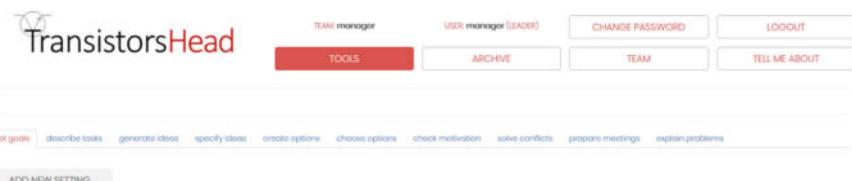


Fig. 3. Dashboard of the management tools platform

The screenshot shows the "set goals" tool interface. On the left, there is a sidebar with buttons: "VIEW" (grey), "EDIT" (orange, selected), "SHARE" (grey), "HIDE" (grey), "DELETE" (grey), "SAVE" (green), and "CLOSE" (red). The main area contains the following fields:

- Goal Name:** FICC 2019
- Period:** PERIOD (button) and DEADLINE (button) are shown, with a date field "2019-09-05 20:00" and a calendar icon.
- Measures:**
 - Measure 1: paper about the system of organizational terms
 - Measure 2: 20 pages
 - Measure 3: checked by a native
- Buttons:** ADD NEW MEASURER (grey) and a large green "SAVE" button.

Fig. 4. Goal named “FICC 2019” being edited in the “set goals” tool

With these management tools, it is possible to record each managerial action and describe it with a t-dimensional feature vector. This feature vector consists of two parts. The first one, which has a stable length, describes managerial in time domain (“when, “who” etc.). The second part of the vector describes the content of the managerial action (especially, a derivative organizational term) answering to the wide and complex question “what”.

In the left of the dashboard there is the ADD NEW function which means that in every tool a manager can create a new item, e.g. a new goal in SET GOALS. Below this button there is a list of items created in the chosen tool, e.g. lists of goals in SET TOOLS.

In the middle vertical part is the universal area containing the same buttons for every tool (VIEW, EDIT, SHARE, DELETE, HIDE). Below this area there are also universal buttons of action confirmation where a manager can save the item in the tool (the derivative organizational term) or close the tool without saving. Save confirmation uploads the data base with new data about the item, e.g. new goal parameters in the SET GOALS, and it creates the representation of a particular managerial action. In the right vertical part there is an area for forms, buttons, text areas or combo lists which a manager uses to establish the content of the tool item, e.g. a goal's name, deadline and measures. This vertical part contains different elements for every tool depending of the designed derivative organizational term parameters.

TransistorsHead platform, containing 10 different management tools, covers main managerial actions in team management. All items created in tools a team leader can share with team members and they can implement and save any changes in goals, tasks, ideas descriptions etc.

The architecture of t-dimensional feature vector which contains data on time domain and content domain of every managerial action depends on this managerial action. The length of the part which included time description of the managerial action (time domain) is stable and defined. The length of another part of this vector, containing features of the manager action (content domain), is different from every managerial action. Such construction of the feature vector is universal and it lets to plug in other measurements not only implemented as online management tools but also focused on recording some other areas of managerial actions such as tone of voice, location, face recognition etc.

In the last few years several research projects were done along with the system of organizational terms and TransistorsHead research tools. In the Sect. 3.1 there are examples of research results in different areas of manager's behavior.

3 Preliminary Research on Manager's Behavior Aimed at Application of Artificial Managers

3.1 Examples of Research Based on the System of Organizational Terms

The first experiment based on the system of organizational terms was done along with the non-participating observation technique from beginning of April to the end of June 2013.

The observation group consisted of students of Managing within the managerial specialization at the University of Economics in Katowice, who were tasked with preparing a complex project of management innovation consisting of a managerial tool in its IT version and a description of the techniques of using this tool. To determine the

objectives and tasks in this project, participants used managerial tools—Goaler (set goals) and Tasker (describe tasks), respectively—which are the measurement tools of primary organizational values: the goal and the task.

In the conducted observation, the activity of 8 managers was registered and data on two primary organizational quantities were obtained, which allowed to visualize the primal and derivative organizational terms in the form of an organizational reality model as a graph. Despite the fact that the participants of the experiment planned the same undertaking, the actions undertaken (the process of setting goals and the process of determining tasks) of each of them had a different course in time. This meant that individual organizational terms appeared at different times and in different order [46].

The collected data allowed not only to check in which order organizational terms appear in the function of time, but also contained information about the values of measured terms that were assigned to objectives and tasks as primal organizational terms. As a result of editing previously set goals or tasks, values of the measured terms also changed as a function of time. In this way it was possible to observe causal relations or relations between co-occurrences between the values of measured terms, although the collected data did not allow to clearly state which of these relations was present. It was also not possible to depict these relations as mathematical functions [47].

After the experiment, the participants were asked about how they perceived their work and what they remember about the course of their work. Differences, between the perception of one's work and the actual actions taken by the participants in the experiment, were quite significant. For example, one of the participants in the experiment believed that each of its team members received other tasks to accomplish, aimed at achieving the set goals. In reality, the manager assigned all tasks to the same person. Similar mistakes related to, for example, the number of objectives set, the frequency of their correction, the assessment of the degree of changes in the content of goals (i.e. the value of the measured organizational terms), etc. [48].

The comparison of managers' opinions with the results of observation of their work allowed to validate two hypotheses set in the study. The first was that managers are not fully aware of their activities during the management process they create. This means that in the organizational reality, primal and derivative organizational terms appear which are a result of a greater degree of unconscious, rather than the conscious actions of the manager. As a result, of the verification of the second hypothesis, it was assumed that the size of the managed enterprise, measured by the number of processes performed (in the system of organizational terms—the number of derivative organizational terms created in a given time interval) does not affect how much the manager is aware of the actions taken [48].

The results of the study confirmed the thesis that the memory about a human behavior is related to the objects (primal organizational terms) that man creates and this memory somehow fits in the relations between these objects and in the relationship between man and these objects [49]. A comparison of managers' opinions with the results of observations also revealed the weaknesses of traditional research techniques, such as questionnaire technique or intelligence technique, used in the field of manager's behavior testing.

The second research started in 2015. It was conducted among 64 students of Management at the University of Economics in Katowice working in groups of 4 persons (a team leader and 3 team members).

The main conclusions derived from such an analysis were as follows:

- the more thought processes the participants had to perform, the more precise their descriptions of the goal or task were. They were placed in managerial tools, which means that more measured terms of these primal organizational terms were registered by the managerial tool, and that the accuracy of the values of particular measured terms was bigger,
- in a situation where the project to be carried out was given in a very unstructured manner, the reconstruction of specific tasks and objectives to be performed was a fairly complex mental process and consumes a relatively long time of managers,
- the more details the task included, the less details the participants included in the final effect (the latest versions of the objectives and tasks), while creating themselves the content not appearing in the task [50].

Participants of the experiment received a comprehensive case study of a company intending to change the existing office to another, together with the characteristics given by the experimenter. This venture had to be planned by the participants by the means of the Goaler (set goals) and Tasker (describe tasks) management tools.

After completing the experiment, a linguistic analysis of the value of measured terms, aimed at assessing the integrity of the intertext of the task, which was given to the participants of the experiment with the effects they achieved at the end of the experiment. These effects were the latest versions of objectives and tasks to be implemented in the undertaking (primal organizational terms).

It can be added that the accuracy of the description of the objective or task in the project increased with the increase in the number of thought processes that participants had to perform.

The same experiment allowed answering the research question, if there are dominant linguistic models in planning of the same project by different managers. As a result of the collected data consisting of the value of the size of measured goals and tasks (primal organizational terms), it was possible to formulate a response that such data were not present in the collected data [51]. However, to conduct further research in this area, it was concluded that a CCC linguistic analysis model (correspondence, consistency, correctness) may be useful [52].

The third research based on the system of organizational terms was also conducted in 2015. The participants came from one of Silesian business schools and their goal was to plan an implementation of a new salary system in a company.

Students worked in groups of three or four, represented by a selected person who was using managerial tools. The aim of the experiment was to check whether the application of the goal management method, used together with the Goaler (set goals) and Tasker (describe tasks) management tools, would measure the effectiveness of the team.

Primal organizational terms—the goal and the task—have become indicators of the team's effectiveness in this project. To assess the effectiveness of the team's work, a group of parameters describing the appearance of individual primal organizational

terms being a result of the operation of these teams was selected. Table 1 presents these parameters and their values on the example of two teams participating in the experiment.

From the experimentally recorded data regarding the primal organizational terms presented in Table 1, it could be concluded that team 2 worked more effectively than team 1 because:

Table 1. Parameters of the team's effectiveness and results of the experiment on the example of two teams

Parameter	Meaning of parameters	Team 1	Team 2
A	Number of created objectives	3	4
B	Number of created tasks	6	7
C	Number of created operations	215	172
D	Worktime of the team from the first to the last login (in minutes)	60847	48987
E	Worktime of the team with the managerial tools (in minutes)	253	246
F	Number of the fractions of worktime of the team from the first to the last login	7	4
G	Number of the modifications of objectives	16	14
H	Number of the modifications of the tasks	33	26
I	Ratio of the modification of objectives (number of modifications for one objective)	2.67	1.75
J	Ratio of the modification of tasks (number of modifications for one task)	2.75	2.0

Source [54]

- worked shorter, both in terms of worktime from the first to the last login (parameter D), as well as the worktime with managerial tools (parameter E),
- in a shorter time set more goals and tasks (parameters A and B),
- had smaller values of ratio of the modification of objectives and tasks (parameters I and J), which indirectly indicated a smaller workload in determining them [53].

It should be added that the above parameters do not indicate which plan was more realistic or would have better results. Neither the effects of the plan's implementation nor the content of the plan were analyzed. Only the effectiveness of the project planning process was examined.

The fourth of experiments was focused also on planning projects. The participants of the experiment were, as before, students of Management in one of private business schools. The students of this university prepared their diploma theses in a rather unusual way, working on them in three-person teams. In the experiment, they were

given the task of planning, with the use of management tools, Goaler (set goals) and Tasker (describe tasks) for a real undertaking that stood in front of them—a team preparation of the diploma thesis.

The aim of the experiment was to assess the impact of management tools on the project planning process and attempt to find hidden rules in the way of planning the same project by 10 groups of students that could be identified with participants of 10 different organizations in the organizational reality. During the experiment, the participants first received a blank piece of paper and a pen as a “management tool”, and when they decided that the plan had already been drawn up, they were given access to the management tools: set goals and set tasks. The experiment time was limited for all groups and was 120 min.

The assessment of the impact of management tools on the work of teams has been achieved in two ways. First of all, the final versions of goals and tasks prepared in the set goals and set tasks tools by each group were compared with the record of goals and tasks in the first “management tool”—on a blank piece of paper. Secondly, the participants were asked by the means of a survey method for their opinions on the change in the way of planning due to the use of Goaler and Tasker management tools compared to the “blank piece of paper” tool. General conclusions summarizing the test results are provided in Table 2.

Table 2. Features of project planning with and without managerial tools

Assessment criterion	Planning without managerial tools	Planning with managerial tools
Planning time	Short	Long
Results (primal organizational terms—objectives and tasks)	Unclear and chaotic values of the measured terms	Precise and clearly described values of the measured terms
Flexibility of planning	Low	High
Availability of the created primal organizational terms (objectives and tasks)	Full	Full
Creativity of a group in a planning process	High	Low

Source [55]

An attempt to find the hidden rules in the way the same project was planned by 10 groups of students did not result in a dominant planning pattern. It was all the more puzzling that each of the 10 groups participating in the experiment planned the same real and feasible project. Yet the quantitative results regarding the number, nature and occurrence of primal organizational terms (objectives and tasks) in a function of time, despite some small similarities, did not allow to determine the dominant pattern of behavior of participants in the experiment.

In 2016 it was the fifth experiment based on the system of organizational terms. It was also focused on assessment how the management tools influence on the planning process and its content.

The selected parameters of group managers' work were compared in the situation of using online manager tools and without these tools, but only with a "blank piece of paper". These experiments confirmed that management tools strongly influenced the course of the planning process as well as the content of the objectives and tasks written in the project plan [55].

Based on data from this experiment in 2017, theoretical work on the search for a method of measuring the similarity of managers, whose activity can be registered, based on the designed features of the system of organizational terms through management tools, posted on TransistorsHead (transistorshead.com), was initiated. The aim of this work was to develop a method of automatic search for similarities between management methods undertaken by team managers. The assumption was made that the team manager could be described by his activity possessing a certain vector of traits.

Using the achievements of the field of pattern recognition in images [56], the method of calculating the degree of similarity of team managers has been developed, which has been positively verified by comparison to the classical statistical analysis of activities undertaken by managers. However, compared to the classic statistical analysis, this method, called "manager partial matching", allows to perform real-time calculations on any large or growing data set, and also indicates the similarity of managers on a scale from 0 (a completely different way of action) to 1 (identical way of operation) [35].

The sixth research was conducted in spring 2017. I was attended by 41 students of the Faculty of Management at the University of Economics in Katowice. They were divided into 5–6 people teams as a part of the subject Human resources management. Each of 7 teams identified a team manager who led the team during the observation. The teams started working on May 18th 2017 at 22:18:01 (the first time one of team managers logged in) and ended on May 30, 2017 at 20:19:12 (logged out by another team manager). The study was conducted by the means of the non-participant, long-term observation. The goal of the study was to recognize participative and authoritarian managers in the research group.

In this research it was distinguished 6 main managerial actions which described participative and authoritarian styles of managing. These managerial actions were recorded by 6 online management tools. It was analyzed these features by ratios between particular managerial actions of a team managers and all his team members. The results in percentage for one team are presented in Fig. 5 (team managers—blue color, team members—green color).

Team manager 1 were nearly fully authoritarian in setting goals (Goals 84.13%) and much participative in preparing meetings (Communication 23.08%). Comparing the results of each team manager to one another it comes the conclusion that the individual styles of management were completely different.

The seventh study was conducted between 26th of September and 20th of December 2017 among 50 BA business students at Haaga-Helia UAS, in Helsinki, Finland. The aim of the study was to answer the following research questions:

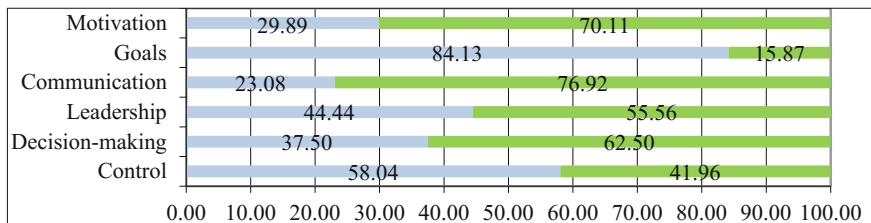


Fig. 5. Ratio between managerial actions of a team manager 1 and all his team members

- how does culture of the team members influence their communication?
- how does communication in multicultural teams influence generating solutions?
- do culturally homogenous teams communicate better than culturally heterogeneous teams?
- is there a correlation between the level of communication and the number of generated solutions?

Students were divided into eight purposeful teams of five. Three teams were homogenous culturally and linguistically. Five teams were heterogeneous both culturally and linguistically. The common working language was English. Their task was to generate ideas for a development project aimed at improving communication among their teachers, both during planning and implementing semester modules. The result of the teams' work was a written report containing two parts: training programme details (number of participants, venue, duration, goals of the project, benefits for the participants, training methods) and teamwork process (individual reflections on the work process, possible difficulties, benefits). To complete the assignment, the students were asked to use managerial tools, available from TransistorsHead (transistorshead.com).

This study expanded the above mentioned findings by investigating the impact of communication. It showed that communication in hetero- and homo-cultural teams influences generating solutions. Three homo-cultural teams, from low-context cultures, communicated more effectively and generated most solutions. However, four international, hetero-cultural groups showed frequent communication during the observation, but did not generate many solutions. The study revealed a weak correlation between the intensity of communication and the number of generated solutions. Although the intensity of communication process was high in some cases, the number of solutions was low. Additionally, the culture of the team members played a significant role in their communication. Mono-cultural teams, sharing native language, did not use management tools too often but achieved high output results. Conversely, in multi-cultural teams the intensity of communication was high, but the output of results was low.

This series of research based on experiment and long-term observations let check the usability of the system of organizational terms as a theoretical concept in managerial actions research. The research tools implemented as online management tools in TransistorsHead recorded data, which let to conclude about some patterns of managers' behavior. However, even having such data about real managerial actions there are still several problems in implementing artificial managers in practice. These challenges were described in Sect. 3.2.

3.2 Challenges in Application of Artificial Managers

There are 5 main research challenges in the implementation of an artificial manager.

First of all, one should answer the question how to build measuring tools useful in the non-participant observation method [57], which was used as the main research method in the system of organizational terms.

Secondly, there is the problem of how to avoid the effect that during the scientific research the researcher constructs a measuring tool expecting certain features of the phenomenon (in the system of organizational terms—values of measured terms), which narrows the scope of information collected about the examined reality [58]. In other words, due to this effect, some of the phenomena in the organizational reality, potentially described by organizational terms, may never be investigated, despite the fact that they will occur.

Thirdly, one can ask who should be the initiator of creating management tools, which are measuring tools [59]. It should be added that in the system of organizational terms, the choice applies to the manager, participants of the organization, users of the study or the researcher.

The fourth research problem that arises from the literature on the subject of automation of human activities is whether in the case of solving organizational problems in the organization it will work well—as in other areas of life—the method of imitation by the manager-robot of the current behavior of the manager-man [60].

The last, fifth research problem arises from the fact that the manager-robot would face a solution to the organizational problem, some of which are a significant obstacle to the manager-robot learning behavior of the human. Namely, organizational problems are characterized by:

- unstable parameters describing a given problem and undefined requirements as to its solution,
- high complexity of relations between elements of the problem,
- particularly large dependence on human activity [61].

The ability of an imitation depends on the ability to repeat the behavior of an original object [62]. In the literature on the subject, two types of imitation of the original object can be distinguished. The first type means accurately reproducing the original behavior in a given situation, without analyzing the context of this situation [63]. The second type requires analysis of both the behavior of the original and the object to which this behavior is manifested [64]. Because it seems that the imitation skill is crucial for the manager-robot to be able to interact with the participants of the organization [65], a research question arises, which kind of imitation of the manager's activity is possible to be implemented in the IT system, which is also the organization's management system.

Thus, a research question arises, how should an IT system, being a manager-robot, learn from a manager-man? The question, though linguistically simple, carries many detailed questions in areas of knowledge such as pattern recognition [66] and machine learning [67].

It seems that solving these research problems determines further work on replacing the manager's work with a robot.

4 Conclusions

The summing up the above outline of further directions of research using the system of organizational terms aiming at the manager's work automation, it should be emphasized that these studies are not about creating a human-like manager and at the same time being its higher evolutionary copy (in original "sophisticated superhuman machine"—author's note) [62].

The aim of the above-mentioned research problems is to improve already existing IT systems used in management, which, however, currently seem to occupy only lower levels in the hierarchy of management systems (in original "system of control"—according to S. Beer) [68].

There are also many other possibilities of using this methodology in business, science and arts. Firstly, in business in Human Resources Management there is a problem of people performance measurement. It is claimed that a subject of human work has always come to a simple question: what makes a man at work? [33]. Using the presented method of manager representation in area of representing work of any employee gives great opportunities in comparison of employee performance in order to us the results:

- systems of employee measurement,
- motivating systems,
- recruitment and development systems,
- time and production management systems.

If the similarities between people working on the similar positions are able to describe, it would be possible to design more efficient methods of paying for their job, motivate employees with similar needs characteristic by non-financial factors, conduct more adequate recruitment for similar positions. The advantage of this methodology would also increase time accuracy in production lines and let compare physical workers to one another.

Secondly, in science there is a strong need of conducting more efficient scientific research with less funds, achieving better and more accurate results. The question is what is the most efficient method of scientific research in a certain discipline which results in most cited and relevant publications?

If two or more scientists in the same discipline could be tracked what they really do, there could be some similarities in their work discovered by this methodology. Of course, another question is what kind of measure tools could be used for tracking scientists work; especially abstract work of designing, reasoning or drawing conclusions. The tools implemented in the TransistorsHead are not enough, because they use only the web page technology. There is a need of designing other tools which could measure changing location, an acceleration of a body, barometric pressure for a blood pressure, an ambient body temperature, a heart rate or a skin conductance. However, the methodology of representing work or other issues of activities is the same as it was describe in the paper.

Thirdly, in arts, especially in movie and television production there is a lack of concrete and stable knowledge how to design and produce movies and tv programs.

Tracking directors or producers work could be useful to formulate the best practices in fields, which are still covered under the misty art performance.

It is said that in approximately 125 years all human jobs will be automated in the term of “High-level machine intelligence” (HLMI) which means the state when unaided machines can accomplish every task better and more cheaply than human workers [69]. Despite the fact that the managerial profession is not listed in these jobs, the preliminary research on the use of the system of organizational terms as a concept of management sciences, however, provides the basis for formulating a vision of the development of management sciences, in which the gradual automation of the manager’s work may be an important element, and probably in the future, now undetermined, form of competition between the manager-man and the manager-robot [67].

Acknowledgements. Research activities leading to this work have been supported by the Faculty of Radio and Television at the University of Silesia in Katowice (Poland) and FoKoS at the University of Siegen (Germany). Olaf Flak greatly thanks to Prof. Marcin Grzegorzek from University of Siegen for his significant help in the experiments.

References

1. Ewenstein, B., Hancock, B., Komm, A.: Ahead of the curve: the future of performance management. McKinsey Q. (May 2016). <http://www.mckinsey.com/business-functions/organization/our-insights/ahead-of-the-curve-the-future-of-performance-management>
2. Fidler, D.: Here’s how managers can be replaced by software. Harv. Bus. Rev. (April 2015). <https://hbr.org/2015/04/heres-how-managers-can-be-replaced-by-software>
3. Drucker, P.F.: the manager and the moron. McKinsey Q. (Dec 1967). <http://www.mckinsey.com/business-functions/organization/our-insights/the-manager-and-the-moron>
4. Klein, G., Feltovich, P.J., Bradshaw, J.M., Woods, D.D.: Common ground and coordination in joint activity. In: Rouse, W.R., Boff, K.B. (eds.) *Organizational Simulation*. Wiley, New York (2005)
5. Christoffersen, K., Woods, D.D.: How to make automated systems team players. *Adv. Hum. Perform. Cogn. Eng. Res.* **2**, 1–12 (2002)
6. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (eds.) *Perspectives on Socially Shared Cognition*. American Psychological Association, Washington (1991)
7. Scherbaum, C.A., Meade, A.W.: Measurement in the organizational sciences. In: Buchanan, D., Bryman, A. (eds.) *Handbook of Organizational Research Methods*, pp. 636–653. Sage, London (2009)
8. Goebel, E., Mills, I.M., Wallard, A.J.: The International System of Units (SI), 8th edn. Organisation Intergouvernementale de la Convention du Mètre, Paris (2006)
9. Koontz, H.: The Management Theory Jungle. *The Journal of the Academy of Management* **4** (3), 174–188 (1961)
10. Witczak, H.: Wstęp do systemu nauk o zarządzaniu. *Contemp. Manag. Q.* **2**, 27–40 (2013)
11. Sobczyk, J.R.: Kryzys podstaw metodologicznych nauk o zarządzaniu—kryzysem powinowactwa z naukami społecznymi. *Acta Univ. Lodz. Folia Oecon.* **234**, 335–345 (2010)
12. Sulkowski, Ł.: Cognitive challenges of epistemology in management sciences. *Entrep. Manag. t. 14, z. 5, cz. 1, pp. 271–282* (2013)

13. Niemczyk, J.: The development of management science from perspective of paradigms of economic sciences. *Organiz. Manag.* Nr **1A**, 167–175 (2014)
14. Douglass, B.: Economic methodology and nobel laureates: confirmation of a methodological paradigm shift. *Am. J. Econ. Sociol.* **71**(5), 1205–1218 (2012)
15. Callaghan, C.: Contemporary insights from social sciences theory: implications for management. *S. Afr. J. Bus. Manag.* **48**(4), 35–45 (Dec 2017)
16. Guercini, S.: New qualitative research methodologies in management. *Manag. Decis.* **52**(5), 662–674 (2014)
17. Hicks, H.G., Goronzy, F.: On methodology in the study of management and organization. *Acad. Manag. J.* **10**(4), 371–384 (Dec 1967)
18. Sudoł, S.: Main dilemmas of management science. *Organiz. Manag.* **1**(139), 7–22 (2010)
19. Sułkowski, Ł.: About utopias in Management. *Contemp. Manag. Q.* **3**, 23–33 (2011)
20. Darmer, P.: The subject(ivity) of management. *J. Organiz. Change Manag.* **13**(4), 334–351 (2000)
21. Hodge, R.: Towards a postmodern science of language. *Soc. Semiot.* **13**(3), 241–262 (2010)
22. Edmondson, A.C., Mcmanus, S.E.: Methodological fit in management field research. *Acad. Manag. Rev.* **32**(4), 1158–1159 (2007)
23. Sułkowski, Ł.: Problem niewspółmierności koncepcji w zarządzaniu. *Prz. Organiz.*, Nr 4, s. 6–8 (2004)
24. Gleiser, M.: *The Island of Knowledge: The Limits of Science and the Search for Meaning*. Basic Books, New York (2014)
25. Fayol, H.: *Administration industrielle et générale*. Dunod, Paris (1916)
26. Koontz, H., O'Donnell, C.: *Principles of Management*. McGraw-Hill, New York (1964)
27. Katz, R.L.: Skills of an effective administrator. *Harv. Bus. Rev.* **52**(5), 90–102 (1974)
28. Kaiser, R.B., Craig, S.B., Overfield, D.V., Yarborough, P.: Differences in managerial jobs at the bottom, middle, and top: a review of empirical research. *Psychol. Manag. J.* **14**(2), 76–91 (2011)
29. Ullah, F., Burhan, M., Shabbir, N.: Role of case studies in development of managerial skills: evidence from Khyber Pakhtunkhwa Business Schools. *J. Manag. Sci.* **8**(2), 192–207 (2011)
30. Mintzberg, H.: *The Nature of Managerial Work*. Prentice-Hall, New York (1980)
31. Pavett, C.M., Lau, A.W.: Managerial roles, skills, and effective performance. In: *Academy of Management Proceedings*, pp. 95–99 (1982)
32. McCan, D., Margerison, Ch.: Managing high-performance teams. *Train. Dev. J.* **10**, 53–60 (1989)
33. Sinar, E., Paese, M.: The new leader profile. *Train. Mag.* **46**, 46–50 (2016)
34. Pavett, C.M., Lau, A.W.: Managerial work: the influence of hierarchical level and functional specialty. *Acad. Manag. J.* **26**(1), 170–177 (1983)
35. Flak, O., Yang, C., Grzegorzek, M.: Action sequence matching of team managers. In: *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods ICPRAM*, pp. 24–26 (2017)
36. Cheung, L.K.C.: The unity of language and logic in Wittgenstein's Tractatus. *Philos. Invest.* **29**(1), 22–50 (2006)
37. Zalabardo, J.: Representation and reality in Wittgenstein's Tractatus. Oxford University Press, Oxford (2015)
38. Rios, D.: Models and modeling in the social sciences. *Perspect. Sci.* **21**(2), 221–225 (2013)
39. Backlund, A.: The definition of system. *Kybernetes* **29**(4), 444–451 (2000)
40. Kulieshev, A.: Simplicity and complexity in metaphysics. *Path Sci.* **4**(6), 3001–3006 (2018)
41. Barney, J.B.: Firm resources and sustained competitive advantage. *J. Manag.* **17**(1), 99–120 (1991)

42. Flak, O.: Theoretical foundation for managers' behavior analysis by graph-based pattern matching. *Int. J. Contemp. Manag.* **12**(4), 110–123 (2013)
43. Beshears, J., Gino, F.: Leader as decision architects. *Harv. Bus. Rev.* **5**, 51–62 (2015)
44. Curtis, B., Kellner, M., Over, J.: Process modelling. *Commun. ACM* **35**(9), 75–90 (1992)
45. Flak, O.: Concept of managerial tools based on the system of organizational terms. In: Knosala, R. (ed.) *Innovation in Management and Production Engineering*, pp. 187–197. Oficyna Wydawnicza Polskiego Towarzystwa Zarządzania Produkcją, Opole (2013)
46. Flak, O.: Results of observations of managers based on the system of organizational terms. In: Nalepka, A., Ujwary, A. (eds.) *Business and Non-profit Organizations Fading Increased Competition and Growing Customers' Demands*, pp. 89–102. Wyższa Szkoła Biznesu, Nowy Sącz (2013)
47. Flak, O.: Ustalanie celów i zadań za pomocą narzędzi menedżerskich online - wyniki obserwacji," *Studia Ekonomiczne*, Nr 199. Uniwersytet Ekonomiczny w Katowicach, Katowice, pp. 46–57 (2013)
48. Flak, O., Pyszka, A.: Differences in perception of the participants in the management process and its real trajectory. *J. Entrep. Manag. Innov.* **9**(4), 65–66 (2013)
49. Shepherd, A.M., Gibbs, MCh.: Remembering things. *Inf. Soc.* **24**, 47–53 (2008)
50. Flak, O., Alnajjar, J.: Ocena spójności intertekstowej w planowaniu projektu. Wyniki badania z wykorzystaniem układu wielkości organizacyjnych. *Lingwist. Stosow.* Nr 13, 7–13 (2015)
51. Flak, O., Alnajjar, J.: Linguistic analysis of managers' behaviour aimed at replacing human managers with robots. In: *Proceedings of the 15th European Conference on Research Methodology for Business and Management Studies* Kingston Business School. Kingston University London, London, pp. 1–8 (2016)
52. Renkema, J.: Improving the quality of governmental documents: a combined academic and professional approach. In: Cheng, W., Kong, K.C.C. (eds.) *Professional Communication: Collaboration Between Academics and Practitioners*, pp. 173–190. Hong Kong University Press, Hong Kong (2009)
53. Hoffmann-Burdzińska, K., Flak, O.: management by objectives as a method of measuring HR teams effectiveness. *J. Posit. Manag.* **6**(3), 67–82 (2015)
54. Flak, O., Hoffmann-Burdzińska, K.: Management techniques and tools in project planning—Part 2. Qualitative results of research. In: Knosala, R. (ed.) *Innowacje w zarządzaniu i inżynierii produkcji*, vol. 1, pp. 288–298. Oficyna Wydawnicza Polskiego Towarzystwa Zarządzania Produkcją, Opole (2016)
55. Flak, O., Hoffmann-Burdzińska, K.: "Managerial Tools' Influence on a planning process. results of the experiment. In: Nalepka, A., Ujwary, A. (eds.) *Business and Non-profit Organizations Fading Increased Competition and Growing Customers' Demands*, pp. 119–139. Wyższa Szkoła Biznesu, Nowy Sącz (2016)
56. Bronstein, A.M., Bronstein, M.M., Bruckstein, A.M., Kimmel, R.: Partial similarity of objects, or how to compare a centaur to a horse. *Int. J. Comput. Vis.* **84**(2), 163–183 (2009)
57. Russell, B.: *An Inquiry into Meaning and Truth*. Routledge, London (1996)
58. Storozhuk, A.: Perception: mirror-image or action? *J. Gen. Philos. Sci.* **38**(2), 369–382 (2007)
59. Bourne, M., Neely, A.: Implementing performance measurement systems: a literature review. *Int. J. Bus. Perform. Manag.* **5**(1), 1–24 (2003)
60. Dayan, P., Hinton, G.E., Neal, R.M.: The Helmholtz machine. *Neural Comput.* **7**, 889–904 (1995)
61. H. Rittel and M. Webber, Planning problems are wicked problems. In: Cross, N. (ed.) *Developments in Design Methodology*, pp. 135–144. Wiley, New York (1984)
62. Schaal, S.: Is imitation learning the route to humanoid robots? *Trends Cognit. Sci.* **3**, 233–242 (1999)

63. Dautenhahn, K.: Getting to know each other: artificial social intelligence for autonomous robots. *Robot. Auton. Syst.* **16**, 333–356 (1995)
64. Atkeson, C.G., Schaal, S.: Learning tasks from single demonstration. In: IEEE International Conference on Robotics and Automation (ICRA 97), pp. 1706–1712. IEEE Press (1997)
65. Breazeal, C., Scassellati, B.: Robots that imitate humans. *Trends Cognit. Sci.* **6**(11), 481–487 (2002)
66. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Elsevier, Burlington (2009)
67. Kelleher, J.D., Namee, B.M., D'Arcy, A.: Fundamentals of Machine Learning for Predictive Data Analytics. MIT Press, Massachusetts (2015)
68. Leonard, A.: The viable system model and its application to complex organizations. *Syst. Pract. Action Res.* **22**(4), 223–233 (2009)
69. Katja, G., John, S., Allan, D., Baobao, Z., Owain, E.: When will AI exceed human performance? Evidence from AI experts (3 May 2018). [arXiv:1705.08807v3](https://arxiv.org/abs/1705.08807v3)



Interruption Timing Prediction via Prosodic Task Boundary Model for Human-Machine Teaming

Nia Peters^(✉)

Air Force Research Laboratory, Wright Patterson Air Force Base,
Dayton, OH 45433, USA
nia.peters.1@us.af.mil
<https://petersnias.com>

Abstract. Human-machine teaming aims to meld the human cognitive strengths with the unique capabilities of smart machines to create intelligent teams adaptive to rapidly changing circumstances. One major problem within human-machine teaming is a lack of communication skills on the part of the machine such as the inability to know when to communicate information to or interrupt human teammates. To address this issue, an intelligent interruption system that monitors the speech within human-machine teaming interactions and predicts when to interrupt based on where human teammates are within the primary task is proposed. The intelligent interruption system leverages the raw audio within a simulated human-machine teaming interaction, extracts prosodic information, and predicts task boundaries as candidate interruption timings. Various machine learning techniques are evaluated as a prosody-only task boundary model and their task boundary detection performance is compared. The prosody-only task boundary model is implemented in real-time and the system latency and task boundary detection performance is evaluated. The final results indicate that although prosodic information processes with a low latency making it tractable in real-time, the prosody-only task boundary model performance is degraded in robust dialogues of human-machine teaming interactions.

Keywords: Human machine teaming · Intelligent interruption · Speech prosody · Collaborative communication

1 Introduction

Human-machine teaming research focuses on the efficient and effective integration of humans with complex machines where human factors and software engineering are combined to incorporate artificial intelligence techniques into human teams to provide intelligent information and interaction capabilities [1]. This research area has attracted the attention of the U.S. military from the implementation of human-robot teams to unmanned aerial vehicle (UAV) operations.

In addition to military endeavors, commercial development includes supporting overworked and multi-tasking nurses in hospitals, systems for spotting anomalies and preventing cyber-attacks, and systems to help design components in manufacturing systems [2]. In transportation, human-machine teaming systems are being developed for systems that team up with traffic enforcement officers helping them to balance their workloads [2]. Other domains that benefit from human-machine teaming are operator control stations for power and chemical processing plants, air traffic control stations, commercial and military pilots in cockpits, power grid management, locomotion transportation monitoring, and human-computer technical support teams.

Human-machine teaming will enable human and machines to communicate and share information. If the objective is for machines to initiate actions on their own, without explicit human consent but they do not possess the communication skills that are required to know when and how to share information with human teammates concerning their intentions, actions, and limitations, then one major contributor to the problem of human-machine teaming interaction is a lack of communication skills on the part of the machine. In an effort to overcome this communication deficiency, a primary communication strategy that has been extensively explored within human factors and engineering literature for autonomous collaborative communications are *interruptions*. In this research an *interruption* is defined as an unanticipated request for task switching from a person, an object, or an event while multitasking [3].

Interruption science is a research area focused on how interruptions affect human performance and the development interventions to ameliorate the disruption caused by these interruptions. Within the current scope of the interruption science literature, a primary intrusion technique is the development of *intelligent interruption systems*. These systems leverage information from single and multitasking interactions and apply hand-crafted or machine learning techniques to disseminate information at appropriate times. The purpose of these systems is to interrupt at points that have minimal cost of disrupting the overall interaction. It is imperative that machines adhere to appropriate communication and *interruption* strategies that do not hinder the overall task goals.

The work in *intelligent interruption system* development primarily explores interruptions in human-machine tasks from the perspective of *single-human, multitasking* and *multi-human, single task* environments. A *single-human, multitasking* interaction is one in which one human is engaged in a primary task and is interrupted with information by the machine relevant to an orthogonal secondary task. A *multi-human, single task* environment is one in which multiple humans are engaged in a primary task and interrupted by the machine with information related to that primary task.

One limiting factor within the current literature is with respect to the exploration of the development of *intelligent interruption systems* within *multi-human, multitasking interactions*. A *multi-human multitasking communication interaction* is one in which multiple humans are engaged in a primary task and interrupted with information from the machine related to an orthogonal secondary task where the primary modality is natural human speech. Within these

exchanges, humans are not only multitasking but also collaborating. In multi-tasking environments, humans are simultaneously working on one or more unrelated tasks. While collaborating, switching tasks could affect interdependencies with other teammates (human or machine). Providing awareness information to machine collaborators could be beneficial in helping align their tasks and interactions.

Although there are numerous methods of communication such as text and visual, the focus of this work is with respect to natural human speech, human's faster and most natural form of communication. Within this interaction, the proposed *intelligent interruption system* is tasked with predicting appropriate interruption timings and disseminating user-interruptions which are the intention of the system to convey new information to a member participating in a collaborative communication task, which aids in one task but may disrupt another.

The proposed solution is the development of an *intelligent interruption system* for multi-human multitasking human-machine teams. Since this intelligent system will be integrated into a collaborative interaction and leverages information from the dialogue of human teammates', the overall system goal is to leverage low-level communication information and map it to high-level constructs indicative of appropriate task interruption timings. The proposed *intelligent interruption system* maps features of the raw audio within human-machine interactions onto constructs indicative of appropriate interruption moments such as the end of a task or *task boundaries*. The overall contribution of this work is a proposed approach to the exploration of an *intelligent interruption system* that can infer when to communicate with human teammates. The system is evaluated on how accurately it can map low-level information streams such as prosody (how a person says something) into higher constructions indicative of interruptability such as humans completion of a task or task boundaries and its tractability in real-time. The use of only prosodic information to infer task boundaries is an exploratory measure of how well one can do in predicting task boundaries using only features derived from the raw audio. Making predictions from the raw information within a communication channel has potential for quick data processing and modeling, but may hinder detection accuracies.

2 Background

Why is it necessary to deploy an intelligent mechanism for disseminating interruptions within multi-human, multitasking human-machine teams? Studies show that interrupting primary tasks can negatively impact productivity [4–7] and affective state [8,9]. Within these contexts there have been proposed methods of intelligent system-mediated interruptions. There is empirical research dedicated to manipulating time on the delivery [4–6] of system-mediated interruptions [10] in multi-task environments [11]. There is also literature that explores immediate interruption dissemination [5,12,13] within dual-task scenarios. Studies have shown that delivering interruptions at random times can result in a decline in

performance on primary tasks [4,5,13–15]. Other studies show similar results [5,9,16,17] and the differences in cost of interruptions are typically attributed to differences in workload at the point of interruption [4]. Additionally, studies have illustrated that interrupting users engaged in tasks has a considerable negative impact on task completion time [5,7,14,18–20]. Interrupting tasks at random moments can cause users to take up to 30% longer to resume the tasks, commit up to twice the errors, and experience up to twice the negative affect than when interrupted at boundaries [4,8,21]. Studies in [22] explores the benefits of providing a form of intelligent interruption dissemination within multi-human, multitasking interactions in terms of task performance across multiple tasks.

There has been considerable research in developing systems that use intelligent methods to disseminate interruptions in multitasking environments. Machine learning techniques have continually been explored in the development of intelligent interruption systems. The Lookout system predicts a user's dwell time on a communication message based on an analysis of its content and predicts a scheduled time to automate assistance [23]. The Notification Platform leverages messages from multiple device sources and performs a decisions analysis to deliver notifications that are most beneficial to the user [24]. Similarly, the BESTCOM system uses social and task context, communication preferences, and available channels to predict the best timing for interpersonal communication [25]. Other systems and models include Lilsys [26], MyVine [27], and a Cost of Interruption (COI) model [28]. The literature continues to emphasize intelligent system-mediated interruptions and its effect on user performance and affect. Previous and current work has made great strides in alleviating the cost of interruptions, but are primarily applied to *single-human multitasks* and *multi-human single task* interactions which are only a subset of the possible interactions in which interruption techniques could be useful. Additionally one could imagine human-machine teams would be comprised of multiple teammates performing multiple tasks.

There is an immense amount of literature that recommends appropriate points of interruptibility as *boundaries within task execution*. A *task boundary* can be defined as a time instance between two moments of task execution. Task boundary modeling has been used in *single-human, multitasking intelligent interruption systems* to indicate appropriate points of interruptibility [4,5,8,28] and shown that deferring delivery of notifications until a boundary is reached can meaningfully reduce costs of interruptions. Conversely, interrupting tasks at random moments can cause users to take up to 30% longer to resume the tasks, commit up to twice the errors, and experience up to twice the negative affect than when interrupted at boundaries [4,8,21]. There is also work that states defining points of appropriate interruption at course task boundaries [8] which can be thought of as boundaries that exist at a higher level of task execution compared to fine breakpoints which exist at a lower level. Some studies [5,18] place moments for interruption towards the beginning, middle, or end of a task. This kind of strategy relates most to Miyata and Norman [29] who explain that task execution occurs in three phases: planning, execution, and evaluation which could be logically extended to each of the subtasks of a task. As tasks in them-

selves, every subtask would then contain moments of planning, execution, and evaluation, making task execution a repeated loop of these phases. Other studies place interruptions between instances of repetitive sequences or, more generally, at breakpoints in a task sequence [6,30,31].

From preliminary work [32] done on multi-human, multitasking collaborative interactions, results indicate humans interrupt closer to the end of a task. From this, task boundaries as predictors of interruptibility is explored. Since the system only has access to low-level speech cues, it is necessary to map low-level speech information to higher constructs of interruptibility which in this case are task boundaries. Within the speech community literature, speech prosody has been used to detect boundaries in sentences, discourse structure, and grounding [33–36]. The aim of this research is to use similar techniques to use speech prosody to infer task boundaries as moment of appropriate interruption within multiple-human, multitasking interactions.

3 Human-Machine Teaming Corpus

Prior to developing an *intelligent interruption system*, it is necessary to simulate a multi-human, multitasking interaction. There is a lot of variability in multi-human, multitasking human-machine teams in terms of the number of human and machine teammates and tasks. Within this research the *intelligent interruption system* is developed for the a dual-human, dual-task team. In the simulated dual-human, dual-task interaction, two humans communicate using a push-to-talk communication interface regarding information related to a human-human task. The intelligent interruption system “listens” to this speech communication stream, processing information at the push-to-talk level, and makes inferences on when to send information related to an orthogonal human-machine task or the *interruption task* which is simultaneously being performed by both human teammates as illustrated in Fig. 1.

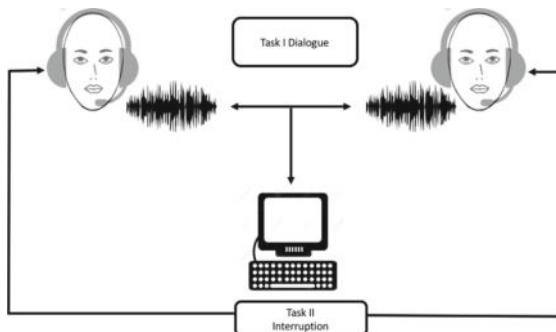


Fig. 1. Human machine teaming interaction

The aim of the data collection is to best simulate an information alignment task where users may have a differing vocabulary for the same object and must

take turns to align their knowledge while simultaneously simulating a human-machine task that monitors the interaction of the primary task to make interruption timing decisions and send information related to a human-machine task. The two multi-human, multitasking tasks include the *Tangram Task* and the *Uncertainty Map Task* (UMT). System integration into the Tangram task is built using data from the *Tangram task* interaction and the same is true for the Uncertainty Map Task (UMT).

For both tasks it is necessary to establish ground truth task boundaries in an effort to evaluate the *Prosody-Only Task Boundary Model*. In order to get task boundary ground truth within the experimental design of both tasks, several tasks were run successively in order to establish task boundaries. Figure 2 is an example illustrating the flow of information in both the Tangram and UMT tasks.

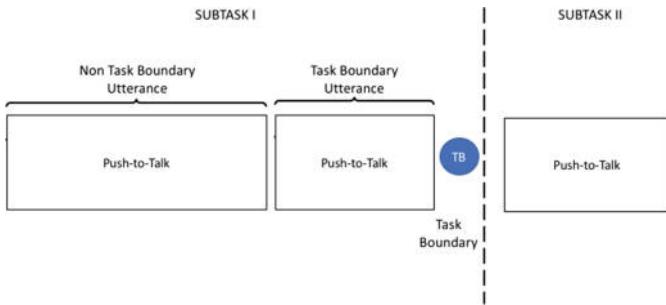


Fig. 2. Information flow for all data collection interactions

The dotted line illustrates the transition from one task to another. This succession of tasks is continually presented to both teammates until either a predetermined number of interruptions has been delivered or until a predetermined amount of time has elapsed. Within a task, users are engaged in several exchanges using a push-to-talk communication interface. Within both experimental datasets, the definition of a task boundary is when both users press a DONE button indicating their completion of the current ask and the intent to progress to the next task until the entire experiment is completed. The reason for waiting for both users to press a DONE button is to get a ground truth indicator that both users agree they are done with the current task and ready to proceed to another. With this being a team task, it is necessary to get a confirmation from both teammates that they are indeed done. The method for extracting task boundary ground truth is consistent in both the Tangram and UMT tasks.

In subsequent sections the human-human tasks (*Tangram Task* and *Uncertainty Map Task*) and simultaneous human-machine interruption task (*Keeping Track Task*) are detailed. Both teammates are performing the human-human and human-machine tasks simultaneously and a correct response on a human-human

task results in a point for that task and the correct response on the human-machine task results in a point for that task. Both task scores are combined for a total team score. The two teammates engaged in the experimental tasks were informed that performance in both the human-human and human-machine tasks are equally important and asked to avoid emphasizing good performance on one task over the other. From the perspective of the experimenter, most participants and teams were equally motivated by both tasks and did their best on both.

3.1 Tangram Task - Human-Human Interaction

The first human-human task or *Tangram task* involves two users corresponding over a push-to-talk communication network to arrange abstract shapes (Tangrams) within a column into corresponding order to simulate aligning knowledge from two perspectives. Tangrams are objects composed of geometric shapes so some objects can be described based on their shape composition or what the object actually resembles. From an experimental design standpoint, Tangrams were used in order to display abstract objects to users and establish a back and forth dialogue of uncertainty that would allow users to align their vocabulary and knowledge in order to successfully arrange their Tangrams in corresponding order. A succession of columns with random Tangram shapes was presented to the team until a predefined number of interruptions was sent from the interruption task. Figure 3 is an example of the graphical user interface (GUI) of the task for both participants. Teammates use the “Push-to-Talk” button to speak to their partner, the “Answer Query” button to respond to UAV update request from the interruption task, and the “Done” button (pressed by both participants) to indicate the completion of knowledge alignment on the current Tangram. From there the system will proceed to a new set of Tangrams.

A sample dialogue exchange within this task is illustrated in Fig. 4.



Fig. 3. Tangram task interface

In Fig. 4 in (A1) Teammate 1 begins by rehearsing an interruption (UAV state update) from the human-machine interruption task then proceeds to

Teammate 1: Seventy-percent. Ok I gotta upside down flamingo, a turtle, a duck facing to the left, and a goose facing upwards (**A1**)

Teammate 2: Got it (**A2**)

***Both users press the DONE button to move on to the next subtask or the set of shapes (**A3**)*

Teammate 1: Aw man, I thought we had it. Ok I have a uh a fox. I have like a dog running to the left then I have a um ram I think it's been called and then I have this weird animal that's on one leg (**A4**)

Fig. 4. Example dialogue from the tangram task

describing the current column (left pane of Fig. 3). In (**A2**), Teammate 2 confirms that he/she understands the shapes Teammate 1 described. From (**A3**) both users press DONE which is considered a *task boundary* in which another set of abstract shapes is generated for the players to describe. In (**A4**) Teammate 1 makes a reference to the feedback that potentially illustrates the teammate's inability to correctly arrange the Tangram shapes in the corresponding order. Teammate 1 then proceeds to describe the next column of shapes.

The dialogue within the primary task seems to follow the theory of Conversational Games where conversations consist of a series of GAMES, each of which involves an INITIATING MOVE (such as an instruction or query), followed by either a RESPONSE MOVE (such as acknowledgment or reply) or possibly an embedded game (e.g. a query may be followed by a clarification sub-dialogue) [37]. Because both of the users are getting interruptions from the *interruption task*, the primary communication strategy is taking turns describing the current Tangram shapes. This was useful for the purpose of our data collection to get a more dyadic conversation flow from the exchange.

To introduce more complexity into the Tangram task, the set of Tangram objects generated for each column are similar in appearance. For instance, objects that look like humans were generated together to avoid dialogue exchanges such as, dog, person, boat, square. There are a total of 250 Tangram images (50 images × 5 classes).

From this data collection, the raw audio of push-to-talk (PTT) audio, task boundary timestamps, and the class of Tangram shapes are generated for each human-human task. Although this is the data used for the overall analysis, information such as the Tangram score (how many Tangram columns were correctly coordinated), time-of-completion (TOC) for each Tangram subtask in each team run is also available in this corpus. Overall this is a corpus composed of audio data, annotated task performance and task boundaries data that can be used to assess various aspects of communication and performance within multi-human,

multitasking interactions such as grounding, dialogue turn-tasking, and teaming interaction.

3.2 Uncertainty Map Task - Human-Human Interaction

The Uncertainty Map Task (UMT) is the second human-human task. The UMT is similar to the Tangram task in that there is a succession of tasks within a single experimental run for a team. In the UMT, two users communicate via a push-to-talk (PTT) to describe their respective interfaces, ground their knowledge, and identify a target house. In the UMT there are 4 aerial maps and 12 houses for each aerial map ($12 \times 4 = 48$ target houses). Each target house has 4 street-view different perspectives. The UMT is comprised of 4 different tasks that were integrated into the design of the experiment in order to establish complexity. A succession of tasks were presented to the team until the team has received 10 randomly generated tasks.

The four UMT tasks include:

AERIAL TARGET/ STREET VIEW IDENTIFICATION - One user is given the target *aerial view* map and the number of the target house. The partner is given three different street view houses with two different perspectives and must identify the target house their partner is describing from the target aerial view and select the target house.

STREET VIEW TARGET/AERIAL IDENTIFICATION One user is given the target *street view* house and their partner is given an aerial view map with different houses labeled from 1 to 12. The user with the street view target interface is tasked with describing their house and the partner must identify which house on the aerial view the partner is referring.

STREET VIEW TARGET/STREET VIEW IDENTIFICATION One user is given the target *street view* house. The partner has three different street view houses from two different perspectives and must select the target house in which their partner is describing.

HOUSE IDENTIFICATION Both users have two *street view* pictures of the target house from two different perspectives in addition to an aerial view of the same target house. The teammates are tasked with identifying the target house based on the street view and aerial view perspectives.

Figure 5 illustrates the GUI for the *AERIAL TARGET/STREET VIEW IDENTIFICATION* task and Fig. 6 is a sample dialogue for the AERIAL TARGET - STREET VIEW IDENTIFICATION task within the Uncertainty Map Task.

In the dialogue illustrated in Fig. 6, Teammate 1 has the target aerial view and Teammate 2 has to identify the corresponding street view house. Teammate 1 has a GUI similar to the one illustrated in the left pane of Fig. 5 and Teammate 2 has a GUI similar to that illustrated in the right pane of Fig. 5. Within the dialogue (A1) is Teammate 1 explaining that he/she is answering a query related to the human-machine *interruption task*. Teammate 1 proceeds to inform



The Target House is 1
(a) Aerial Target GUI



(b) Street View Identification GUI

Fig. 5. Uncertainty map task - AERIAL TARGET/STREET VIEW ID

Teammate 1: okay um let me see sorry I was answering a query **clicking tongue** okay I have an overhead this house is on a corner and it has a sidewalk in front of it and a walkway that intersects the sidewalk um there's a tree on one side of it and then nothing on the other side of it basically the hou.. the house has like a light grey roof and it looks like it has probably a bunch of like landscaping around the prop like right around the house line is like a lot of bushes or like mulch or something right all around the like house line (**A1**)

Teammate 2: is there a sidewalk on both sides of it or just on one side (**A2**)

Teammate 1: it's on both sides of it and the sidewalk like crosses at the corner I don't know if you can see that but it crosses so it makes an X at the corner then it makes an X where or cross where the um walkway from the house intersects the sidewalk (**A4**)

Teammate 2: That helped I got it (**A4**)

***Both users press the DONE button to move on to the next subtask (**A5**)*

Fig. 6. An example dialogue from the AERIAL TARGET - STREET VIEW IDENTIFICATION task within the uncertainty map task

his/her teammate that he/she has an “overhead” view and then describes the target house from the perspective of a view similar to the left pane in Fig. 5. (**A2**) Teammate 2 asks a clarification question regarding the number of sidewalks on the side of the house. In utterance (**A3**), Teammate 1 responds to his/her teammate’s question. In (**A4**) Teammate 2 confirms that he/she now understands. Teammate 2 would more than likely subsequently select the check box for the row of houses he/she believes to be the target house from the perspective of the street view from the present three options of houses in the right pane of Fig. 5. Finally in (**A5**) both teammates press the DONE button indicating the comple-

tion of the current *AERIAL TARGET - STREET VIEW IDENTIFICATION* task.

3.3 Interruption Task - Human-Machine Interaction

Along with the aforementioned human-human tasks (Tangram and Uncertainty Map Task), teammates were simultaneously engaged in a human-machine task or the *Keeping Track Task* which is a task that involves the teammates receiving *interruptions* or synthesized audio of status updates and queries related to varying unmanned aerial vehicle (UAV) states. The Tangram-Keeping Track Task and UMT-Keeping Track Task are both examples of the simulated multi-human, multitasking human machine teaming tasks simulated in this research. For the *intelligent interruption system*, the interruption timing decisions are based on modeling the push-to-talk audio stream of the human-human task and predicting the time to send UAV updates or related queries. Both users receive 3–5 updates about various UAV states before being queried about the current state of a UAV previously presented. For example, a user is only required to keep track of 3–5 UAV states prior to a query. After a user receives a query and responds, a different set of 3–5 UAV states is presented. Below is an example of an Update/Query block:

- **Update I:** Hawk-88’ LOCATION is Point Bravo.
- **Update II:** Raven-3’s FUEL-LEVEL is 30%
- **Update III:** Falcom-11’s ALTITUDE is 1900 ft.
- **Query:** What is Raven-3’s current FUEL-LEVEL?

This is an example of a 3-block set of updates where the query refers to the UAV RAVEN-3 and asks about its current FUEL-LEVEL. Once a pre-specified number of UAV queries is sent to both users, the overall dual Tangram/Keeping Track task is over. For the dual UMT/Keeping Track task, regardless of the number of queries that are sent, the task ends when a pre-specified number of successive UMT tasks has been completed regardless of the number of updates/queries sent to each user. On average for the UMT/Keeping track task, a completion of 10 UMT subtasks results on average in each teammate receiving approximately 8–12 interruptions comprised of 6–9 updates and 2–3 queries. Each of the updates or queries (interruptions) is about 1–2 s in duration. Teammates press the “Answer Query” button illustrated in Fig. 3 to answer the queries and must response to a query before another interruption (UAV update) is sent otherwise any late response to a query will be scored as incorrect. There is an audio/visual presentation of the interruptions within the human-machine task because the human-human task is an audio/visual task and the objective is to have teammates attend to both task equally. Because the primary emphasis of this work is on interruption timing, rather than the presentation of the interruption, there is no analysis of how the interruption overlaps with the push-to-talk dialogue, but is present in the data collection via the audio channel data and can be explored in future work.

4 Task Boundary Prediction via Prosodic Speech Model

The proposed *intelligent interruption system* integrated into the multi-human, multitasking human-machine teaming interaction is a *Prosody-Only Task Boundary Model*. This prosody modeling component takes the raw audio of push-to-talk (PTT) utterances from both teammates engaged in the human-human task, maps this audio into a prosodic feature space, and uses a binary classification model (Task Boundary model) to detect the presence or absence of a task boundary based on the push-to-talk utterance that proceeds in order to make interruption decisions.

4.1 Data and Features

The predictors in this classification problem are prosodic features extracted as a 989-dimensional feature vector from the emotion detection feature set in OpenS-mile [38]. These features are derived from the raw audio of the push-to-talk utterances within the human-human tasks. These features are a composition of the utterance signal energy, loudness, Mel-/Bark-/Octave-spectra, MFCCs, PLPs, Pitch, voice quality (jitter, shimmer), formants, LPCs, Linear Spectral Pairs (LSPs), and spectral shape descriptors. Additionally, statistical functions or feature summaries are included in the feature set: means/extremes, moments, segments, samples, peaks, linear and quadratic regressions, percentiles, durations, onsets, DCT coefficients, and zero-crossing. Each utterance is sampled at 32 K and the features are extracted from the audio partitioned into 40 ms windows with 10 ms in overlap, common in audio processing and modeling.

Three different datasets are used to evaluate the performance of the *Prosody-Only Task Boundary Model*:

MODEL: This dataset is composed of a preliminary data collection from both the Tangram and the Uncertainty Map Task. The models built from this dataset are integrated into the REAL-TIME implementation of the Prosody-Only Task Boundary Model in both the Tangram and Uncertainty Map Task interactions.

REAL-TIME: The Tangram and Uncertainty Map tasks are run with the Prosody-Only Task Boundary model from the data in MODEL integrated into the interaction as the *intelligent interruption system*, an autonomous mechanism of making interruption decisions based on correctly predicting a task boundary from a push-to-talk utterance.

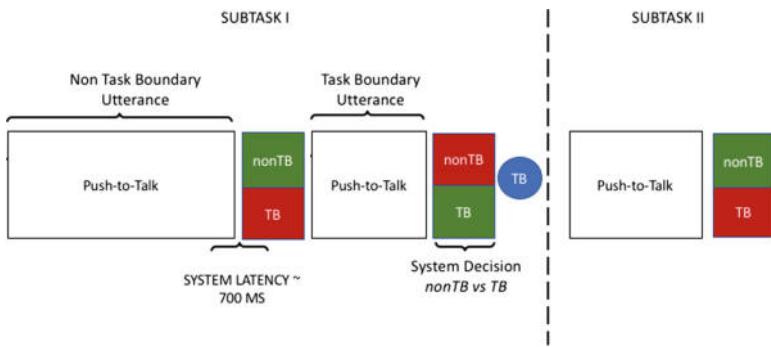
ALL: This dataset is a combination of the MODEL dataset and REAL-TIME dataset. The push-to-talk utterances in the REAL-TIME dataset are labeled in accordance with the labeling process of the MODEL dataset and the two are concatenated to form an ALL dataset for both the Tangram and Uncertainty Map tasks.

Table 1 illustrates all the data from the aforementioned datasets.

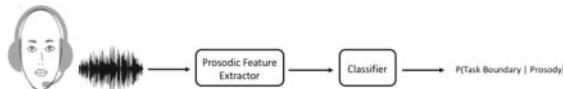
An illustration of the information flow of the real-time system can be illustrated in Fig. 7.

Table 1. Human-human task dataset summary

Human-human task	Dataset	Push-to-talk (N)	Push-to-talk audio (h)
TANGRAM	MODEL	10,300	17
	REAL-TIME	2,016	4
	ALL	12,316	21
UMT	MODEL	4,180	13
	REAL-TIME	4,478	11
	ALL	8,658	24

**Fig. 7.** Information flow for the human-human task of all data collection interactions to illustrate real-time system decisions based on the human-human task

This illustration is similar to Fig. 2 except it shows the human-human task communication flow within the context of how the real-time system makes decisions. A decision about the task boundary is made after every push-to-talk utterance. In the real-time system, a push-to-talk utterance is sent to a prosody feature extractor module which extracts features using OpenSmile [38]. The task boundary model is integrated into the Tangram and Uncertainty Map tasks where the interruption timing decisions related to the interruption task are based on the detection of a task boundary by leveraging prosodic information from the primary Tangram task. The system extracts the push-to-talk utterances from the primary task, generates a 989-dimensional feature vector from OpenSmile [38], and makes decisions on the probability of a task boundary using a Random Forest [39] model based on the data from the MODEL dataset. This process is illustrated in Fig. 8.

**Fig. 8.** Prosody-only task boundary model

From Fig. 7 there are several things to note in terms of how the real-time prosody-only model is evaluated. First the ground truth task boundary is a timestamp associated with a button click from the two teammates who agree they completed the current human-human task. All utterances that proceed this timestamp are task boundary utterances and all others are non-task boundary utterances. A task boundary detection decision is made after every utterance so the red boxes illustrate the mistakes the system can make (false positives and false negatives) and the green boxes illustrate the correct task boundary detection decisions (true positives and true negatives). Finally the prosody-only task boundary model is tractable in real-time because of the low average latency, approximately 700 ms, for making detection decisions from the prosody data of an utterance. More specifically, from the time the system received an utterance, processed the prosody of the audio, and made a decision to interrupt based on the detection of a task boundary took approximately 700 ms which is an indicator of how close the system is making decisions to its intended point of interruption. The smaller this latency value, the better especially in fast pace interactions where it is ideal for the system to make interruption decisions closest to its intended point in time.

4.2 Method

A preliminary proposition for predicting candidate interruption timings is via task boundary detection. A *task boundary* is defined within the context of the data collection as a timestamp associated with when two users collaborating on a human-human task pressed the DONE button as an indicator of completing a task. Utterances preceding these points are labeled as task boundaries (TB) and others are labeled as non task boundaries (nonTB).

The use of only prosodic information to infer task boundaries is an exploratory measure of how well one can do in predicting task boundaries using only features derived from the raw audio. Making predictions from the raw information within a communication channel has potential for quick data processing and modeling, but may hinder detection accuracies. Several binary classifiers that use supervised learning techniques can be used to solve this problem. The Random Forest [40] is selected as the task boundary classifier using specifically the implementation by Banerjee [39]. Preliminary studies [41] indicate this modeling method only slightly outperforms other methods such Nave Bayes [42] and Support Vector Machines [43].

Decision trees are a popular method for various machine learning tasks. In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, for instance, they may have a low bias, but very high variance. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. [36] This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance of the final model.

It was hypothesized that classification techniques that model predictor dependencies will perform better at predicting task boundaries, similar to other boundary detection work [44, 45] using prosody which illustrates the combination of pitch, energy, and their contours are useful at sentence and discourse boundary detection. Additionally, it was hypothesized that utterances preceding boundaries will be shorter in duration and lower in energy corresponding to confirmation utterances such as “done,” “got it,” “ok,” “finished,” “copy that”. These utterances can be confused with backchannels, shorter descriptions of shapes, or turn-taking confirmations in the middle of a task. With this confusability, there is the potential for more false positives.

A 10-fold cross-validation method was used to generalize each model. For a 10-fold cross-validation, the entire dataset is divided into 10 random subsets of the data which was then partitioned 90–10% on the train-test data so that $10 \times (\text{number of samples in the test set}) = \text{entire dataset}$. For each fold, the training set was resampled to a uniform distribution of class labels and the class distribution was maintained for the test set. The rationale for this is that within the UMT dataset there is an 85–15% on negative and positive classes. Without balancing the classes in the training data, one runs the risk of only learning the dominant class. The test data remains the same to reflect the actual class distribution of the task. Overall the objective is to see how well a task boundary detection algorithm can perform via a prosody-only model by validating the model offline and finally integrating the model into a live system and evaluating its performance.

4.3 Results

The first analysis was the performance of the Tangram MODEL and UMT MODEL results to get a preliminary idea of how well a Prosody-Only Random Forest classification model could do at discriminating task boundaries from non-task boundaries. In these preliminary results, we used the area under the curve (AUC) metric from the receiver operating characteristics (ROC) as a performance metric for each tasks’ MODEL data. The AUC metric is the trade-off between the true-positive rate (TPR) and the false-positive rate (FPR) or how well the systems predicts each class with respect to the presence of that class within the interaction. An AUC of 100% is perfect classification. The results are presented in Fig. 9 where there is an AUC for the Tangram task (96%), UMT (90%), and a GLOBAL (96%) dataset which is the concatenation of both datasets.

The rationale for extracting the ROC characteristics for the MODEL is to get an optimal operating point for each of the task to make better decisions in developing the real-time *Prosody-Only Task Boundary Model*. For the Tangram task, the optimal operation threshold of the classifier is 0.48 so a classification output probability of greater or equal to 0.48 is a task boundary, otherwise a non-boundary. This was adjusted to 0.5 in the real-time system. For the Uncertainty Map Task the operation threshold of the classifier is 0.58 so a classification

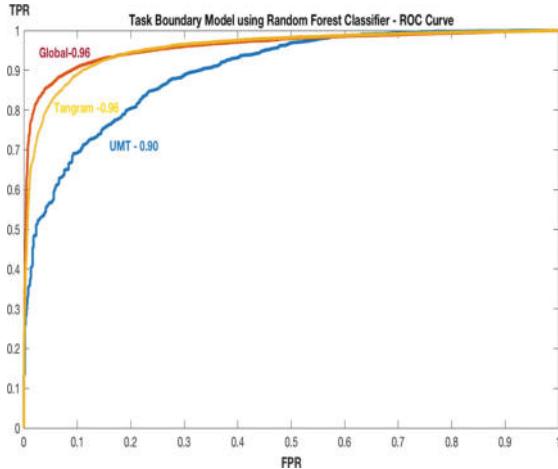


Fig. 9. ROC curve of the tangram and UMT

output probability of greater or equal to 0.65 is a task boundary, otherwise a non-boundary. This was adjusted to 0.65 in the real-time system.

The evaluation of these models was used as a potential indicator of how well the *intelligent interruption system* would integrate into their respective tasks. Since the GLOBAL model do not seem to perform better than the Tangram task model in which the GLOBAL dataset was comprised of 70% of the data was from the Tangram task and only 30% was from the UMT task. With this information, we decided to use the Tangram MODEL within the Tangram real-time system and the UMT MODEL in the UMT real-time system to get a baseline for how well the data used to model the respective tasks performed in the exact same interaction.

Table 2. Prosody-only task boundary modeling results via random forest classifier

	Data	Precision (%)	Recall (%)	F1 (%)	AUC (%)
TANGRAM	ALL	84.3	92.7	88.2	96.0
	MODEL	85.0	91.8	88.3	96.1
	REAL-TIME	79.1	70.8	74.7	–
UMT	ALL	58.6	72.6	64.9	89.9
	MODEL	74.2	47.2	57.7	90.2
	REAL-TIME	44.7	75.6	56.2	–

The results from the ALL, MODEL, and REAL-TIME performance is summarized in Table 2 and a comparison of the MODEL and REAL-TIME F1 scores is summarized in Fig. 10. It should be noted that the ALL model is just to give

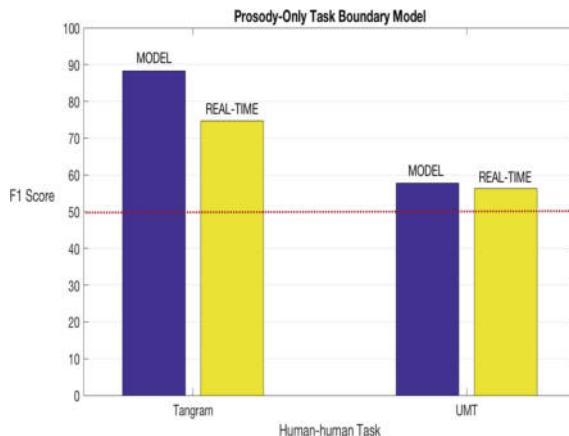


Fig. 10. Comparison of the F1 scores for the MODEL data and REAL-TIME data: Since the F1 score is the average between the RECALL and PRECISION, it is used as the primary metric in identifying how well the model and system does as predicting task boundaries that are present (RECALL) and not imagining task boundaries (PRECISION)

us general idea of how well the Tangram and UMT models perform base on all the data collected.

4.4 Discussion

The MODEL and REAL-TIME data for both tasks give an idea of how well the data used to build the real-time system performs and how well the integration of the MODEL into the REAL-TIME system performance. One observation in comparing the PRECISION and RECALL of the MODEL and REAL-TIME systems for both tasks in Table 2 is that the type of errors the system makes changes from the MODEL to the REAL-TIME system. This is probably attributed to the decisions of the optimal operating point chosen for the real-time system based on the MODEL. Choosing the optimal operating point is more of an executive decision based on which errors one would want the system to be more sensitive to. For the Tangram MODEL, the optimal operating point was 0.48, but in the real-time, it was adjusted to 0.5 which may have caused the real-time system to classify the non-task boundaries as task boundaries, causing more false positives. Additionally, for the UMT MODEL, the optimal operating point was 0.58 and was adjusted to 0.65 because in a pilot study, keeping the optimal operating point at 0.58 was resulting in far too many false-negatives so it was adjusted. With this adjustment, it caused the system to move forward in delivering interruptions but at the cost of detecting task boundaries or interruption points that were not really there and also causing more false-positive.

Additionally, overall the Prosody-Only task boundary model for the Tangram task outperforms the UMT overall. This could be attributed to the fact

that the UMT tasks in composed of several different tasks which introduce more variability in the overall task dialogue whereas since the tasks in the Tangram experiment are similar in nature, the overall interaction and dialogue may be a bit more predictable. This does not however in anyway negate the usefulness of task boundary modeling strategies. For tasks that have more predictable dialogues, such modeling strategies may be useful. Additionally, one can imagine cases in which team members interact on a task over a long period of time and begin to develop more predictable exchanges. Within these contexts, task boundary model integration into the overall intelligent interruption system could be useful.

In analyzing the important detection features in the Random Forest algorithm, the loudness_Pcm_Min_Pos feature is a key indicator in discriminating these classes where 61.3% of classification decisions were based on this cue for the Tangram task. This does not necessarily mean that how loudness is perceived is a predictor of task boundaries, but could suggest that information correlated with energy and potentially duration could be useful in predicting task boundaries. This suggestion is corroborated by the results from [41] that show that augmenting the prosody-only model with duration rules results in a 10.84% improvement in overall accuracy. From these results duration and energy based cues seem to be potential indicators of predicting task boundaries within multi-human, multitasking interactions with more predictable dialogue exchanges, but may not be as useful in interaction with much more robust dialogue as indicated by the overall performance of the prosody-only task boundary model within the UMT task which only slightly outperforms a 50% random baseline.

Prosody is proposed as a useful speech information stream because its features process rapidly and can be implemented in real-time dialogue systems. This was illustrated by evaluating the average system latency as 700 ms which allowed perception of a real-time system from a human perspective. Overall the results indicate that extracting prosody from the natural speech of multi-human, multitasking interaction results in a lower system latency and for interactions with predictable dialogues, can result in accurate task boundary detection as candidates for interruption timings. The task boundary detection performance is degraded for robust dialogues and interactions.

5 Conclusion and Future Work

Overall in this research two multiple humans, multitasking human-machine teaming scenarios are simulated and aid in evaluating mapping low-level speech information onto higher constructs indicative of interruptibility such as task boundaries via a Task Boundary Prosody Model which is evaluated on (1) how they perform within two different simulated human-machine teaming scenarios and (2) the speed vs. accuracy tradeoffs for each module in addition to other limitations. Overall the Task Boundary Prosody Model is tractable within a real-time system because of the low-latency in processing prosodic information but is less accurate at predicting task boundaries except within human-machine interactions with more predictable dialogue. The results from the Task Boundary The

Prosody-Only Model gives some indication that utterance energy and duration may be good predictors of a task boundaries because these utterances could be characteristics of confirmations or knowledge aligning indicators that a user is ready to continue to the next task. Unfortunately, this was only evident in tasks whose dialogue is simple such as the Tangram task. Task boundary confirmation utterances can be confused with backchannels and mid-task confirmations, which was evident in evaluating the Prosody-Only Task Boundary Model within the Uncertainty Map Task (UMT). Additionally, within UMT, longer utterances were present before task boundaries with teammates explaining their decisions and adding other caveats that were indicative of uncertainty before continuing to a new task.

In future work, the use of lexical information could aid in task boundary detection performance, but this may come at the expense of system latency. To extract lexical information from the dialogues of these multi-human, multitasking interactions, it is necessary to map the raw audio into some word feature space via automatic speech recognition (ASR) and then perform lexical modeling such as N-gram modeling. Each processing step adds latency to the overall system design so it is crucial to evaluate the latency within each step to see how the latency accumulates and the feasibility within real-time. The aim is to extend this work by exploring how to detect task boundaries as candidate interruption timings using lexical, dialogue, topic, and part-of-speech modeling techniques. In addition to these techniques, a multimodal system is proposed to augment the contribution of various speech information streams. The long-term goal is to design and evaluate an *intelligent interruption system* that can be integrated into multi-human, multitasking human-machine teaming interactions and provide users with information and notifications within these interactions without hindering the overall multitasking performance.

References

1. Adams, J.: Human machine teaming research (2017)
2. Steifik, M.: Half-human, half-computer? meet the modern centaur (2017)
3. Arroyo, E., Selker, T.: Attention and intention goals can mediate disruption in human-computer interaction, pp. 454–470 (2011)
4. Bailey, B.P., Konstan, J.A.: On the need for attention-aware systems: measuring effects of interruption on task performance, error rate, and affective state. Comput. Hum. Behav. **22**(4), 685–708 (2006)
5. Czerwinski, M., Cutrell, E., Horvitz, E.: Instant messaging and interruption: influence of task type on performance. In: OZCHI 2000 Conference Proceedings, vol. 356, pp. 361–367 (2000)
6. Monk, C.A., Boehm-Davis, D.A., Gregory Trafton, J.: The attentional costs of interrupting task performance at various stage. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 46, pp. 1824–1828. SAGE Publications, Los Angeles, CA (2002)
7. Czerwinski, M., Cutrell, E., Horvitz, E.: Instant messaging: effects of relevance and timing. People and Computers XIV: Proceedings of HCI **2**, 71–76 (2000)

8. Adamczyk, P.D., Bailey, B.P.: If not now, when?: the effects of interruption at different moments within task execution. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 271–278. ACM (2004)
9. Zijlstra, F.R.H., Roe, R.A., Leonora, A.B., Krediet, I.: Temporal factors in mental work: effects of interrupted activities. *J. Occup. Organ. Psychol.* **72**(2), 163–185 (1999)
10. Scott McCrickard, D., Chewar, C.M., Somervell, J.P., Ndiwalana, A.: A model for notification systems evaluation?assessing user goals for multitasking activity. *ACM Trans. Comput.-Hum. Interact. (TOCHI)*, **10**, 312–338 (2003)
11. McFarlane, D.C., Latorella, K.A.: The scope and importance of human interruption in human-computer interaction design. *Hum.-Comput. Interact.* **17**(1), 1–61 (2002)
12. Dabbish, L., Kraut, R.E., Controlling interruptions: awareness displays and social motivation for coordination. In: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, pp. 182–191. ACM (2004)
13. Latorella, K.A.: Investigating interruptions: an example from the flightdeck. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 40, pp. 249–253. SAGE Publications, Los Angeles, CA (1996)
14. Kreifeldt, J.G., McCarthy, M.E.: Interruption as a test of the user-computer interface (1981)
15. Rubinstein, J.S., Meyer, D.E., Evans, J.E.: Executive control of cognitive processes in task switching. *J. Exp. Psychol.: Hum. Percept. Perform.* **27**(4), 763 (2001)
16. Altmann, E.M., Trafton, J.G.: Task interruption: resumption lag and the role of cues. In: Proceedings of the Cognitive Science Society, vol. 26 (2004)
17. Sasse, A., Johnson, C., et al.: Coordinating the interruption of people in human-computer interaction. *Hum.-Comput. Interact.* **99**, 295 (1999)
18. Horvitz, E.C.M.C.E., Notification, disruption, and memory: effects of messaging interruptions on memory and performance. In: Human-Computer Interaction: INTERACT'01: IFIP TC. 13 International Conference on Human-Computer Interaction, 9th–13th July 2001, Tokyo, Japan, p. 263. IOS Press (2001)
19. McFarlane, D.C.: Interruption of people in human-computer interaction: a general unifying definition of human interruption and taxonomy. Technical Report, Office of Naval Research, Arlington VA (1997)
20. Bailey, B.P., Iqbal, S.T.: Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Trans. Comput.-Hum. Interact. (TOCHI)*, **14**(4), 21 (2008)
21. Iqbal, S.T., Bailey, B.P.: Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In: CHI'05 Extended Abstracts on Human Factors in Computing Systems, pp. 1489–1492. ACM (2005)
22. Peters, N., Romigh, G., Bradley, G., Raj, B.: When to interrupt: a comparative analysis of interruption timings within collaborative communication tasks. In: Advances in Human Factors and System Interactions, pp. 177–187. Springer (2017)
23. Horvitz, E., Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 159–166. ACM (1999)
24. Horvitz, E., Apacible, J.: Learning and reasoning about interruption. In: Proceedings of the 5th International Conference on Multimodal Interfaces, pp. 20–27. ACM (2003)
25. Horvitz, E., Koch, P., Kadie, C.M., Jacobs, A.: Coordinate: probabilistic forecasting of presence and availability. In: Proceedings of the Eighteenth Conference on

- Uncertainty in Artificial Intelligence, pp. 224–233. Morgan Kaufmann Publishers Inc. (2002)
- 26. Begole, J.B., Matsakis, N.E., Tang, J.C.: Lilsys: sensing unavailability. In: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, pp. 511–514. ACM (2004)
 - 27. Fogarty, J., Lai, J., Christensen, J.: Presence versus availability: the design and evaluation of a context-aware communication client. *Int. J. Hum.-Comput. Stud.* **61**(3), 299–317 (2004)
 - 28. Iqbal, S.T., Bailey, B.P., Leveraging characteristics of task structure to predict the cost of interruption. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 741–750. ACM (2006)
 - 29. Miyata, Y., Norman, D.A.: Psychological issues in support of multiple activities. In: User Centered System Design: New Perspectives on Human-Computer Interaction, pp. 265–284 (1986)
 - 30. Bannon, L., Cypher, A., Greenspan, S., Monty, M.L.: Evaluation and analysis of users' activity organization. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 54–57. ACM (1983)
 - 31. Miller, S.L.: Window of opportunity: Using the interruption lag to manage disruption in complex tasks. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 46, pp. 245–249. SAGE Publications, Los Angeles, CA (2002)
 - 32. Peters, N., Romigh, G., Bradley, G., Raj, B., A comparative analysis of human-mediated and system-mediated interruptions for multi-user, multitasking interactions. In: International Conference on Applied Human Factors and Ergonomics, pp. 339–347. Springer (2017)
 - 33. Mushin, I., Stirling, L., Fletcher, J., Wales, R.: Discourse structure, grounding, and prosody in task-oriented dialogue. *Discourse Process.* **35**(1), 1–31 (2003)
 - 34. Mixdorff, H., Quantitative analysis of prosody in task-oriented dialogs. In: Speech Prosody 2004, International Conference (2004)
 - 35. Syrdal, A.K., Kim, Y.-J.: Dialog speech acts and prosody: considerations for TTS. In: Proceedings of Speech Prosody, pp. 661–665 (2008)
 - 36. Hastie, H.W., Poesio, M., Isard, S.: Automatically predicting dialogue structure using prosodic features. *Speech Commun.* **36**(1), 63–79 (2002)
 - 37. Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J.C., Anderson, A.H.: The reliability of a dialogue structure coding scheme. *Comput. Linguist.* **23**(1), 13–31 (1997)
 - 38. Eyben, F., Weninger, F., Gross, F., Schuller, B.: Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 835–838. ACM (2013)
 - 39. Banerjee, S.: Simple example code and generic function for random forests. <https://www.mathworks.com/matlabcentral/fileexchange/51985-simple-example-code-and-generic-function-for-random-forests> (2016)
 - 40. Ho, T.K.: Random decision forests. In: Proceedings of the Third International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282. IEEE (1995)
 - 41. Peters, N., Raj, B., Romigh, G.: Topic and prosodic modeling for interruption management in multi-use multitasking communication interactions. In: AI-HRI (2018)
 - 42. Russell, S., Norvig, P., Intelligence, Artificial: “A modern approach”. In: Artificial Intelligence, vol. 25, p. 27. Prentice-Hall, Englewood Cliffs (1995)

43. Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learn.* **20**(3), 273–297 (1995)
44. Swerts, M., Ostendorf, M.: Prosodic and lexical indications of discourse structure in human-machine interactions. *Speech Commun.* **22**(1), 25–41 (1997)
45. Shriberg, E., Stolcke, A., Hakkani-Tür, D., Tür, G.: Prosody-based automatic segmentation of speech into sentences and topics. *Speech Commun.* **32**(1), 127–154 (2000)



Computer Science Education: Online Content Modules and Professional Development for Secondary Teachers in West Tennessee—A Case Study

Lee Allen^(✉)

University of Memphis, Memphis, TN 38152, USA
Lee.Allen@memphis.edu

Abstract. With ongoing efforts in the United States to further develop the availability of computer science education in the public schools, federal, state, and local educational agencies are increasing efforts to encourage and promote the inclusion of computer science and programming skills in the middle school curriculum (Grover and Pea in Comput. Sci. Educ. 25:199–237, 2015) [1]. The goal for the Online Content Modules: Computer Science in the Middle Grades project was the development of five online content modules with a focus on computer science instruction in three public school districts in West Tennessee, and disseminated through a week-long professional development summer institute.

Keywords: Online learning · STEM · Computer science · First section

1 Introduction

Far too often, the “T” in STEM (Science, Technology, Engineering, and Math) education is provided only nominal support via the use of the latest technologies available in classroom settings in searching for information via the Internet and using common productivity applications such as word processing, spreadsheet creation, and presentation software [2]. In recent years, increasing interest and effort has been dedicated to introducing computer programming, often referred to as “coding”, to students in K-12 settings, especially at the middle and high school level [3, 4]. While computer education in the upper secondary grades (10–12) is becoming more prevalent where additional instructional resources are available, the middle secondary grades (6–9) have lagged somewhat in introducing computer science to younger students [5]. However, several recent studies have provided evidence that introducing computer science and programming skills at the middle school level can help pave the way for students’ future educational and even career goals [1, 6–9].

In June 2016, President Obama announced the start of an educational initiative, Computer Science For All, whose stated purpose is to “empower all American students from kindergarten through high school to learn computer science... both educators and business leaders are increasingly recognizing that computer science (CS) is a ‘new basic’ skill necessary for economic opportunity and social mobility” [10]. With a

national effort to further the availability of computer science education, state and local educational agencies are increasing efforts in disseminating the inclusion of computer science and programming skills in the K-12 school curriculum [11]. Such efforts are based, in part, on the documented successes in integrating game-based programming and coding in STEM-based coursework for middle and high school students [12–15].

Due to the increasing emphasis and acknowledgement of the importance of introducing computer science education in the secondary school grades, the goal and focus of the Online STEM Content Modules: Computer Science in the Middle Grades project was to enhance computer science content taught in the Middle School (grades 6-8) Science, Technology, Engineering, and Mathematics (STEM) curriculum.

While the current Middle School Career and Technology Education Coursework provided by the [16] does not specifically identify Computer Science or computer programming/coding, aspects of computer science instruction are embedded in the Computer Applications and, to a certain degree, the STEM Designers coursework.

The STEM Designers course (the third course in the Middle School STEM sequence of Coursework) description cites the P21: Partnership for 21st Century Skills Framework for 21st Century Learning; the P21 Framework was developed as an ongoing collaborative effort “with input from teachers, education experts, and business leaders to define and illustrate the skills and knowledge students need to succeed in work, life and citizenship, as well as the support systems necessary for 21st century learning outcomes” [17]. The P21 Framework aspires to represent “both 21st century student outcomes... and support systems” (*ibid.*). Regarding computer education, the P21 Framework addresses several facets of desirable student proficiency achievement within the 21st Century Student Outcomes [17].

2 The Study

The primary goal of the Online Content Modules: Computer Science in the Middle Grades project was the development of five online content modules with a focus on computer science instruction in middle schools. These online modules were developed with the direct input of five identified expert secondary computer science teachers employed in—and identified by—two of the partner county school districts, plus one each from two municipal school systems. The development team met from January to May 2017 on one Saturday each month at the University of Memphis’ campus in Jackson, TN, for a total of five face-to-face meetings. Other activities and communications took place online via email and uploads to a mutually accessible web-based repository. The development team was also responsible for instruction using the online computer science modules during teacher professional development workshops.

Upon completion of the online Middle School Computer Science (MSCS) content modules, a program of workshops was provided for all secondary STEM and STEM-related teachers identified by—and recruited from—partnering school districts were conducted at the University of Memphis’ campus located in Jackson, TN from in June 2017. The summer institute workshops introduced the online MSCS content modules to all teacher/participants representing the three partner school districts and provided examples of how these can be integrated to enhance existing or new STEM and/or

computer science classes offered in the partnering district middle and high schools. Forty-three teacher participants attended the five-day workshop, with instruction based on the online modules provided by the five teachers who comprised the modules' development team.

Also, during the five-day summer institute workshop guest speakers representing leadership in the partner school districts, the TN Department of Transportation, the TN STEM Innovation Network (TSIN), the National Youth Cyber Education Program, the Robotics Education & Competition (REC) Foundation, and the West Tennessee STEM Hub provided informative presentations and demonstrations on various topics associated with STEM education throughout the state.

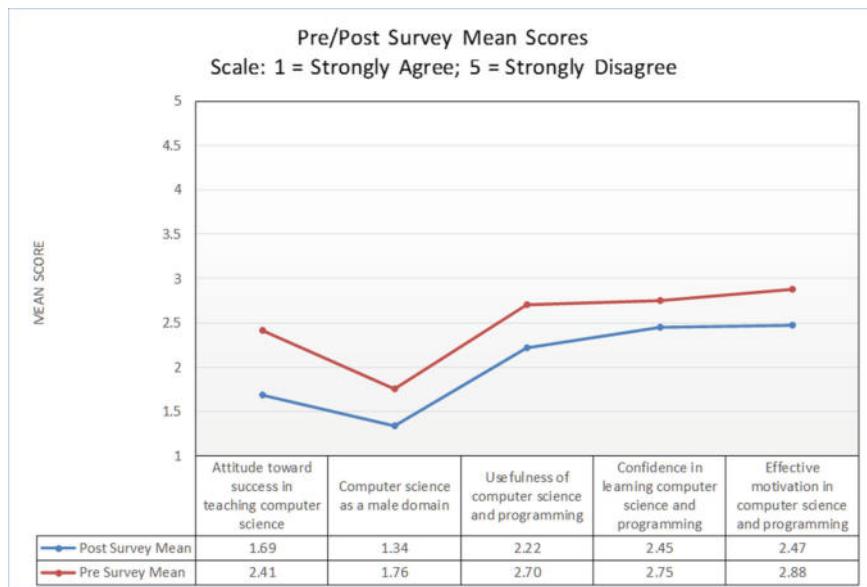
In recruiting teachers/participants from the partnering school districts, leadership for the three partner school districts were asked to provide email contact information for all secondary STEM and STEM-related (science, math, career and technology education) teachers from their school districts. The project director communicated with the teachers by emailing the information regarding the opportunity to participate in the June workshops and the compensation provided for participation. Of those who were contacted, 57 total responses were received; of those, 43 teachers participated in the summer institute.

The 43 teachers participating in the summer institute professional development workshops were administered pre-institute and post-institute assessments regarding their knowledge, skills, and attitudes towards computer science instruction in the classroom. A survey was developed to measure attitudes towards computer programming and computer science in general; this instrument was derived from the Fennema-Sherman mathematics attitudes scales [18], modified to reflect programming and computer science rather than mathematics. The survey consisted of 49 pre- and post-professional development questions requiring corresponding positive and negative statements to a Likert-type scale with 1 indicating they strongly agreed to a 5 indicating that they strongly disagreed with the statement. The survey consists of five subgroups, with several items in each category required interpreting the data in reverse, as they were written as negative items. The negative statements are reverse coded prior to summing the subscale scores. The survey uses five of the seven subscale categories used in the Fennema-Sherman instrument and, in addition, the survey starts with a statement concerning the participant's intent to teach computer science. The reliability of the instrument was evaluated for internal consistency of the subscales [19]. The final results are displayed in Table 1; a line graph to depict positive mean score changes for all categories are shown in Fig. 1.

In addition, a follow-up survey assessing classroom use of the computer science knowledge and skills acquired during the Summer Institute work sessions was distributed to the participant teachers to determine if, when, and how the online computer science modules' content was being taught in their classrooms. The online survey was made available to participants in October 2017; 29 of the 43 total participants responded to the second survey. Key results from the follow-up survey are shown in Figs. 2, 3, 4, 5, 6 and 7.

Table 1. Pre- and post- computer science teacher attitude survey

Sub-scales	Survey statement numbers	Pre	Post
Attitude toward success in teaching computer science	2–13	2.41	1.69
Computer science as a male domain	14–21	1.76	1.34
Usefulness of computer science and programming	22–29	2.70	2.22
Confidence in learning computer science and programming	30–39	2.75	2.45
Effective motivation in computer science and programming	40–50	2.88	2.47
Overall		2.50	2.03

**Fig. 1.** Pre- and post- computer science teacher attitude survey chart: mean scores

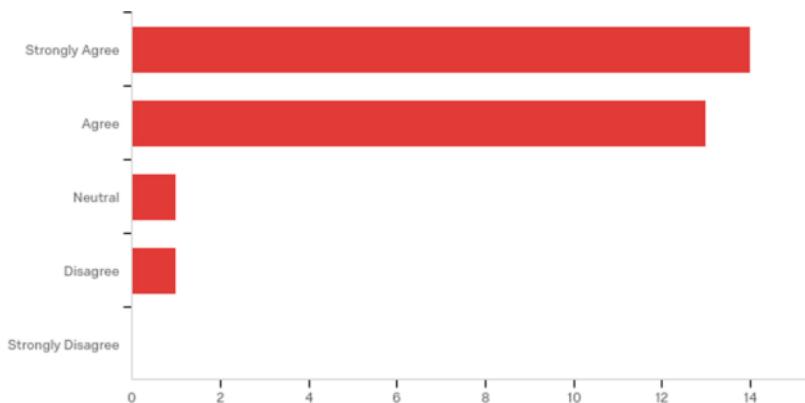


Fig. 2. Question 1: The Computer Science Summer Institute professional development activities enhanced my understanding of computer science and programming

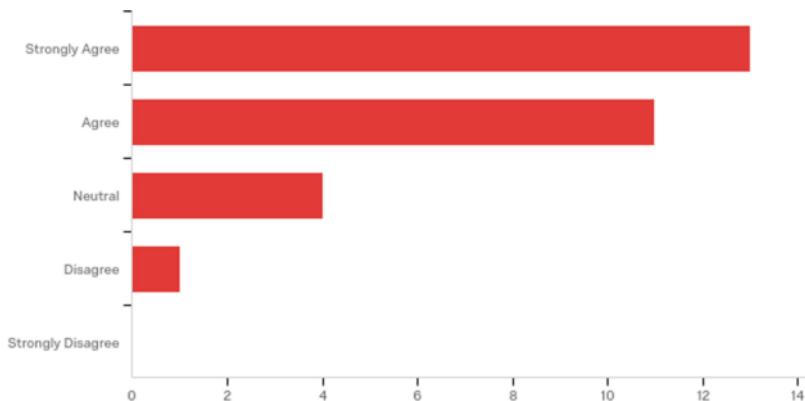


Fig. 3. Question 2: The Computer Science Summer Institute professional development activities enhanced my overall interest in computer science

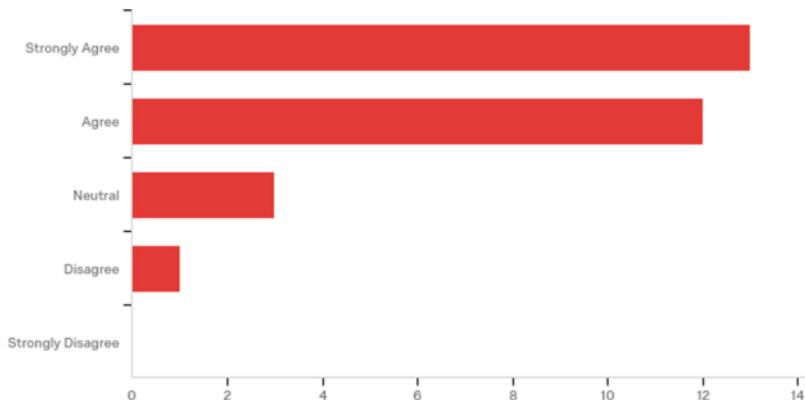


Fig. 4. Question 3: The Computer Science Summer Institute professional development activities increased my awareness of different computer science applications in the professional world

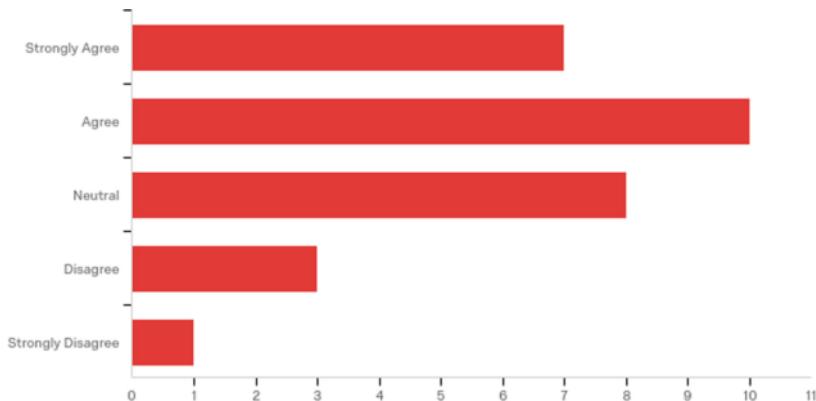


Fig. 5. Question 4: The Computer Science Summer Institute programming activities this summer were helpful and are useful in my teaching this fall

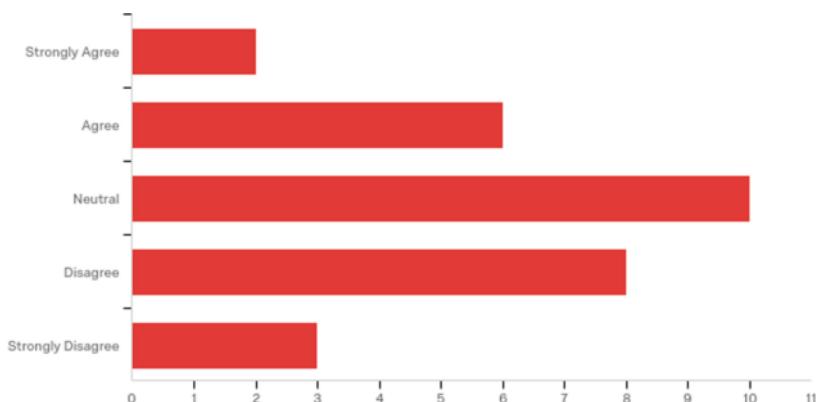


Fig. 6. Question 5: I integrate and use the Computer Science content modules available via BFK/TN and Summer Institute professional development activities in my teaching this Fall

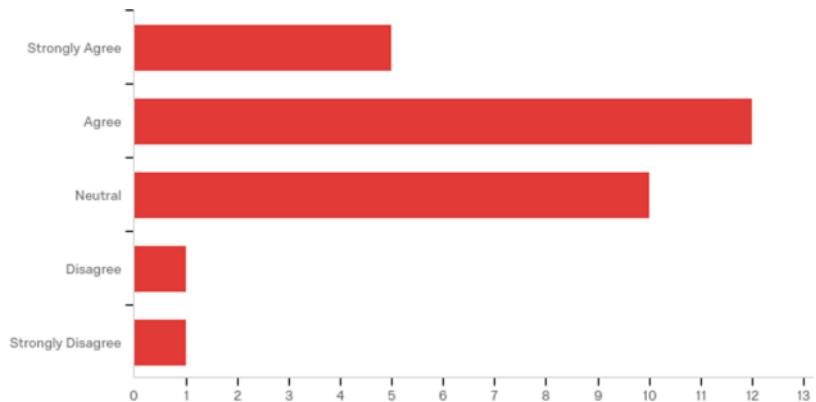


Fig. 7. Question 6: The online Computer Science content modules are more helpful than traditional textbooks to assist in my teaching Computer Science or other STEM-related topics

3 Findings

The purpose of the summer institute professional development pre- and post- survey was to measure if and how the participant teachers' attitudes towards computer science as an instructional as well as professional content area might have changed due to the use of the online computer science modules, in addition to exposure to other computer science content of the five-day summer institute. As can be viewed in Table 1, the difference between attributional positive attitudes towards computer science pre- and post-workshop responses overall was relatively small: Pre: 2.50 and Post: 2.03, or 0.47 difference towards a more positive attitude after the week-long institute.

Though still relatively minute, the most statistically significant increase in positive responsiveness was regarding the survey's sub-scale regarding the teachers' "Attitude toward success in teaching computer science"; the increase in average positive attitude from 2.41 to 1.69 (from Agree and Neutral, to Strongly Agree or Agree) was 0.72 points.

The most notable difference to the researcher was the slight gain in positive attitude (0.42) towards "Computer science as a male domain", from 1.76 to 1.34. This indicated that the responding teachers increased their perception that computer science as a profession or educational subject was predominantly a male-oriented area.

The follow-up survey was distributed to the 43 participants during the Fall semester was designed to assess the subsequent classroom use of the computer science knowledge and skills acquired during the Summer Institute work sessions. The survey was intended to also determine if, when, and how the online computer science modules were being used in the participants' classrooms. The online survey was made available to participants in October 2017; 29 of the 43 total participants responded to the second survey. The key results from the survey (as shown in Figs. 2, 3, 4, 5, 6 and 7) indicated the following.

In response to Question 1, "The Computer Science Summer Institute professional development activities enhanced my understanding of computer science and programming", 27 of the 29 respondents (93%) strongly agreed or agreed with that statement.

In response to Question 2, "The Computer Science Summer Institute professional development activities enhanced my overall interest in computer science", 24 of 29 respondents (83%) strongly agreed or agreed with that statement.

Responding to Question 3, "The Computer Science Summer Institute professional development activities increased my awareness of different computer science applications in the professional world", 25 respondents (86%) strongly agreed or agreed with the statement.

In response to Question 4, "The Computer Science Summer Institute programming activities this summer were helpful and are useful in my teaching this fall", 17 of the 29 respondents (59%) indicated that they strongly agreed or agreed with the statement.

Responding to the statement provided in Question 5, "I integrate and use the Computer Science content modules available via BFK/TN and Summer Institute

professional development activities in my teaching this Fall”, eight of the respondents (28%) agreed or strongly agreed, while 18 (62%) were neutral or disagreed with the statement.

In responding to the statement in Question 6, “The online Computer Science content modules are more helpful than traditional textbooks to assist in my teaching Computer Science or other STEM-related topics”, 17 respondents (59%) indicated that they agreed or strongly agreed with the statement, however ten (35%) provided neutral responses.

4 Conclusion

The summer institute professional development pre- and post- survey—was used to measure if and how the participant teachers’ attitudes towards computer science may have changed due to the use of the online modules and exposure to other computer science content during the five-day professional development sessions. While differences were statistically insignificant overall, the most surprising difference to the researcher was the slight *gain* in positive attitude (0.42) towards “Computer science as a male domain”, from 1.76 to 1.34. This indicated that the responding teachers, after five days of exposure to several women instructors and presenters specializing in coding, robotics and other STEM-related contents, *increased* their perception that computer science as a profession or educational subject was predominantly male-oriented. While it is not possible to determine from the survey alone, as ca. 80% of the responding teachers were women, there is a possibility that the rural and semi-rural—and thus more traditional—counties that comprised their workplaces and homes could posit an influence on attitudes regarding male and female roles in the workplace and school environments. Because the pre- and post- surveys were distributed after only a five-day professional development institute, a long-term study with participants with similar demographic backgrounds could prove interesting as a future study.

The results of the Fall semester follow-up survey, with 29 of the 43 participants responding to the statements assessing their use of the computer science modules and knowledge and skills acquired during the Summer Institute in their own classrooms, indicated that the participants were mostly satisfied with the content they were provided. However, somewhat unsurprisingly, a majority of the respondents also indicated that they were not using the online computer science content modules in their classrooms, and many were neutral, or uncommitted, in finding the modules to be more useful than traditional textbooks in their classrooms.

A component of this study that was not reported here—however important to include in this discussion—consisted of phone interviews with three teacher/participants in the Summer Institute. This qualitative data provided a possible glimpse into why at least some teachers in one of the districts represented (the largest) may not be using the online modules. All three interviewees indicated that the content they were allowed to present in their classrooms was tightly controlled and restricted by their districts and administrators, and the content taught must reflect the district-mandated curriculum as established and required by the TN State Department of Education.

The irony of this last finding is that the study being discussed here was funded by the United States' Department of Education via the Tennessee Higher Education Commission (THEC), in an ongoing effort to stimulate educational innovation and resources to improve students' understanding of STEM and STEM-related content in general, and, in this project's example, computer science in particular. If teachers are exposed to innovative, alternative resources—funded at the federal and state level—but subsequently not allowed to incorporate these in their classrooms, the question as to the ultimate purpose of these projects remains.

References

1. Grover, S., Pea, R., Cooper, S.: Designing for deeper learning in a blended computer science course for middle school students. *Comput. Sci. Educ.* **25**(2), 199–237 (2015)
2. Howley, A., Wood, L., Hough, B.: Rural elementary school teachers' technology integration. *J. Res. Rural Educ.* **26**(9), 1 (2011) (Online)
3. Richtel, M.: Reading, writing, arithmetic, and lately, coding. In: *The New York Times*, A1 (2014)
4. Settle, A., Franke, B., Hansen, R., Spaltro, F., Jurisson, C., Rennert-May, C., Wildeman, B.: Infusing computational thinking into the middle- and high- school curriculum. In: *Proceedings of the 17th ACM Annual Conference on Innovation and Technology in Computer Science Education*, pp. 22–27. ACM (2012)
5. Grover, S., Pea, R., Cooper, S.: Remedyng misperceptions of computer science among middle school students. In: *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*, pp. 343–348. ACM (2014)
6. Rodger, S., Dalis, M., Gadwal, C., Hayes, J., Li, P., Wolfe, F., Liang, L.: Integrating computing into middle school disciplines through projects. In: *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*, pp. 421–426. ACM (2012)
7. Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Gallagher, L.P.: Integration of technology, curriculum, and professional development for advancing middle school mathematics three large-scale studies. *Am. Educ. Res. J.* **47**(4), 833–878 (2010)
8. Taub, R., Armoni, M., Ben-Ari, M.: CS unplugged and middle-school students' views, attitudes, and intentions regarding CS. *ACM Trans. Comput. Educ. (TOCE)* **12**(2), 8 (2012)
9. Woolley, M.E., Rose, R.A., Orthner, D.K., Akos, P.T., Jones-Sanpei, H.: Advancing academic achievement through career relevance in the middle grades: a longitudinal evaluation of CareerStart. *Am. Educ. Res. J.* **50**(6), 1309–1335 (2013)
10. Whitehouse.gov.: Computer Science for All <https://www.whitehouse.gov/blog/2016/01/30/computer-science-all>. Last accessed 29 Aug 2016 (2016)
11. Cheung, A.C., Slavin, R.E.: The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: a meta-analysis. *Educ. Res. Rev.* **9**, 88–113 (2013)
12. Brown, Q., Mongan, W., Kusic, D., Garbarine, E., Fromm, E., Fontecchio, A.: Computer aided instruction as a vehicle for problem solving: scratch programming environment in the middle years classroom. College of Engineering, Drexel University, Philadelphia, PA (2013). http://www.pages.drexel.edu/~dmk25/ASSE_08.pdf. Last accessed 22 Sept 2016
13. Burke, Q.: The markings of a new pencil: introducing programming-as-writing in the middle school classroom. *J. Media Lit. Educ.* **4**(2), 121–135 (2012)

14. Repenning, A., Webb, D., Ioannidou, A.: Scalable game design and the development of a checklist for getting computational thinking into public schools. In: Proceedings of the 41st ACM Technical Symposium on Computer science education, pp. 265–269. ACM (2010)
15. Werner, L., Denner, J., Campe, S., Kawamoto, D.C.: The fairy performance assessment: measuring computational thinking in middle school. In: Proceedings of the 43rd ACM Technical Symposium on Computer Science Education, pp. 215–220. ACM (2012)
16. Tennessee Department of Education.: Tennessee Educator Acceleration Model: <http://teamtn.org>. Last accessed 8 Sept 2016 (2016)
17. Partnership for 21st Century Learning. Retrieved 9 Aug 2016 <http://www.p21.org/our-work/p21-framework> (2009)
18. Fennema, E., Sherman, J. A.: Fennema-Sherman mathematics attitudes scales. Instruments designed to measure attitudes toward the learning of mathematics by females and males. JSAS: Catalog Sel. Doc. Psychol. **6**(31), (Ms. No. 1225) (1976)
19. Williams, L., Wiebe, E., Yang, K., Ferzli, M., Miller, C.: In support of paired programming in the introductory computer science course. Comput. Sci. Educ. **12**(3), 197–212 (2002)



Enterprises and Future Disruptive Technological Innovations: Exploring Blockchain Ledger Description Framework (BLDF) for the Design and Development of Blockchain Use Cases

Hock Chuan Lim^(✉)

Faculty of Engineering and Information Sciences (FEIS),
University of Wollongong in Dubai (UOWD), Dubai, United Arab Emirates
hclim@uowdubai.ac.ae

Abstract. This digital age is viewed as a space that will herald greater innovation focus and spirit. Kindled by Blockchain Technologies (BT) or Distributed Ledger Technologies (DLT) potentials for disruptive innovations across a wide range of industries, the question of how enterprises can prepare for future, next-generation digital technological innovations is raised. This paper examines the current approaches in Enterprise Modelling (EM), Enterprise Architecture (EA) design and Enterprise Governance of IT (EGIT) and deploy them as useful models and tools to aid management of new innovative technologies. The scope is restricted to the domains of Blockchain or Distributed Ledger Technologies (DLT) as the underlying platform for technological innovations in modern enterprises. The contributions of this paper include: (a) mapping of EM/EA for new digital innovations exemplified by the case of DLT; (b) applying suitable EM/EA framework for design and building of blockchain use cases; (c) validating of the framework via participatory actions research method and (d) discussing challenges and research areas that impact enterprises and new disruptive digital innovations.

Keywords: Blockchain distributed ledger technology · Participatory action research · Enterprise modelling · Disruptive innovations · Use cases

1 Introduction

The uncertain future of business entities and enterprises is one aspect that continues to challenge leaders and managers in our innovation context. Enterprises face several jeopardies when it comes to dealing with disruptive innovations. On the one hand, there is the ignore and perish danger, and we do not have to look afar; enterprises that are not ready for disruptive and innovative technologies will have to pay a high price and likely will perish just as reported: "...Sperry Univac, Honeywell, Control Data, Digital Equipment Corp., Wang, Data General, Prime, Kodak, Polaroid, Lucent, Nortel, Compaq, Gateway, Lotus, Ashton Tate, Borland, Novell, Nokia, Tower Records, Borders, Barnes & Noble, and Blockbuster among companies that suffered because

they failed to craft an appropriate response to disruptive technology..." [1]. On the other hand, enterprises who judiciously allocate resources and focus to innovations may not necessarily do better as in "...executives of successful enterprises that carefully monitor and manage the near-term health of their established businesses while focusing adequate resources on disruptive technologies, could ultimately lead to the enterprises' downfall..." [2].

Making matter worse is that in our modern digital age, disruptive technological innovations continue to unfold and evolve; here again, without having to look afar, we are intrigued by the potentials of recent trends of Blockchain Technology (BT) or Distributed Ledger Technology (DLT) for disruptive innovations across a wide range of industries. This begs the question of: how enterprises can build sustainable capabilities that will address coming digital technological innovations. Enterprises have adopted varying approaches to address these concerns. In the wake of BT/DLT advances, the way forward is evolving and focusing on achieving a balance on business versus technological values via the "business Use Cases approach" is a vital key to unlock the innovation potentials.

The benefits and contributions of this effort is (a) to stress that new innovations such as BT/DLT require more than mere enterprise information system planning, it is an area where Enterprise Governance of IT (EGIT) and EGIT tools should be applied; (b) to generate deeper awareness in the generation of Blockchain use cases and Blockchain business case as an important phase of BT/DLT innovation management process; and (c) to highlight essential observations on the challenges and issues faced by enterprise in the wake of future digital technological innovations.

This paper is organized into the following sections. Section 2 addresses background preliminaries of related works and concepts; Sect. 3 introduces the formulation of the Blockchain Ledger Description Framework (BLDF); Sect. 4 briefly describes the outcomes of a validation workshop and Sect. 5 concludes the paper with short discussion of issues and challenges faced by enterprises in the wake of digital technological innovations such as BT/DLT.

2 Background and Preliminaries

2.1 Related Works

There has been an explosion of interest in BT/DLT. BT/DLT has gone past its inception/ideation phase and as BT 3.0, BT/DLT has spread into a wide range of industries [3, 4]. It has topped the list of business meeting agendas; it is seeing various forms of trials in different sectors; and without doubt, has undergone or will continue to undergo extensive executive level debates covering issues such as return on investment and the real value to an enterprise.

BT/DLT led transformation in government and business is not expected now [5], however, as seen from the numerous BT/DLT-based initiatives, it is timely to address how this innovation can be embraced by enterprises.

2.2 Essential Concepts

The decision-making and management of knowledge (those data and information resources that underpin the business knowledge) are core activities of modern business entities and enterprises. The reception of new innovations as part of existing enterprise's infrastructure and systems do not differ from existing systems planning and roll-out. Hence, methodological challenges of managing disruptive innovations do not differ much from enterprise information systems planning and implementation. At the design level, there are several similarities as in:

Compatibility and Interoperability: There are concerns of interoperability of new and old systems and the compatibility of new innovations with and within existing infrastructures. Are these systems compatible? Will new innovations work alongside with the existing systems?

Ripple Effects: There are concerns of ripple effects of adding new systems, just as there are concerns of ripple effects of adding new innovations. What should be the adjustments for existing systems when new innovations are added?

Suitability and Appropriateness. There are concerns on the suitable and appropriate properties for the new innovative systems.

While similarities exist, there are at least two core differences. Firstly, BT/DLT requires additional component of enterprise governance of IT (EGIT). The governance aspect for an enterprise system, although can be viewed as a recent perspective, is likely to be applicable to future enterprise-wide systems and services. EGIT calls into play the use of important management tools and frameworks such as COBIT 5 and IT balanced Scorecard (IT BSC). Within the context of EGIT management tools, portfolio management are useful for management of IT-enabled investments such as BT/DLT. As such, as part of portfolio management processes, careful generation of use cases and the development of life-cycle of business case are essential activities.

Secondly, unlike inclusion of new enterprise systems, BT/DLT as innovative technologies require greater proof-of-concept, trials and testing. This would likely to require the setting up of separate and independent testing environment that does not impact the business flow. Our focus here is on addressing the first issue of generation of Blockchain use cases for preparation of suitable Blockchain business case using the Blockchain Ledger Description Framework (BLDF). Various related concepts are addressed before introducing the BLDF. These related concepts set the stage for a better understanding of the design and formulation of the BLDF. For a detailed listing of BT/DLT, see [6, 7].

Enterprise Modelling (EM) and Enterprise Architecture (EA) Design

The EM and EA are complex areas and discipline of studies intended to ensure smooth management and operations at the enterprise level. Proponents of these disciplines develops concepts, theories and tools to ensure that business enterprises are adequately designed, planned and managed. Within these disciplines are tools intended for organizing the enterprise resource needs and to allow for rapid visualization and appropriate decision-making. Some of the core tools that are common in EM and EA and they

include large pool of modelling tools. Use case modelling are subsets of the many tools and methods used. See [8] for a more comprehensive listing of modelling tools.

Enterprise Governance of IT (EGIT) and Portfolio Management (PM)

Enterprise Governance of Information Technology (EGIT) looks at not the IT governance of IT resources, but in addition, it goes on to address business and IT alignment. It focusses on the value that IT-enabled investment can help bring to an enterprise given the risk, culture and business contexts [9].

Use Cases and Description Frameworks

Use case approach has been extensively used in software engineering, enterprise modelling and enterprise architecture design. A simple view and concept of use case can be gleam from one of its many definitions:

...The specification of sequences of actions, including variant sequences and error sequences, that a system, subsystem, or class can perform by interacting with outside actors... [10]

What is clear is that use cases serve many purposes, ranging from specification requirement elicitation to scenario planning and seen from the perspective of use case description, it align well with the description frameworks used in system planning.

3 Description Framework and Blockchain Ledger Description Framework (BLDF) Components

The concept of using description frames for information system planning is not new. In fact, description, layers and components are various means used to organize complex information and to facilitate enterprise modelling and enterprise architecture design. An important usage of description frames is the application of description frames for different views of the same product suggested by [11] as part of information system architecture. The grand concept of Zachman Framework for Enterprise Architecture (ZFEA) [12] and enterprise modelling tools embed this concept and is an essential component.

We adopted and marry the existing description frame from the EM and EA disciplines and apply it to the context of portfolio management (PM) of EGIT. Specifically, we adjust the use case generation phase of PM by developing a series of use case descriptions that present different views of BT/DLT. These views form the framework components of BLDF. Here, we outline an overview of each of the components.

3.1 Overview of BLDF

One of the primary area of concern in dealing with BT/DLT is that of achieving business value: How to ensure that new innovation delivers the proposed value? Another concern is that the technological solution can be morphed or integrated into existing operational framework and structures. The BLDF design is aimed at addressing these concerns. The BLDF applies the component approach comprising of business centered use cases scenarios and infrastructure modules (see Fig. 1).



Fig. 1. Overview of BLDF modules

3.2 Use Case Scenarios

Use case scenario component spells out the full range of scenario that BT/DLT is expected to operate in, for example, to what extend will BT/DLT be used in an enterprise existing business model [5]. This component essentially provides the scope and constraints that will allow the Infrastructures Module component to be formulated.

Example 1. Sample BT/DLT Use Case Scenario—Blockchain in Asset Management

Asset management offers an example of focused scenario. This area covers a wide range of industries, ranging from engineering assets, financial assets to art assets. In essence, a tangible products of value. The value of a product can be further organized into business/strategic value, economic value and/or social value. A typical use case in asset management workflow involves the actor—An Asset Manager with typical assess to data that will allow the deletion of asset; creation of asset, schedule an asset valuation, delete asset policies, create asset policies and query asset databases and policies.

Example 2. Sample BT/DLT Use Case Scenario in Healthcare

Healthcare provides another example of multiple high-level use case scenarios. Unlike Asset management, where the use case is more specialized and focused. The use case scenarios for a complex field like Healthcare may be of a higher and broader level, such as use case for: reducing waiting time in emergency rooms; ensuring availability and accessibility of hardware; tracking inventories and/or addressing drug management and chronic disease management.

Enterprise will explore the use of BT/DLT for use in current business model. Transactions process that requires audited, third party intermediaries will be mapped to proposed BT/DLT workflow that may include public/private DLT.

Base Use Case

This component describes the general flow of interactions and process amongst the various actors. It is at this level or layer that we specific the common interactions to be expected from the new BT/DLT. Unlike concrete or essential use cases [13], base use case does not focus on concrete or specific interactions and avoid going into implementation details.

Use Case Narrative

Narratives supply the purpose and scope for the BT/DLT use case formulation. They are designed usually in parallel with the base use case, however, the narratives go beyond the base use case and addresses other views within the context of enterprise

business strategies and enterprise IT strategies. Either descriptive styles may be used, for example, continuous narrative or numbered narrative. The descriptive narrative allows for quick assessment of business and IT alignment and avoid the pitfall of silo design and development.

3.3 Infrastructure Module

The infrastructure module contains the proposed approach to manage mutable and immutable blockchains and data. One of the strength of BT/DLT is in the ability to ensure the historical track of immutable data and data chain. Business scenarios however inform the need for both flexibility as well as the need to handle changes. These needs will require that data chain be retained as mutable at some point of their life span. To handle these requirements, the infrastructure module contains two key concepts: (a) A user-based validating module and a staging/synchronization module (see Fig. 2).

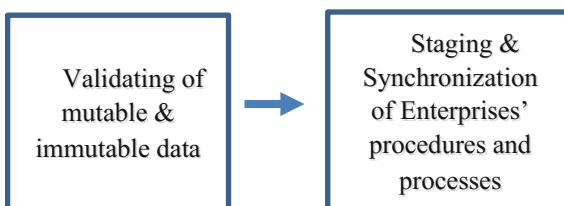


Fig. 2. Components of infrastructure module

4 Validation Workshop

Participatory action method is an establish approach to validate design by group of experts and industry practitioners. The proposed BLDF was presented to a group of industry experts as part of the overall 2018–2019 Blockchain Workshop Series. This participatory approach sees the involvement of users and BT/DLT experts in a round-table discussion. Each member was assigned participatory resources and bid for the order and weights of the framework components. Feedbacks from the workshop suggest the following considerations should be added:

4.1 Mutable and Immutable Data States

Careful consideration of mutable and immutable data and data chains is required especially for specific cases and scenarios. A balance is needed between generic framework versus localization and customization for specific industry and specific enterprises' internal processes and infrastructures. Even within the same industry, different enterprises may still require careful validation and synchronization of operations and processes. A “one-size fit” all approach may not be the ideal solution for BT/DLT.

4.2 Validation of Business Values

The validation of mutable and immutable data blocks module can be further clarified using existing business use cases. Data that needs to be mutable may be removed from inclusion into the BT/DLT. Only those sets of immutable data groups will be added to the BT/DLT. In order to manage the mutable data sets, it is suggested that terminating protocols be added to the mutable data chain and for new data states, new set of data chain will be formed. Each of these data chain is then identified by the complete sequence of base data plus terminating protocols. One of the implication of this suggestion is the need to have an archive for multiple data states; a need to address historical states and to address the time/space dimension of data. One simple way may be not to allow any adjustment of a given time/space arrangement.

4.3 Bi-directional Synchronization

The validation of mutable and immutable data blocks module can be further clarified using existing business use cases. Data that needs to be mutable may be removed from inclusion into the BT/DLT. Only those sets of immutable data groups will be added to the BT/DLT. In order to manage the mutable data sets, it is suggested that terminating protocols be added to the mutable data chain and for new data states, new set of data chain will be formed. Each of these data chain is then identified by the complete sequence of base data plus terminating protocols. One of the implication of this suggestion is the need to have an archive for multiple data states; a need to address historical states and to address the time/space dimension of data. One simple way may be not to allow any adjustment of a given time/space arrangement.

5 Discussions and Concluding Remarks

Some of the old hypes on BT/DLT and disruptive innovations are seen to be paving the way for newer issues and challenges. It was and still is easy to turn to high level concepts of innovative use BT/DLT and hope that they bring the needed innovations and value to the industries and enterprises. As the trend progresses, more management and business case concerns surfaces. As seen in [2] "...cases of well-managed firms such as those cited above, good management was the most powerful reason they failed to stay atop their industries. Precisely because these firms listened to their customers, invested aggressively in new technologies that would provide their customers more and better products of the sort they wanted, and because they carefully studied market trends and systematically allocated investment capital to innovations that promised the best returns, they lost their positions of leadership....".

Here, noticed that it is not the lack of management, but rather good enterprises are failing when it comes to disruptive innovations. The main idea behind preparing for disruptive innovation appears to be continuous and careful adjustment of innovations. This calls for the application of design science by developing descriptive use cases and judicious application of techniques of EGIT such as portfolio management of IT-enabled investments.

While we have addressed a small part of EGIT and IT-enabled investments for digital innovations. We have by no means covered all aspects of EGIT and investments of IT-enabled resources. Future work will address the implementation of BT/DLT project and refine the BLDF components as needed. We look forward to improve upon the outcomes of the Blockchain Workshop Series and to introduce concrete Use Cases within selected domains such as digital asset management; management of information systems (healthcare) and education sectors.

Acknowledgements. UOWD had provided funding and support for the Blockchain Workshop Series, held at Dubai.

References

1. May, T.: Disruptive technology: dead companies do tell tales, 20 Nov 2014. [Online]. Available: <https://www.computerworld.com/article/2849847/disruptive-technology-dead-companies-do-tell-tales.html>
2. Clayton, C.M.: The Innovator's Dilemma: The Revolutionary Book that Will Change the Way You Do Business. Collins Business Essentials, USA (1997)
3. Zamani, E.D., Giaglis, G.M.: With a little help from the miners: distributed ledger technology and market disintermediation. Ind. Manag. Data Syst. **118**(3), 637–652 (2018)
4. Hofmann, E., Strew, U.M., Bosia, N.: Discussion—how does the full potential of blockchain technology in supply chain finance look like? In: Supply Chain Finance and Blockchain Technology. Springer, Cham (2018)
5. Iansiti, M., Lakhani, K.R.: The truth about blockchain. Harv. Bus. Rev. **95**(1), 118–127 (2017)
6. Christidis, K., Devetsikiotis, M.: Blockchains and smart contracts for the internet of things. IEEE Access **4**, 2292–2303 (2016)
7. McConaghay, M., McMullen, G., Parry, G., McConaghay, T., Holtzman, D.: Visibility and digital art: blockchain as an ownership layer on the Internet. Strateg. Change **26**(5), 461–470 (2017)
8. Rittgen, P.: Enterprise Modeling and Computing with UML, Hershey PA 17033. Idea Group Publishing, USA (2007)
9. Van Grembergen, W., De Haes, S.: Enterprise Governance of Information Technology: Achieving Strategic Alignment and Value. Springer Science & Business Media, USA (2009)
10. Rumbaugh, J., Booch, G., Jacobson, I.: The Unified Modeling Language Reference Manual. Addison Wesley, Reading, MA (2017)
11. Zachman, J.A.: A framework for information systems architecture. IBM Syst. J. **26**(3), 276–292 (1987)
12. Lapalme, J., Gerber, A., Van der Merwe, A., Zachman, J., De Vries, M., Hinkelmann, K.: Exploring the future of enterprise architecture: a Zachman perspective. Comput. Ind. **79**, 103–113 (2016)
13. Constantine, L.L., Lockwood, L.A.: Structure and style in use cases for user interface design. In: Object Modeling and User Interface Design, pp. 245–280 (2001)



Human Superposition Allows for Large-Scale Quantum Computing

Bruce Levinson^(✉)

Center for Regulatory Effectiveness, Washington, DC, USA
levinson@thecre.com

Abstract. Quantum computing is not confined to qubits in carefully controlled environments, it also takes place on a human scale. This paper will show that the process by which the Inuit elders make food security predictions on behalf of their people—and the instructions for how to test those predictions—constitute a quantum computing exercise. The elders' food security predictions require calculations about environmental systems and their complex interactions based on (1) a millennium or more of the Inuit people's experiences and (2) a continually evolving understanding of these experiences. When the elders engage in their decision-making process, they achieve superposition—that is to say that the elders, collectively, are able to achieve a unified perspective that encompasses the sum total unique understanding of the world of each elder. Inuit Indigenous Knowledge, which has produced robust and reliable food security predictions for centuries, is consistent with quantum mechanics. Conclusion: Inuit Indigenous Knowledge is a scientific system comparable to—and deserving of acceptance by scientists and governments as being of equal stature to—"Western" science.

Keywords: Environmental · Food security · Inuit people · Indigenous knowledge · Quantum computing

1 Purpose

The purpose of this paper is to demonstrate that Inuit Indigenous Knowledge is, by commonly accepted standards, a scientific system that produces reliable and robust predictions about atmospheric, hydrological, and biological systems and thus is deserving of equal stature to standard model science in Arctic governance decisions.

2 Background

2.1 About Indigenous Knowledge (IK)

IK is a system. IK is “a systematic way of thinking applied to phenomena across biological, physical, cultural and spiritual systems. It includes insights based on evidence acquired through direct and long-term experiences and extensive and multi-generational observations, lessons and skills. It has developed over millennia and is still

developing in a living process, including knowledge acquired today and in the future, and it is passed on from generation to generation.” [1].

IK is integrative. IK focuses on understanding how systems such as air, water, and humanity work in concert. This is in contrast to the usual practice of Western scientists, which seeks to understand the laws of nature on as finely splintered a level as possible.

IK does not use measurement. The Inuit people have developed an understanding of nature that uses experience instead of measurement as its fundament. Each participant in the IK process recognizes that the other participants also assign unique meanings to observations. For example, while modern observers consider the night sky at a given place and point in time to be an objective phenomenon, the Inuit do not.

The archaeoastronomer Clive Ruggles explains: “Traditional knowledge of the skies can be very localised, and even personal, as among the Inuit, for whom knowledge of celestial phenomena is ‘in varying degrees, specific to communities, families... When imparting information, elders frequently made it plain that they were speaking for themselves, that their opinions were not necessarily correct in any absolute sense, and that other elders might, and in probability did, have different views.’” [2].

The Inuit people use no measuring instruments (compasses, sextants, barometers, etc.) and thus neither take measurements nor use any analog of “objective” data in their scientific system.

2.2 IK Produces Useful, Reliable, and Counterfactually Robust Results

Science is observation- and experiment-based study of the laws of nature. Philosophers explain that the laws of nature are distinguishable from other regularities that the universe happens to be conforming with by being “counterfactually robust,” which is to say that these laws would exist under a wide range of alternative conditions and scenarios [3].

The Arctic provides a dynamic laboratory for testing meteorological and oceanographic predictions. Testing of IK by communities across the Arctic coast for at least 800 years demonstrates that the Inuit are able to make useful predictions based on the laws of nature [4].

Anthropogenic climate change and government regulations have provided additional and ongoing counterfactual robustness testing of IK, since the warming Arctic and government restrictions on hunting and other food security decisions have changed regularities that the region has conformed with for centuries. “For example, shifts in animal migration patterns and shifts in vegetation are occurring as a result of changes in temperatures, salinity levels, precipitation rates, snow coverage, soil integrity (erosion), ice coverage, etc. Such changes require adjustments in gathering, hunting and fishing strategies. Additionally, we face new dangers as we attempt to navigate through storms with increased intensity, rotting ice, timing of sea and/or river ice formation and change in ice thickness.” [5].

The survival of the Inuit people attests that IK is a robust and reliable scientific system which is unsurpassed for making useful collective predictions based on the laws of nature; the Arctic does not grant mulligans.

3 Human Superposition

Quantum systems, unlike classical systems, are systems in which information is incomplete and incompletable. Quantum information is incompletable because, as Niels Bohr explained, reality is too rich to be completely captured from any single perspective; there is room for mutually exclusive perspectives that offer partial, valid perceptions.

A quantum state is an observer's personal—that is, Bayesian—probabilistic expectation for the outcome of an experiment on a quantum system [6]. This personal probability is subject to continual updating via ongoing experience. A quantum state, thus, exists only as a quantified expression of an observer/experimenter's personal belief, not as a physical reality. The quantification of belief allows an observer's personal expectations to be analyzed via mathematical formalism. Quantum states are additive, and superposition refers to the sum total of a set of quantum states.

In the case of IK, each Inuit elder possesses a quantum state, that is to say, has probabilistic beliefs about the Arctic and the planet understood as a single system—a system about which they possess imperfect information. These beliefs are based not only on a lifetime of experience but also on the information about the observations and understanding of their families and friends that are communicated in daily discourse and the experiences of their people going back centuries that are stored and communicated in oral histories.

When the Inuit elders gather to make food security decisions for the coming year, they achieve human superposition; that is, they collectively develop and communicate a prediction on how to achieve food security based on the sum total of the unique perspectives and knowledge bases of each.

Quantum theory can provide us with keen insights into the elders' decision-making process. Quantum theory, as understood via QBism, says that observers are not incidental to the world but matter as much as electrons and atoms. The physicist Christopher Fuchs explains that “the greatest lesson quantum theory holds for us is that when two pieces of the world come together, they give birth. [Bring two fists together and then open them to imply an explosion.] They give birth to FACTS in a way not so unlike the romantic notion of parenthood: that a child is more than the sum total of her parents, an entity unto herself with untold potential for reshaping the world. Add a new piece to a puzzle not to its beginning or end or edges, but somewhere deep in its middle and all the extant pieces must be rejigged or recut to make a new, but different, whole. That is the great lesson.” [7].

By recognizing that reality is participatory, not passive, we can infer that the Inuit elders are not merely making predictions, they are also providing their people with a road map of the actions that will help create the future that is predicted.

4 Integrating the Observer into the Observation

In Western, i.e., observer-abstracted science, there is a “tradition of removing the observer from the description in order to guarantee objectivity.” The problem with this approach is that observers attach meaning to data and thus cannot be truly objective.

The computational scientist Russell K. Standish explains that by “explicitly recognising a role for the observer of a system, an observer that attaches meaning to data about the system, contradictions that have plagued the concept of complexity can be resolved.” Standish further explains that “Explicitly acknowledging the role of the observer helps untangle other confused subject areas. … Quantum Mechanics can also be understood as a theory of observation. The success in explaining quantum mechanics, leads one to conjecture that all of physics may be reducible to properties of the observer.” [8].

Explicitly recognizing the role of the observer also opens the door to our understanding IK. In IK, observers are recognized as integral to their personal observations. IK uses disparate perceptions among the elders as the basis for complex food security decisions that incorporate predictions about numerous phenomena, including weather and the migration patterns of various species.

The collective ability of the Inuit elders to make useful predictions for their communities based on the personal observations and beliefs of each elder demonstrates that integrating observers into observations does not result in solipsism. Instead, the mutual recognition of non-definitive perspectives is a basic condition that allows the elders to store and make use of millennia of Arctic observations [9].

Consistent with IK’s observer-integrated approach, these predictions—as well as the hunting and other actions that test the predictions—are made without sextants, barometers, or other devices used to capture and communicate objective data. The solid state physicist N. David Mermin explains that the “term ‘measurement’ plays no fundamental role in” quantum mechanics. Rather, the measurements “that play so central a role in the orthodox theory are just particular examples of actions taken by a user of science, usually with the help of a large piece of apparatus.”

5 Language Reflects IK’s Integrative Approach

The complexity of observer-integrated science explains the ability of the Inuit people to use an oral tradition to preserve, communicate, analyze, and make predictions based on environmental information that has been developed across centuries. Integrating observers into observations requires that the language used be able to accommodate great complexity. IK’s integrative approach to understanding the world is reflected linguistically in the Yup’ik language, which can combine multiple types of information in a single word. For example, there is a Yup’ik name for a certain fish, Imangaq, which is translated into English as “black fish.” The Yup’ik name, however, encodes more information than the English and Latin names for the fish.

English ordinary proper names, names that are potentially generic, such as “black fish,” can and do routinely refer to a singular thought, for example, a specific type of fish. The ordinary proper name identifies the fish; however, it provides no additional information. The Imangaq’s Latin name, its binomial nomenclature, encodes more information because it identifies the specific type of fish and also links it to related species.

The Yup'ik name Imangaq is more complex than binomial nomenclature in the connections that it references, because it connects the fish to its environment—including humans.¹ Imangaq designates a specific type of fish and also refers to “the education youth gain when taught how to obtain this fish ...”

Inuit Indigenous Knowledge embodies a unitary concept of the world as a single system. Although this may, mistakenly, be viewed as an overly simplified view, the physicist Erwin Schrödinger pointed out that when multiple systems engage with each other, “the best possible knowledge of a *whole* does not necessarily include the best possible knowledge of all its *parts*, even though they may be entirely separate and therefore virtually capable of being ‘best possibly known,’ i.e., of possessing, each of them, a representative of its own.”

It is because the Inuit people understand the environment—including their role in it—on a non-divisible basis that they are able to store and process observations without measurement, since the very concept of measurement is based on understanding the laws of nature in a fractionated format.

Quantum computers seek to take advantage of the superposition of quantum bits—qubits—to solve calculations far faster than transistor-based computers could be capable since transistors are capable of existing in only a single state at a time. However, since qubits and, thus, quantum computers operate only under rigorously controlled laboratory conditions, quantum computers remain in rudimentary states of development.

With IK, the superposed elders bring to the prediction-making table are beliefs that take into account a sum of information that would take storage technologies and information processing power far beyond current technologies in order to evaluate. The Inuit elders make predictions about the Arctic environment—and about how humans should behave as part of the system in order to maximize the likelihood of our survival—that far surpass in detail and accuracy the outputs from even the most sophisticated computer-aided environmental modeling of the Arctic.

6 Conclusion

IK is a scientific system that sets the gold standard for making predictions about the effects of human activities in the Arctic. Accordingly, IK should be given at least equal status with other scientific systems in Arctic governance decisions.

References

1. Daniel, R., Behe, C.: Co-production of knowledge: an Inuit Indigenous Knowledge perspective. American Geophysical Union, Fall Meeting 2017, abstract #C13H-04 (2017)
2. Ruggles, C.: Indigenous astronomies and progress in modern astronomy. In: Norris, R., Ruggles, C. (eds.) Accelerating the Rate of Astronomical Discovery (Special Session 5, IAU General Assembly, Aug 11–14, 2009, Rio de Janeiro, Brazil). Proceedings of Science, PoS (sps5) (2009)

¹ In accordance with conventional wisdom, the Inuit have many words for snow, but fewer than 50.

3. Roberts, J.: Laws, counterfactuals, fine-tuning and measurement. Collected in *Laws of Nature: Their Nature and Probability*, Perimeter Institute (2010)
4. Raff, J., et al.: Mitochondrial diversity of Iñupiat people from the Alaskan North Slope provides evidence for the origins of the Paleo- and Neo-Eskimo peoples. *Am. J. Phys. Anthropol.* **157**(4), 603–614 (2015)
5. Inuit Circumpolar Council-Alaska, “Alaskan Inuit Food Security Conceptual Framework: How to Assess the Arctic From an Inuit Perspective. Technical Report.” Anchorage, AK, (2015)
6. Caves, C.M., Fuchs, C.A., Schack, R.: Quantum probabilities as Bayesian probabilities. *Phys. Rev. A* **65**, 022305 (2002)
7. Fuchs, C.A.: “On Participatory Realism” v3, [arXiv:1601.04360v3](https://arxiv.org/abs/1601.04360v3) [quant-ph] to appear in “Information & Interaction: Eddington, Wheeler, and the Limits of Knowledge”, edited by Durham, I.T., Rickles, D. (2016)
8. Standish, R.K.: The importance of the observer in science. In: *Proceedings the Two Cultures: Re-considering the Division Between the Sciences and Humanities*, UNSW (2005)
9. Riedlinger, D., Berkes, F.: Contributions of traditional knowledge to understanding climate change in the Canadian Arctic. *Polar Record* **37**, 315–328 (2001)



Student User Experience with the IBM QISKit Quantum Computing Interface

Stephan Barabasi, James Barrera, Prashant Bhalani, Preeti Dalvi, Ryan Dimiecik, Avery Leider^(✉), John Mondrosch, Karl Peterson, Nimish Sawant, and Charles C. Tappert

Seidenberg School of Computer Science and Information Systems, Pace University,
Pleasantville, NY 10570, USA

barabasi@us.ibm.com,
{jb05881n,pb21845n,pd21567n,rk43626p,aleider,jm58593n,
kp53279n,ns33992n,ctappert}@pace.edu

Abstract. The field of quantum computing is rapidly expanding. As manufacturers and researchers grapple with the limitations of classical silicon central processing units (CPUs), quantum computing sheds these limitations and promises a boom in computational power and efficiency. The quantum age will require many skilled engineers, mathematicians, physicists, developers, and technicians with an understanding of quantum principles and theory. There is currently a shortage of professionals with a deep knowledge of computing and physics able to meet the demands of companies developing and researching quantum technology. This study provides a brief history of quantum computing, an in-depth review of recent literature and technologies, an overview of IBMs QISKit for implementing quantum computing programs, and two successful programming examples. These two programs along with the associated Jupyter notebook pages will provide additional intermediate samples for IBM Q Experience.

Keywords: Quantum computer · Qubit · QISKit · Tutorial · Student

1 Introduction

In 1959, shortly after the arrival of the classical computer and the beginning of the digital era, theoretical physicist Richard Feynman gave a lecture series on electronic miniaturization titled “There’s Plenty of Room at the Bottom.” Feynman proposed in these lectures that the exploitation of quantum effects could create more powerful computers. Quantum computers, machines that operate on qubits rather than traditional bits, are the manifestation of Feynman’s proposal [1].

Thanks to the IBM Faculty Award that made this research possible.

The discovery of quantum mechanics in the early 20th century laid the foundations for Feynman's proposal. In 1905, Albert Einstein published a paper proposing the photon concept of light [2]. In the paper, he described light as quantum particles that are “localized points in space, which move without dividing, and which can only be produced and absorbed as complete units [2].” This challenged the accepted model at the time, which viewed light only as a wave. Ultimately, Einstein's revelation on the physical properties of light laid the groundwork for the development of quantum physics.

Following Einstein's quantum explanation of light, the field of quantum mechanics developed gradually with research by Werner Heisenberg, Niels Bohr, and Erwin Schrödinger [3]. Quantum mechanics states that the position of a particle cannot be predicted with precision as it can be in Newtonian mechanics. The only thing an observer could know about the position of a particle is the probability that it will be at a certain position at a given time [3]. By the 1980s several researchers including Feynman, Yuri Manin, and Paul Benioff had begun researching computers that operate using this concept.

The quantum bit or qubit is the basic unit of quantum information. Classical computers operate on bits using complex configurations of simple logic gates. A bit of information has two states, on or off, and is represented with a 0 or a 1. The qubit has a $|0\rangle$ state, called zero-ket and a $|1\rangle$ state called one-ket. When zero-ket is measured it produces a classical 0. When one-ket is measured it produces a classical 1. The most notable difference between classical and quantum bits is that the qubit can also be in a state that is a linear combination of both $|0\rangle$ and $|1\rangle$. This vector property of the qubit is called superposition. Superposition allows for many calculations to be performed simultaneously [4].

Another distinct property is that qubits can become entangled. Entanglement is a property of many quantum superpositions and does not have a classical analog [4]. Observation of two entangled qubits causes random behavior in the observed qubit, however, the observer can tell how the other would behave if observed in the same manner [4]. Random behavior between entangled qubits is responsible for the extra computing power of quantum machines.

Quantum computers are useful for large scale problems that classical computers lack the computational power to solve in a reasonable amount of time. Superposition creates opportunities to implement more efficient factoring, searching, sorting, modeling, and simulation algorithms. According to Almudever , factoring a 2000 bit number using Shor's algorithm on a quantum computer would take one day. Factoring the same number using a classical computer would take a data center the size of Germany one hundred years to complete [5].

Quantum computing promises breakthroughs in fields that deal with large data sets and require massive amounts of processing power such as genetics, molecular modeling, astrophysics, chemistry, data science, artificial intelligence, and cryptography. This expanding field will need many trained professionals proficient in both physics and computing. This study will document the student experience using the IBM Q Experience and QISKit to pursue training in programming for quantum computers.

2 Study Overview

This study explores the educational material available online for students interested in programming for quantum computers focusing on the interface offered by IBM that allows users to run code on publicly available quantum computers or quantum simulators.

The study uses IBM Q Experience to create two simple quantum programs and documents the experience working with the development kits, programming languages, user interfaces, online documentation, and tutorial information. The study utilizes IBMs QISKit, a Python based development kit, to create a simple random password generation program and a card game program that runs on IBMs publicly available quantum computer in Yorktown, New York.

The goal of this study is to produce an informational review of a students early experience with quantum computing focusing on improving the currently available material in order to attract more students to the growing field. Toward this goal, the study includes a supplemental quantum tutorial using Jupyter Notebook that incorporates the example programs and serves as a guide for students interested in quantum computing.

3 Literature Review

In 1965, Gordon Moore, the co-founder of Fairchild Semiconductor and Intel proposed that transistors (the main component of central processors) would keep shrinking every year due to engineering advancements in the semiconductor industry. Specifically, Moore claimed that the industry would double the number of transistors in CPUs approximately every year. He would later revise this to every 2 years in 1975, while David House, an Intel executive at the time, noted that the computing power would also double every 18 months [6]. In the late 1950s, some chips contained 200 transistors; and by 2005, “Intel would produce chips with 1 billion transistors [6].” The semiconductor revolution gave rise to the internet, smart phones, and the Internet of Things.

Consumer demand for more powerful devices has increased, and manufacturers strive to push the limits of Moore’s law. However, simple physics dictates that transistors can only be so small. “No physical quantity can continue to change exponentially forever,” Moore said in 2003 [6]. In March of 2016, the semiconductor industry would formally acknowledge the nearing end of Moore’s law [7]. This was due in part to the excess heat generated by densely packed silicon circuitry in small devices, and that when transistors shrink to only 10 atoms across, “electron behavior will be governed by quantum uncertainties that will make transistors hopelessly unreliable [7].” Researchers hope that quantum computing will allow a new phase of exponential growth in computing speeds. IBM announced in a press release in May of 2016, “with Moore’s Law running out of steam, quantum computing will be among the technologies that could usher in a new era of innovation across industries [8].”

In his influential paper, “The Physical Implementation of Quantum Computation” David DiVincenzo proposed five criteria that quantum computer implementation must meet. First, the system must utilize qubits for making computations. Second, the system must have the ability to initialize the qubits to a known state. Third, the computer must use a universal set of quantum gates. Fourth, the gates must operate faster than the information stored in the qubits can be lost due to interaction with the surrounding environment. And finally, measurement of information contained in qubits must be possible [9]. A machine that meets these criteria is classified as a quantum computer.

The first quantum computer was built at Oxford University in 2004. The machine was a nuclear magnetic resonance (NMR) machine that was able to solve simple calculations twice as fast as a classical computer. NMR machines utilize magnetic field emissions created by electrons orbiting a nucleus and radio waves of varying frequencies to manipulate the electron fields causing the electrons to act as qubits [3].

In the last five years, large corporations such as Google and IBM, and start ups such as Rigetti and Quantum circuits have had major success implementing superconductor quantum computers. Superconductor computers require temperatures of millikelvin to operate. Qubits can store information through the charge of the particles. Superconductor machines use pulses of radio frequencies to control qubits in order to execute quantum operations [3].

Researchers are also working on other implementations of quantum computers. Linear optic machines have potential for future breakthroughs due to low infrastructure cost. Diamond machines use impurities in diamonds to capture single electrons and use them as qubits. Diamond machines have the advantage over superconductors in that they can operate at 4 K. The potential benefits of Diamond implementations are promising, however the diamond is not as well researched as other technologies.

The cutting edge of quantum computer implementations is the anyon computer. Researchers theorize that “under some circumstances certain material can behave as if they are holding particles that do not actually exist, known as quasi-particles [10].” These quasi-particles, or anyons, can be used as qubits. The physical properties of anyons protect from decoherence and give anyon machines an advantage over other implementations. Anyon research is promising although still in the theoretical phase.

There are many engineering challenges in quantum computing. As mentioned above, decoherence describes a process in which information stored in qubits is lost due to its interaction with the environment. All implementations of quantum computers discussed above suffer from decoherence. Decoherence must be accounted for using quantum error correction (QEC). QEC is necessary for accurate calculations but drastically increases the necessary number of qubits required for computations [5].

Quantum computers must also be capable of fault tolerant (FT) computations. Fault tolerance means that small errors do not lead to information loss or larger errors in calculations. C. G. Almudever et al. suggest that a number of

qubits in the millions may be required to perform QEC and FT computations successfully [5].

Another challenge is that quantum computers are expensive, large, and difficult to maintain. For most organizations, it is impractical to implement a machine with even a small number of qubits. Cooling the quantum machine accounts for much of the cost and difficulty of operation. Similar to classical computers, quantum computers require cooling, but at much lower temperatures [11]. When a qubit changes its state it generates heat, and because qubits change states often during program execution they can generate massive amounts of heat. Cooling quantum superconducting processors requires extremely low temperatures. For example, at D-Wave Systems in British Columbia Canada, the cryogenic system processor operates at 15 mK, “approximately 180 times colder than interstellar space.” In other words, the quantum processor requires -460°F or -273°C to operate [12].

Due to massive operational costs and engineering challenges, R. Van Meter and S. J. Devit have proposed two new criteria to be added to DiVincenzo’s original quantum criteria. First, that quantum systems be small, cheap, and reliable enough to be practical and fast enough to make using one worthwhile. Second, that “implementation limitations make locally distributed computation imperative, which requires system area networks that are fast, high-fidelity, and scalable [10]”. These two criteria have led to research into distributed quantum systems. If one organization is able to maintain a 20 qubit machine, another organization a 15 qubit machine, and a third organization a 12 qubit machine, the ability to use all three machines simultaneously would allow the quantum computation to access 47 qubits. Researchers are currently working on connecting and scaling these machines.

The complexities of creating distributed quantum systems show the necessity for attracting students and professionals from multiple disciplines to the field. Researchers skilled in multiple disciplines of physics are needed to develop quantum chips. Electrical engineers and computer architects must develop quantum instruction set architectures and error correction schemes. Computer scientists must develop quantum compilers and higher level programming languages capable of running on quantum hardware. Chemists and materials engineers must develop hardware capable of maintaining the quantum states of qubits and extremely low temperatures. Quantum computing sits at the crossroads of these disciplines. Attracting skilled and knowledgeable individuals to the field of quantum computing is vital to its future success.

4 Methodology

This study utilizes IBMs QISKit, a Python based software development kit to implement two quantum computing examples. The QISKit requires Python 3.5 or later and runs on Windows, OSX, and Linux. Users who register with IBM can gain access to the Q Experience API which allows users to run code on a publicly accessible quantum computer in Yorktown, NY. Currently, IBM allows

access to three quantum chips. IBMQX2 and IBMQX4 are both 5 qubit chips. IBMQX5 is a 16 qubit chip [13]. The QISKit also allows users to run quantum simulations on a local machine.

The basic structure of a quantum program is organized into three steps: build, compile, and run. In the build step, the user creates a quantum circuit composed of quantum registers. The user can then add quantum gates to manipulate the registers. Gates are essential to quantum programs. They perform operations directly on qubits. In the compile step, the user selects a back-end, either their local machine or a publicly available IBM quantum chip, on which to execute their quantum code. In the final step, the user runs the code and receives a result. The user can select various options that can affect execution in the step.

All quantum programs operate on qubits using gates. A single qubit is represented visually using a Bloch Sphere.

Figure 1 on page 553 is an illustration of a Bloch Sphere. The Bloch Sphere has a radius of 1 and a vector from its center to its surface. The state of the qubit is represented by the point where the vector meets the surface of the sphere. The point at the top of the sphere represents $|0\rangle$ and the point at the bottom of the sphere represents $|1\rangle$. Gates manipulate the position of the qubit on the surface of the sphere to perform calculations [4].

This Bloch Sphere is a representation of a qubit with X, Y, and Z axis labeled. $|0\rangle$ at the top of the sphere. $|1\rangle$ at the bottom of the sphere. The position of the qubit on the surface of the sphere is represented by the solid orange vector. Angle θ represents a state of superposition when the value of the qubit is between $|0\rangle$ and $|1\rangle$. Angle ϕ represents rotation around the Z axis or phase of the qubit.

The quantum gate is the primary tool a developer can use to interact with qubits. Therefore, in order to write a quantum program, a developer must have a thorough understanding of how to use quantum gates. The following sections will explain in detail the two quantum computing program examples produced by this study including how they make use of quantum gates [14].

5 Program I - Quantum War

The first programming example is the quantum version of the children’s card game “War”. This program demonstrates a practical application for the quantum principle of superposition. Figure 2 on page 554 shows how a small number of qubits can do the work of many classical bits by using a real-world example to increase comprehension. It also illustrates how to create and measure two quantum circuits.

To play the game, a deck of playing cards is divided evenly among two players. Each player simultaneously reveals a card, the player with the card of the highest order wins the battle. The winning player claims the losing player’s card. The first player to run out of cards loses. After drawing, the cards are removed from the deck. Traditionally, the rules of War state that the losing player’s cards are added to the winning player’s deck. This study has opted to forego this portion of the rules to keep the example simple for those first learning about quantum

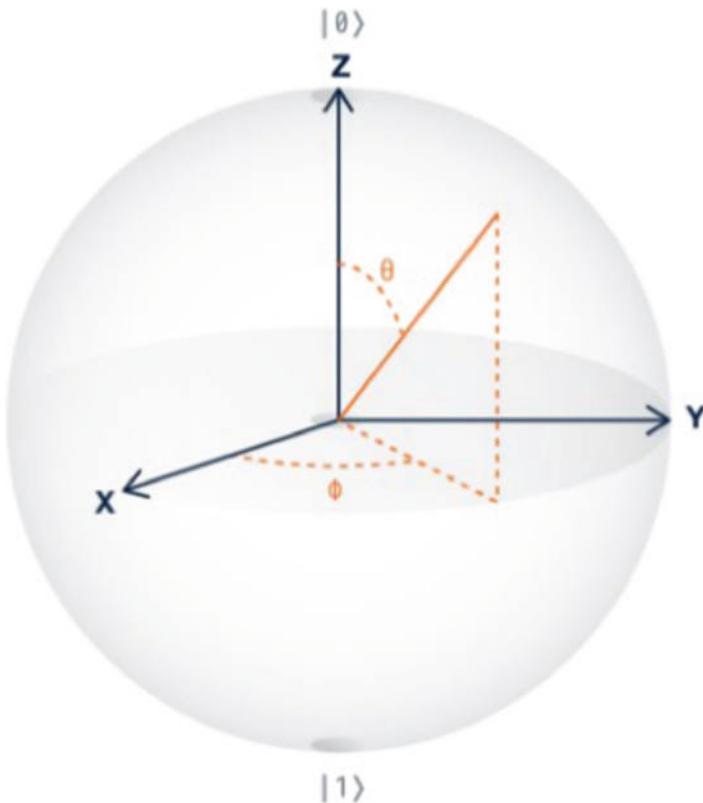


Fig. 1. Bloch sphere [4]

programming. This study has also opted to omit the three card W-A-R draw at the start of each round for simplicity.

A traditional deck of cards contains fifty-two unique cards. In order for a classical computer to represent one card of the fifty-two options, six bits are required as illustrated in Fig. 2. In comparison, a classical computer requires three hundred and twelve bits (six bits per card) to represent all fifty-two options simultaneously. The principle which allows the quantum computer to represent all fifty-two options with only these six qubits is superposition. Superposition, illustrated in Fig. 3 on page 554, is when the value of a qubit is a linear combination of both $|0\rangle$ and $|1\rangle$. This means that the qubit represents all possible values simultaneously. Thus, six qubits, all in a state of superposition, represent all fifty-two possibilities when drawing from a deck of playing cards.

Another means of demonstrating the position of the qubits as well as the gates influencing them is by creating a simple quantum circuit schematic as illustrated in Fig. 4 on page 555. For this program, each deck is configured in the same fashion and can be represented with a single schematic. The schematic allows

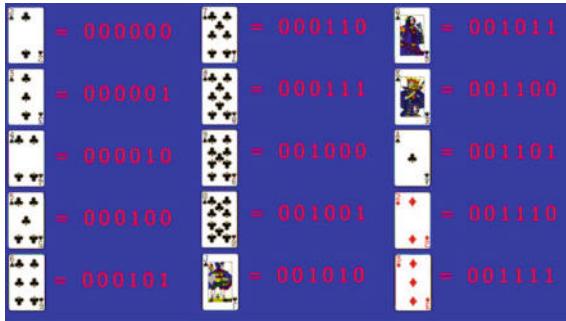


Fig. 2. Visual representation of how cards are identified in the quantum deck

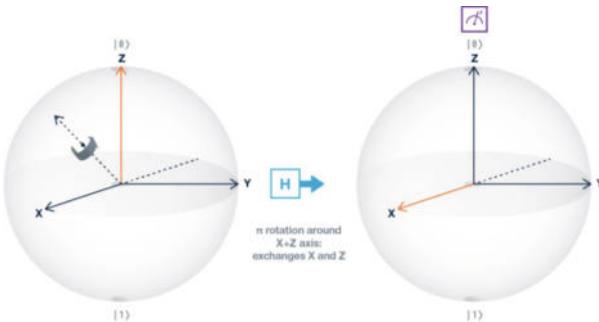


Fig. 3. Bloch sphere representing H gate rotation around the X + Z axis. The qubit, represented by the orange arrow, begins with a state of $|0\rangle$. After the H gate performs the rotation, the qubit is in a state of superposition. The value of the qubit is a linear combination of $|0\rangle$ and $|1\rangle$ [4]

the observer to see the qubits, how they are manipulated, and the projected outcome based on the gates applied to each qubit.

This study slightly modifies the rules of War to illustrate more fully the quantum principles at play. Each player plays with a full deck of fifty-two cards. Each deck is represented by six qubits, twelve qubits in total.

The program begins with two classes that are used to keep track of the cards in each player's deck. The Card class holds information about each playing card including, the suit, name, and value of the card. The Deck class contains logic for creating a deck of cards containing 52 cards with values 2–10, Jack, Queen, King and Ace with suits Clubs, Diamonds, Hearts, and Spades. It also contains the logic for drawing a card and removing it from the deck.

Following the class definitions, the user chooses to run the program on a local quantum simulator or IBM's IBMQX5, a 16 qubit chip. The simulator runs faster than making a connection to the IBMQX5 and is guaranteed to produce a result. Due to high volume of usage, a request to use the IBMQX5 can sometimes timeout.

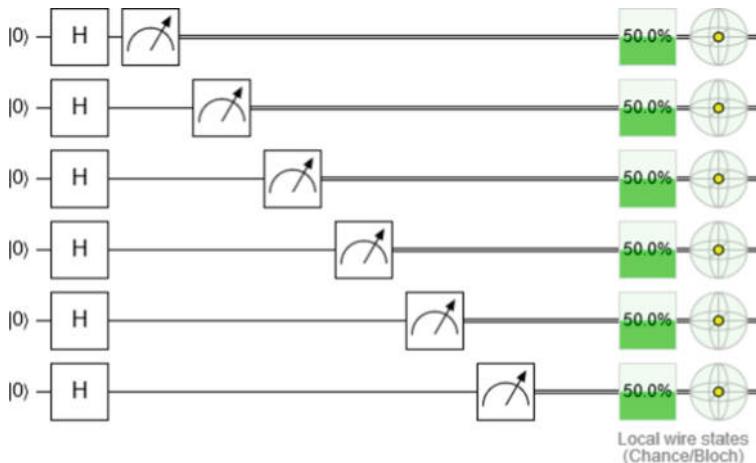


Fig. 4. Visual representation of quantum circuit schematic. Each horizontal line represents one qubit. Along the line gates manipulate the value of the qubit. After the gate manipulate the value of the qubit, the program measures the value of the qubit, represented by the arrow shaped measurement indicator. On the right side, the percentage represents the probability that the qubits value is $|1\rangle$. On the far right, a Bloch sphere represents the qubit with the gate(s) applied

Next, the program creates two quantum circuits to represent each deck. Each circuit contains six qubits. All six qubits have an H gate applied to them to induce a state of superposition represented in Fig. 3 on page 554. This will allow all six qubits to potentially represent any of the 52 cards in the deck. The program now contains two quantum circuits, each of which represents a deck of 52 playing cards.

Next, the battle between players begins. Player One's quantum circuit is measured. This represents drawing a card from the top of the deck. Then, player Two's quantum circuit is measured. The measurement returns the probability of drawing any one of the fifty-two cards from the top of the deck. The card drawn is determined by the measurement taken of the state of the qubits.

The six qubits representing each deck are measured 1024 times. One measurement is referred to as a “shot.” In a single shot, the probability that a qubit is equal to $|0\rangle$ or $|1\rangle$ is measured and recorded. After all 1024 shots, a dictionary is returned that contains the number of times all sixty-four combinations of the six qubits were measured. The value between 000000 and 110011 that is measured with the highest frequency corresponds to the card that the player drew. The drawn card is then removed from the deck and cannot be drawn again. Finally, Python logic compares the two cards and determines a winner for the battle.

The program then presents the players with the cards that they drew and the likelihood that the qubits represented that card at the time of measurement. The likelihood is calculated by dividing the number of times the value was measured

by the total number of shots and represented as a percentage in the output. Sample output is presented in Fig. 5 on page 556.

The program is successfully able to simulate a truly random event. In classical computing, random functions are usually pseudo random. The functions must be seeded with the date and time or other information to produce a result that seems random but is not actually random. In contrast, the behavior of the qubits is truly unpredictable. Thus, when measuring qubits to represent a real world activity, such as drawing a card from a shuffled deck, the result is truly random. The card game simulation employs qubits in a practical situation that requires truly random results.

```
----- Battle 4 -----
Player 1 drew: 3 of Clubs      2.73%
Player 2 drew: 6 of Spades     2.34%
Player 2 wins the battle!
Player 1 Score: 1
Player 2 Score: 2
```

Fig. 5. Sample output of the quantum war program. Player 1 drew the 3 of clubs. This draw was measured in 2.73% of shots. Player 2 drew the 6 of Spades which was measured in 2.34% of the shots.

6 Program II - Random Password Generator

The second programming example is the quantum version of a random password generator. This program demonstrates how to use superposition to generate ASCII characters. It also shows how a small number of qubits can do the work of many classical bits and uses a practical example to increase user engagement. It also shows how to use a single measurement to get multiple results and illustrates how to account for noise when measuring a quantum circuit.

This program uses a quantum circuit to create a truly random password using English ASCII characters. The program uses an 8 qubit quantum circuit to generate a variable length random password. A single ASCII character, represented on a classical computer, uses 8 bits. In the classical implementation, 8 bits represents a single character, and only that character. In order for the classical machine to represent a different character, the value of one or more bits must be changed or more bits must be used. The quantum implementation can utilize superposition to create multiple characters using only 8 qubits by measuring each qubit multiple times. Thus, the quantum circuit represents all ASCII characters simultaneously when in a state of superposition.

The quantum computer has two advantages over the classical computer when it comes to password generation. The first advantage is that because quantum computers use qubits with superposition, they can generate longer passwords in less time than a classical computer. The second advantage is that the quantum computer is capable of generating a truly random password while the classical computer is not. While the classical computer utilizes pseudo random functions to produce results that seem random, only qubits can produce truly random results. As with the Quantum War program, the motion of quantum particles is unpredictable. Thus, using qubits can produce a password that is truly random and thus stronger than a random password generated on a classical computer.

The program begins by creating an 8 qubit quantum circuit. Next, the program uses gates to set the qubits to the proper states for execution. All English ASCII characters begin with the leading bits 01. The following six bits are variable and can be set to either 0 or 1 depending on the character being represented. Thus, to represent an ASCII character with qubits, the leading two qubits, qubits 7 and 6, must be set to 0 and 1 respectively. These values must not change during the execution of the program or non-English characters will appear in the results. The program achieves this by leaving qubit 7 in its initialized state of 0 and using an X gate to set qubit 6 to 1. Next, qubits 0 through 5 are put into a state of superposition using H gates.

This is demonstrated in the quantum circuit schematic in Fig. 4 on page 555. This program is a simple yet efficient representation of the randomness of superposition and demonstrates the advantages over classical computing. When all eight qubits are set to the correct values, the program measures the quantum circuit. This determines the characters that are used to generate the password. By measuring the qubits, the program can determine the probability that they represent a particular character at the time of measurement. The eight qubits representing a single character are measured in 1024 shots. Each shot is a single measurement that represents a single character. After all 1024 shots occur, QISKit returns a dictionary that contains the frequency that each character was measured. Figure 6 on page 558 illustrates a dictionary representation of the measurement results.

Because Quantum measurements can be noisy, the program must also account for noise in the measurement. Noise is unintended environmental interference that may distort the results of a measurement. To account for this, the program eliminates all results that were measured in less than 2% of shots. The 2% percent noise threshold is recommended in several QISKit examples. The program assumes that any character which appears in greater than 2% of shots is a legitimate qubit value and not noise. In Fig. 6 on page 558, the results of 1024 shots are displayed. The character 01101110, the character n, was measured in 13 shots. To determine if the character n was actually measured or is the result of noise, the program divides 13 by 1024 which equals 0.0127. Since the character n only appeared in 1.2% of shots, it is indistinguishable from noise and thus is not included in the password. This noise threshold check is performed for each

[89]	'01101110'	= {int} 13
[89]	'01111000'	= {int} 20
[89]	'01001100'	= {int} 16
[89]	'01101011'	= {int} 11
[89]	'01001101'	= {int} 15
[89]	'01110010'	= {int} 20
[89]	'01010000'	= {int} 15
[89]	'01010100'	= {int} 13

Fig. 6. An excerpt from the dictionary representing the frequency that each character was measured during each of the 1024 shots. On the left, is the binary representation of the character and on the right the number of shots which the 8 qubits represented that character. The first character in the dictionary 01101110 is the character n and it was measured in 13 of the 1024 shots

measured character. All characters that the program measures in more than 2% of shots are added to the password.

Because the program accounts for noise, the final length of the password is variable. Generally the password is between six to eight characters but may be longer or shorter. This is due to the random nature of qubits and the fact that noise can distort measurements. Sometimes the measurement frequency of a few characters rises above the noise threshold. Other times, ten or more characters may be measured over 2% of the time. Figure 7 on page 559 demonstrates three sample passwords generated by the quantum program. Each column represents a single password. On the left is the character, and on the right the percentage of shots that it was measured in. Running the program multiple times will predictably produce varied results.

Password 1		Password 2		Password 3	
H	2.05%	~	2.34%	g	2.44%
P	2.15%	c	2.64%	X	2.05%
S	2.25%	K	2.34%	R	2.15%
N	2.05%	N	2.05%	Q	2.15%
W	2.05%	d	2.25%	Z	2.25%
^	2.05%	i	2.15%		
Y	2.25%		2.15%		
{	2.05%	^	2.25%		
}	2.15%	q	2.54%		
T	2.05%				

Fig. 7. Three output examples of the quantum random password generator. Each character of the password is displayed vertically with the probability that the character was measured displayed across from it. The passwords are variable length depending on how many characters were measured more than 2% of the time to account for noise.

This program can create randomized passwords that meet acceptable password guidelines. The passwords are truly random, use upper and lowercase English characters and most commonly used special characters. Similar to the last program, it illustrates the concepts of superposition and uses qubits to represent multiple outcomes simultaneously. In addition to having real world application, the combination program can also help students understand intermediate quantum concepts like noise, H gates and X gates.

7 Results

This study has successfully developed two working quantum programming examples that return valid output. Both programs illustrate the quantum concepts of superposition and how to utilize gates to perform operations on qubits. The programs also demonstrate the efficiency of performing calculations using qubits rather than using classical bits.

The first program, Quantum War, shows how qubits can simulate a tangible activity such as drawing from a deck of cards. The program runs on the local simulator as well as the 16 qubit IBMQX5 machine available over the internet. The program is more economical in its representation of cards than a classical machine and each card drawn is truly randomized. Figure 5 on page 556 represents a single output example showing a single battle in which each player has drawn a card from a randomized quantum deck.

The second program, Quantum Random Password Generator, successfully produces truly random passwords. It uses all English ASCII characters, both uppercase and lowercase, and commonly used special characters. In 98% of

attempts the program generates a password of 7–14 characters in length, which will satisfy most password requirements. The program can be easily adjusted to increase or decrease the quantity of characters within the output. The program runs on the local simulator as well as the IBMQX5 quantum machine. Figure 7 on page 559 represents three output examples from the Quantum Random Password Generator as it is currently configured.

This study utilized various IBM Q Experience tutorials and the QISKit software development kit to create a development environment using Python 3.5, Conda, and Jupyter Notebook. The IBM tutorials were student friendly, however the difficulty increased significantly after introduction to the basics. The IBM QISKit tutorials require more intermediate level examples that illustrate simple operations using quantum gates with less abstract functionality. This study offers two such practical examples, Quantum War and Quantum Random Password Generator, to the growing body of literature. This study also produced Jupyter Notebook pages that accompany each program and document how the program operates. Jupyter Notebook is an integrated development environment (IDE) that allows developers to create a document that contains runnable code and formatted narrative text tutorials. The notebook pages that accompany each program were written to match the style and substance of existing IBM QISKit tutorials.

The notebooks will serve two purposes. First, they organize the code into understandable blocks and explain how each block functions. This approach is preferable to in-line comments because they are clearer, more visually appealing, easier to understand, and can contain explanations of concepts not normally found in in-line comments. Figure 8 on page 561 provides an example notebook page.

Overall, this study found that Jupyter Notebook pages are better for the student learning experience than parsing in-line comments or reading documentation and code in separate documents. The second purpose of the notebooks is to serve as a tutorial for students interested in quantum computing who have read and understood basic quantum programming examples, but are not ready for more advanced material. This study found that more intermediate material of this nature was needed. As mentioned in the last paragraph, there is a large gap between introductory material and advanced material. This study provides two ready now working intermediate programming examples that can help students bridge the gap from basic quantum programming to understanding advanced examples.

This study has contributed two intermediate quantum programming examples for students interested in advancing their quantum programming knowledge.

Overall, the student quantum experience was a positive one. IBMs tutorials and visual aids to familiarize users with both the programming environment and underlying physical principles are well-crafted and understandable. IBM has clearly documented the functionality of the QISKit on the Q Experience website. IBMs choice to write the QISKit in Python is advantageous because Pythons easily readable syntax clearly demonstrates quantum concepts such as

The circuit contains 6 qubits. All 6 qubits have an H gate applied to them to induce a state of superposition. This will allow all six qubits to potentially represent any of the 52 cards in the deck.

In [7]:

```
# set up quantum program
qcName1 = "deck1"
numQubits1 = 6

qpl = QuantumProgram()
qpl.set_api(Qconfig.APItoken, Qconfig.config["url"]) # :
# declare register of 6 qubits
qr1 = qpl.create_quantum_register("qr", numQubits1)
# declare register of 6 classical bits to hold measurement results
cr1 = qpl.create_classical_register("cr", numQubits1)
# create circuit
qc1 = qpl.create_circuit(qcName1, [qr1], [cr1])

qc1.h(qr1[0])
qc1.h(qr1[1])
qc1.h(qr1[2])
qc1.h(qr1[3])
qc1.h(qr1[4])
qc1.h(qr1[5])
```

Fig. 8. An excerpt from a Jupyter notebook page. This excerpt is from the quantum war program and shows how to create and initialize a quantum circuit. At the top of the figure is the narrative explanation and below a neatly separated block of code that can be run from within Jupyter notebook.

initializing a quantum circuit, applying gates to qubits, and printing the results in a classical format. The only negative of note is that it can be difficult to find clear explanations that bridge the gap between intermediate and advanced quantum topics such as QRAC (Quantum Random Access Code) and Phase Gates.

8 Conclusions

The two examples presented in this paper are available to the researchers at IBM for use in the QISKit tutorial materials. These concrete and familiar examples - a card game and random password generator - have the potential to enrich a students introduction to quantum programming and can assist in filling the gaps between beginner and advanced learning materials.

The future for quantum computing is bright. However, there are still hurdles that must be cleared. High cost, poor error correction, limited access to existing quantum machines for learning and experimentation, and untrained personnel

still stymie development and adoption. While there is a wealth of introductory material available on the web from various publications, news outlets, and educational institutions that covers the background information of quantum computing and quantum mechanics, little content exists outside of IBM QISKit that clearly explains programming for quantum computers. For instance, currently there are none or few tutorials for quantum programming found on popular programming learning sites like Udemy, Udacity, Coursera, and YouTube. As companies such as IBM, Google, and Microsoft scale up their quantum computing research, there will certainly be a need for more quantum programmers.

As argued by R. Van Meter and S. J. Devitt, affordability and scalability of quantum implementations will increase their usefulness in a broad range of fields, including biophysics and security [10]. As the usefulness of quantum computers increases, their availability will increase. As their availability increases, the need for trained quantum programmers will also increase. This study offers simple, high-yield student-to-student teaching examples for introductory courses in quantum computing. Helping students develop literacy in programming for quantum computers will depend on educators who can create clear, applicable, and engaging teaching materials that will attract motivated programmers to the field and help usher in the next revolution in computing.

References

1. Lyshevski, S.E., Iafrate, G.J., Brenner, D., Goddard III, W.A.: There's plenty of room at the bottom: an invitation to enter a new field of physics. In: *Handbook of Nanoscience, Engineering, and Technology*, 2nd edn., pp. 27–36. CRC Press (2007)
2. Arons, A.B., Peppard, M.B.: Einstein's proposal of the photon concepta translation of the annalen der physik paper of 1905. *Am. J. Phys.* **33**(5), 367–374 (1965)
3. Kumar, K., Sharma, N.A., Prasad, R., Deo, A., Khorshed, M.T., Prasad, M., Dutt, A., Ali, A.B.M.S.: A survey on quantum computing with main focus on the methods of implementation and commercialization gaps. In: *2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*, pp. 1–7. IEEE (2015)
4. IBM Q Experience. <https://quantumexperience.ng.bluemix.net> (2018)
5. Almudever, C.G., Lao, L., Fu, X., Khammassi, N., Ashraf, I., Iorga, D., Varsamopoulos, S., Eichler, C., Wallraff, A., Geck, L., et al.: The engineering challenges in quantum computing. In: *2017 Design, Automation and Test in Europe Conference and Exhibition (DATE)*, pp. 836–845. IEEE (2017)
6. Kanellos, M.: Moores Law to Roll on for Another Decade. CNET News.com (2003)
7. Waldrop, M.M.: More than moore. *Nature* **530**(7589), 144–148 (2016)
8. Vu, C.: IBM makes quantum computing available on IBM cloud to accelerate innovation. IBM News Room (2016)
9. DiVincenzo, D.P.: The physical implementation of quantum computation. *Fortschritte der Physik: Prog. Phys.* **48**(9–11), 771–783 (2000)
10. Van Meter, R., Devitt, S.J.: The path to scalable distributed quantum computing. *Computer* **49**(9), 31–42 (2016)
11. Tan, K.Y., Partanen, M., Lake, R.E., Govenius, J., Masuda, S., Möttönen, M.: Quantum-circuit refrigerator. *Nat. Communi.* **8**, 15189 (2017)

12. Lee, P.: Dwave Quantum Computing Company. News Room (2018)
13. (Ginni) Rometty, V.M.: IBM QISKit Github Project. IBM. <https://github.com/QISKit> (2018)
14. Barenco, A., Deutsch, D., Ekert, A., Jozsa, R.: Conditional quantum dynamics and logic gates. Phys. Rev. Lett. **74**(20), 4083 (1995)



Searching for Network Modules

Giovanni Rossi^(✉)

Department of Computer Science and Engineering DISI,
University of Bologna, 40126 Bologna, Italy
giovanni.rossi6@unibo.it

Abstract. When analyzing complex networks, a key target is to uncover their modular structure, which means searching for a family of node subsets spanning each an exceptionally dense subnetwork. Objective function-based graph clustering procedures such as modularity maximization output a partition of nodes, i.e. a family of pair-wise disjoint subsets, although single nodes are likely to be included in multiple or overlapping modules. Thus in fuzzy clustering approaches each node may be included in different modules with different [0, 1]-ranged memberships. This work proposes a novel type of objective function for graph clustering, in the form of a multilinear polynomial extension whose coefficients are determined by network topology. It may be seen as a potential, taking values on fuzzy clusterings or families of fuzzy subsets of nodes over which every node distributes a unit membership. If suitably parameterized, this potential attains its maximum when every node concentrates its all unit membership on some module. Maximizers thus remain partitions, while the original discrete optimization problem is turned into a continuous version allowing to conceive alternative search strategies. The instance of the problem being a pseudo-Boolean function assigning real-valued cluster scores to node subsets, modularity maximization is employed to exemplify a so-called quadratic form, in that the scores of singletons and pairs also fully determine the scores of larger clusters, while the resulting multilinear polynomial potential function has degree 2. After considering further quadratic instances, different from modularity and obtained by interpreting network topology in alternative manners, a greedy local-search strategy for the continuous framework is analytically compared with an existing greedy agglomerative procedure for the discrete case. Overlapping is finally discussed in terms of multiple runs, i.e. several local searches with different initializations.

Keywords: Modularity · Fuzzy clustering · Pseudo-Boolean function

1 Introduction

Networks or graphs are pairs whose elements are a set of nodes or vertices and a set of unordered pairs of nodes, namely the links or edges. In the weighted case, real-valued weights on edges generally measure some intensity of the relation between the two endvertices. A great variety of data may be represented by

means of networks [27], with weighted edges clearly allowing for more flexibility. An oldest and well-known example is found in the social sciences, where nodes are individuals and links formalize friendship/influence relations. More recently, given the increasing availability of genome-scale data on protein interactions, much attention is being paid to PPI (protein-to-protein interaction) networks, where nodes are proteins and links formalize their interaction in metabolic processes [1, 4, 5, 13, 19, 29, 38, 40, 41].

These complex networks describing so different real-world phenomena display several common features [20], and in particular are all organized in modular structures [23, 25]. A module or community basically is a ‘heavy cluster’ of nodes, in the following sense. The subgraph spanned by a subset of vertices includes only these vertices in the subset and those edges whose endvertices are both in the subset. A module then is a vertex subset spanning a subgraph whose edge set is exceptionally dense. In the weighted case, the edges of subgraphs spanned by modules are literally heavy, i.e. collectively receiving a significant fraction of the total weight on all edges. Mathematically speaking, searching for families of network modules is a generalization of graph clustering [36], where the goal is to determine a peculiar family of heavy vertex subsets, namely a partition [3]. A partition of vertices is a collection of non-empty and pair-wise disjoint vertex subsets, called blocks, whose union is the whole vertex set.

1.1 Related Work

The present paper provides an optimization framework and a greedy local-search strategy for graph clustering and network module detection, with focus on fuzzy and/or overlapping modular structures. Objective function-based graph clustering methods determine partitions of vertices relying on optimization, i.e. maximizing or minimizing a score or a cost. Hence the blocks of the generated partition are optimal or maximally heavy vertex subsets as they are found to maximize or minimize a global objective function, where this latter commonly is a so-called additive partition function [12]. This means that a suitable set function assigns a cluster score/cost to every subset of vertices, and then the global score/cost of any partition simply obtains by summing the scores/costs of its blocks. A main example is modularity maximization [8, 24, 25, 35], where in particular the scores of vertex subsets depend only on the scores of the n singletons and the $\binom{n}{2}$ pairs, denoting by n the total number of vertices.

In a partition every vertex is included in precisely one block, while network modules may well be non-disjoint or overlapping [17, 44], and this seems mostly relevant for PPI networks, where modules are protein complexes. The search for overlapping modular structures often resorts to fuzzy clustering algorithms [21, 30, 42, 43, 46], whose output is a family of fuzzy vertex subsets. Hence vertices may be included in different modules, with different $[0, 1]$ -ranged memberships. One way to look at the whole framework formalized in the sequel is to see it as a tool for extending additive partition functions over families of fuzzy subsets. In fact, as the proposed model conforms any chosen objective function-based graph

clustering method employing an additive partition function to the fuzzy setting, it may be step-wise exemplified for modularity maximization, starting with Sect. 2 hereafter. Next Sect. 3 introduces pseudo-Boolean functions [7] in terms of the cluster score of vertex subsets, while Sect. 4 formalizes the implications of evaluating fuzzy clusterings through the polynomial multilinear extension of set functions. Toward tight comparison with modularity maximization, Sect. 5 defines two topology-based quadratic cluster score functions, with focus respectively on (i) the weighted case, and (ii) the clustering coefficient of spanned subgraphs. Section 6 details a greedy local-search strategy for the continuous framework, and compares it with a well-known fast heuristic for modularity maximization. Overlapping is discussed in Sect. 6.2 in terms of multiple local searches with different initializations, while the conclusion in Sect. 7 also briefly examines how to model random networks with overlapping modules, in view of future work.

2 Modularity

Denote by $N = \{1, \dots, n\}$ the n -set of vertices and by $2^N = \{A : A \subseteq N\}$ the 2^n -set of vertex subsets. A network is usually intended to be a simple graph $G = (N, E)$, i.e. with edge set $E \subseteq N_2$ included in the $\binom{n}{2}$ -set of unordered pairs of vertices: $N_2 = \{\{i, j\} : 1 \leq i < j \leq n\} \subset 2^N$, where (\cdot, \cdot) and $\{\cdot, \cdot\}$ respectively are ordered and unordered pairs, while \subset is proper inclusion. Such an edge set E is identified by its characteristic function $\chi_E : N_2 \rightarrow \{0, 1\}$, defined by

$$\chi_E(\{i, j\}) = \begin{cases} 1 & \text{if } \{i, j\} \in E \\ 0 & \text{if } \{i, j\} \in E^c = N_2 \setminus E \end{cases}, \text{ and thus Boolean vector } \chi_E \in \{0, 1\}^{\binom{n}{2}}$$

is an extreme point of the $\binom{n}{2}$ -dimensional unit hypercube $[0, 1]^{\binom{n}{2}}$. Networks with $[0, 1]$ -ranged weights on edges correspond to non-extreme points of this $\binom{n}{2}$ -cube. In other terms, their edge set is fuzzy: it includes unordered pairs of vertices, each with a membership in $[0, 1]$. In this view, weighted networks $G = (N, W)$, $W \in [0, 1]^{\binom{n}{2}}$ thus comprise non-weighted ones as those $2^{\binom{n}{2}}$ special cases where all the $\binom{n}{2}$ weights range in $\{0, 1\}$. Let w_{ij} denote the weight on pair $\{i, j\} \in N_2$ or equivalently the ij -th entry of vector W .

Modularity maximization is a fundamental approach to module detection in complex networks via objective function-based graph clustering, and is usually applied to the non-weighted case. Modularity is an additive partition function, meaning that it takes real values on families $P = \{A_1, \dots, A_{|P|}\}$ of non-empty vertex subsets $A_1, \dots, A_{|P|} \in 2^N$, called ‘blocks’, such that $A \cap A' = \emptyset$ for all $A, A' \in P$ and $A_1 \cup \dots \cup A_{|P|} = N$, where $|\cdot|$ is the number of elements or cardinality of a set. Modularity is denoted by Q and defined, for given network $G = (N, W)$, in terms of the following quantities: $w_i = \sum_{j \in N \setminus i} w_{ij}$ for every vertex $i \in N$ and $w_N = \sum_{\{i, j\} \in N_2} w_{ij}$. In non-weighted networks $G = (N, \chi_E)$, $E \subseteq N_2$ these quantities respectively are the degree $d_i = w_i$ or number of neighbors of every vertex i and the total number $|E| = w_N$ of edges, while $w_{ij} = a_{ij}$ is the ij -th entry of the $(n \times n$ symmetric and Boolean) adjacency matrix [9]. On any

partition P modularity \mathcal{Q} takes value $\mathcal{Q}(P) = \frac{1}{2|E|} \sum_{1 \leq i, j \leq n} \left(a_{ij} - \frac{d_i d_j}{2|E|} \right) \delta_P(i, j)$

with $\delta_P(i, j) = \begin{cases} 1 & \text{if } i, j \in A \text{ for some } A \in P \\ 0 & \text{otherwise} \end{cases}$ and where the sum is over the n^2

ordered pairs of vertices, including those n of the form $(i, i), i \in N$. Clearly, $\delta_P(i, i) = 1$ for all partitions P . Therefore,

$$\mathcal{Q}(P) = \sum_{A \in P} v_{\mathcal{Q}}(A) = \sum_{A \in P} \left[\sum_{i \in A} \left(-\frac{d_i^2}{4|E|^2} \right) + \sum_{\{i, j\} \subseteq A} \left(\frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2} \right) \right].$$

The relation between \mathcal{Q} and function $v_{\mathcal{Q}} : 2^N \rightarrow \mathbb{R}$ defined inside square parenthesis shall soon be looked at from a combinatorial perspective. This expression details how modularity is an additive partition function, since global score is the sum over blocks of their score. In fact, $v_{\mathcal{Q}}(A)$ is a measure of cluster score obtained by comparison with a probabilistic model that can only be mentioned here for reasons of space. Apart from constant terms given by the scores of singletons, $v_{\mathcal{Q}}(A)$ is essentially determined by the difference between the fraction $\sum_{\{i, j\} \subseteq A} a_{ij}/|E|$ of edges whose endpoints are both in A , and the expectation $\sum_{\{i, j\} \subseteq A} d_i d_j/(2|E|^2)$ of such a fraction in the so-called configuration model, namely the random graph with same degree sequence $(d_i)_{i \in N}$ [23, p. 200].

A key feature of additive partition functions such as modularity is that the 2^n values $(v_{\mathcal{Q}}(A))_{A \in 2^N}$ of the underlying cluster score function $v_{\mathcal{Q}}$ are fully determined by the $1 + n + \binom{n}{2}$ values taken on the empty set, where obviously $v_{\mathcal{Q}}(\emptyset) = 0$, on the n singletons $\{i\}, i \in N$, where $v_{\mathcal{Q}}(\{i\}) = -\frac{d_i^2}{4|E|^2}$, and on the $\binom{n}{2}$ pairs $\{i, j\} \in N_2$, where $v_{\mathcal{Q}}(\{i, j\}) = \frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2} - \frac{d_i^2}{4|E|^2} - \frac{d_j^2}{4|E|^2}$.

3 Pseudo-Boolean Cluster Score Functions

Subsets or clusters $A, B \in 2^N$ and partitions or clusterings $P, Q \in \mathcal{P}^N$ of vertex set N are elements of two fundamental posets (partially ordered sets), respectively the Boolean lattice $(2^N, \cap, \cup)$ of subsets of N ordered by inclusion \supseteq and the geometric lattice $(\mathcal{P}^N, \wedge, \vee)$ of partitions of N ordered by coarsening \geqslant , i.e. $P \geqslant Q$ if for every $B \in Q$ there is $A \in P$ such that $A \supseteq B$, where \wedge and \vee respectively denote the ‘coarsest-finer-than’ or meet and the ‘finest-coarser-than’ or join operators [3]. Since these posets are finite, lattice functions $v : 2^N \rightarrow \mathbb{R}$ and $V : \mathcal{P}^N \rightarrow \mathbb{R}$ may be dealt with as points $v \in \mathbb{R}^{2^n}$ and $V \in \mathbb{R}^{\mathcal{B}_n}$ in vector spaces.¹ A fundamental basis of these spaces (apart from the canonical one) is provided by the so-called zeta function ζ , which works as follows: for every $A \in 2^N$ and every $P \in \mathcal{P}^N$, define $\zeta_A : 2^N \rightarrow \{0, 1\}$ and $\zeta_P : \mathcal{P}^N \rightarrow \{0, 1\}$ by $\zeta_A(B) = \begin{cases} 1 & \text{if } B \supseteq A \\ 0 & \text{otherwise} \end{cases}$ and $\zeta_P(Q) = \begin{cases} 1 & \text{if } Q \geqslant P \\ 0 & \text{otherwise} \end{cases}$ (for all $B \in 2^N, Q \in \mathcal{P}^N$).

¹ $\mathcal{B}_k = \sum_{1 \leq l \leq k} \mathcal{S}_{k,l}$ is the (Bell) number of partitions of a k -set, while $\mathcal{S}_{k,l}$ is the Stirling number of the second kind, i.e. the number of partitions of a k -set into l blocks [3, 16, 33].

Then, $\{\zeta_A : A \in 2^N\}$ is a basis of \mathbb{R}^{2^n} and $\{\zeta_P : P \in \mathcal{P}^N\}$ is a basis of $\mathbb{R}^{\mathcal{B}_n}$ (with axes indexed respectively by subsets A and partitions P). Set functions v and partition functions V are linear combinations of the elements of these bases, with coefficients $\mu^v(A), A \in 2^N$ and $\mu^V(P), P \in \mathcal{P}^N$. That is to say, $v(\cdot) = \sum_{A \in 2^N} \zeta_A(\cdot) \mu^v(A)$ and $V(\cdot) = \sum_{P \in \mathcal{P}^N} \zeta_P(\cdot) \mu^V(P)$, or equivalently $v(B) = \sum_{A \subseteq B} \mu^v(A)$ and $V(Q) = \sum_{P \leq Q} \mu^V(P)$.

Functions $\mu^v : 2^N \rightarrow \mathbb{R}$ and $\mu^V : \mathcal{P}^N \rightarrow \mathbb{R}$ are the *Möbius inversions* [3, 34] respectively of v and V , obeying recursions $\mu^v(A) = v(A) - \sum_{B \subset A} \mu^v(B)$ and analogously $\mu^V(P) = V(P) - \sum_{Q < P} \mu^V(Q)$, where $P > Q$ denotes *proper coarsening*: there exist at least two blocks $B, B' \in Q$ and a corresponding block $A \in P$ such that $A \supseteq (B \cup B')$.

A partition function V is additive when $V(P) = \sum_{A \in P} v(A)$ at all $P \in \mathcal{P}^N$ for some set function v , in which case Möbius inversions μ^v and μ^V are of course related. In particular, μ^V takes value 0 on all partitions apart (possibly) from those $2^n - n$ where the number of non-singleton blocks is ≤ 1 [14, 15], namely the *modular elements*² [3, 39] of geometric lattice $(\mathcal{P}^N, \wedge, \vee)$, i.e. the top $P^\top = \{N\}$, the bottom $P_\perp = \{\{1\}, \dots, \{n\}\}$, and for $1 < |A| < n$ those with form $P_\perp^A = \{A, \{i_1\}, \dots, \{i_{n-|A|}\}\}$, where $\{i_1, \dots, i_{n-|A|}\} = N \setminus A = A^c$. Recursively, the values of μ^V on these modular elements are:

- (a) $\mu^V(P_\perp) = \sum_{i \in N} v(\{i\}) = n\mu^v(\emptyset) + \sum_{i \in N} \mu^v(\{i\})$,
- (b) $\mu^V(P_\perp^A) = \mu^v(A) + (-1)^{|A|+1} \mu^v(\emptyset)$ for $1 < |A| \leq n$, with $P_\perp^N := P^\top$.

Additive partition functions admit a continuum of set functions v, v' satisfying $\sum_{A \in P} v(A) = \sum_{A \in P} v'(A)$ at all P , the requirements being: for the bottom, $\sum_{i \in N} v(i) = \sum_{i \in N} v'(i)$ or $n\mu^v(\emptyset) + \sum_{i \in N} \mu^v(i) = n\mu^{v'}(\emptyset) + \sum_{i \in N} \mu^{v'}(i)$, and above the bottom, $\mu^{v'}(A) = v(A) - \sum_{B \subset A} \mu^{v'}(B)$ or $v'(A) = v(A)$ (i.e. for $|A| > 1$). While $v(\emptyset) = \mu^v(\emptyset)$ and $v(\{i\}) = \mu^v(\emptyset) + \mu^v(\{i\})$, if v quantifies the cluster score of vertex subsets, then $v(\emptyset) = 0$ hence $v(\{i\}) = \mu^v(\{i\})$ for all $i \in N$ as well as $\mu^v(\{i, j\}) = v(\{i, j\}) - v(\{i\}) - v(\{j\})$ for all $\{i, j\} \in N_2$.

Geometrically, Boolean lattice $(2^N, \cap, \cup)$ is often dealt with as outlined above, namely as the set $\{0, 1\}^n$ of extreme points of the n -cube $[0, 1]^n$, in that subsets $A \in 2^N$ correspond to characteristic functions $\chi_A : N \rightarrow \{0, 1\}$, where $\chi_A(i) = 1$ if $i \in A$ and $\chi_A(j) = 0$ if $j \in A^c$, with $\chi_A = (\chi_A(1), \dots, \chi_A(n)) \in \{0, 1\}^n$. In this view, set functions $v : 2^N \rightarrow \mathbb{R}$ are pseudo-Boolean functions $\hat{f}^v : \{0, 1\}^n \rightarrow \mathbb{R}$ [7], with polynomial MLE (multilinear extension) $f^v : [0, 1]^n \rightarrow \mathbb{R}$ over the cube

$$\text{given by } f^v(q) = \sum_{A \in 2^N} \left(\prod_{i \in A} q_i \right) \mu^v(A), \text{ where } \prod_{i \in \emptyset} q_i := 1$$

and $q = (q_1, \dots, q_n) \in [0, 1]^n$ is any fuzzy subset of N . This is an extension in that $f^v(\chi_A) = \sum_{B \subseteq A} \mu^v(B) = v(A)$ at all extreme points $\chi_A \in \{0, 1\}^n$. Polynomial

² Modularity \mathcal{Q} is meant to evaluate modular structures in complex networks, while the modular elements of $(\mathcal{P}^N, \wedge, \vee)$ are those partitions \hat{P} realizing, for all $Q \in \mathcal{P}^N$, equality $r(\hat{P} \wedge Q) + r(\hat{P} \vee Q) = r(\hat{P}) + r(Q)$, where $r(P) = n - |P|$ is the rank (see [3] on modular lattices/lattice functions).

$f^v(q)$ is multilinear in n variables q_1, \dots, q_n , with degree $\max\{|A| : \mu^v(A) \neq 0\}$ and coefficients given by the non-zero values $\mu^v(A) \neq 0$ of Möbius inversion. If $\mu^v(A) = 0$ for all $A \in 2^N, |A| > 1$, then f^v is *linear* (and v is a *valuation* [3] of Boolean lattice $(2^N, \cap, \cup)$, i.e. $v(A \cap B) + v(A \cup B) = v(A) + v(B)$ for all $A, B \in 2^N$). Similarly, if $\mu^v(A) = 0$ for $A \in 2^N, |A| > 2$, then f^v is *quadratic*. If $\mu^v(\emptyset) = 0$, then a linear f^v satisfies $v(A) = \sum_{i \in A} \mu^v(\{i\})$, while a quadratic f^v satisfies $v(A) = \sum_{i \in A} \mu^v(\{i\}) + \sum_{\{i,j\} \subseteq A} \mu^v(\{i,j\})$.

Cluster score function v_Q for modularity $Q(P) = \sum_{A \in P} v_Q(A)$ is quadratic, with coefficients $\mu^{v_Q}(\{i\}) = v_Q(\{i\}) = -[d_i/(2|E|)]^2$ for singletons $\{i\}$ and $\mu^{v_Q}(\{i,j\}) = [a_{ij} - d_i d_j / (2|E|)] / |E|$ for pairs $\{i,j\}$. Concerning conditions (a)–(b) above, define cluster score function v'_Q by $v'_Q(\{i\}) = \sum_{j \in N} v_Q(\{j\})/n$ for all vertices $i \in N$. This means that in v_Q every vertex has its own score when considered as a singleton cluster, while in v'_Q all vertices score the same as singletons. Condition (a) holds since $\sum_{i \in N} v'_Q(\{i\}) = \sum_{i \in N} -d_i^2 / (4|E|^2) = \sum_{i \in N} v_Q(\{i\})$. If $\mu^{v'_Q}(\{i,j\}) = v_Q(\{i,j\}) - \mu^{v'_Q}(\{i\}) - \mu^{v'_Q}(\{j\})$ for pairs and $\mu^{v'_Q}(A) = 0$ for $1 \neq |A| \neq 2$, then condition (b) holds too, hence $\sum_{A \in P} v'_Q(A) = \sum_{A \in P} v_Q(A)$ for all $P \in \mathcal{P}^N$.

4 Fuzzy Clustering

Denote by $2_i^N = \{A : i \in A \in 2^N\}$ the 2^{n-1} -set consisting of all subsets where each $i \in N$ is included, and by Δ_i the associated $2^{n-1} - 1$ -dimensional unit simplex whose extreme points are indexed by these subsets $A \in 2_i^N$. Slightly modifying the notation introduced in Sect. 3, from now on let $q_i \in \Delta_i$ be a generic membership distribution, with $q_i^A \in [0, 1]$ quantifying i 's membership in cluster A . That is to say, $q_i : 2_i^N \rightarrow [0, 1]$ with $q_i(A) = q_i^A$ and $\sum_{A \in 2_i^N} q_i^A = 1$.

Definition 1. A fuzzy cover $\mathbf{q} = \{q^A : A \in 2^N\}$ is a collection of 2^n fuzzy clusters $q^A = (q_1^A, \dots, q_n^A) \in [0, 1]^n$, where $q_i^A \in [0, 1]$ if $i \in A$ and $q_j^A = 0$ if $j \in A^c$, while $\sum_{A \in 2_i^N} q_i^A = 1$ for all $i \in N$.

Apart from zero entries, fuzzy covers \mathbf{q} thus essentially correspond to n -tuples $(q_1, \dots, q_n) \in \prod_{i \in N} \Delta_i$ of membership distributions [17, 22]. The originality of the present contribution develops from evaluating fuzzy network modules through the MLE f^v of an underling cluster score function v , entailing that fuzzy covers $\mathbf{q} = \{q^A : A \in 2^N\}$ attain additive global score $F^V(\mathbf{q})$ given by the sum of the 2^n values taken by f^v , i.e.

$$F^V(\mathbf{q}) = \sum_{A \in 2^N} f^v(q^A) = \sum_{A \in 2^N} \sum_{B \supseteq A} \left(\prod_{i \in A} q_i^B \right) \mu^v(A). \quad (1)$$

In pseudo-Boolean optimization [7], the goal is to minimize or maximize a pseudo-Boolean function $\hat{f}^w : \{0, 1\}^n \rightarrow \mathbb{R}$, i.e. a set function $v : 2^N \rightarrow \mathbb{R}$, with MLE $f^v : [0, 1]^n \rightarrow \mathbb{R}$ thus allowing to turn several discrete optimization

problems into a continuous setting. In near-Boolean optimization [32], the objective function has the form of $F^V(\mathbf{q})$ above, and the MLE allows to deal with discrete optimization problems involving additive partition functions (namely maximum-weight set partitioning/packing) into a continuous setting.

Definition 2. A fuzzy clustering is a fuzzy cover \mathbf{q} such that for all $A \in 2^N$ condition $|\{i : q_i^A > 0\}| \in \{0, |A|\}$ holds.

Thus in a fuzzy clustering (or fuzzy partition), for all $A \in 2^N$, the number of those $i \in A$ with strictly positive membership $q_i^A > 0$ is either 0 or else $|A|$. As shown below, the set of values taken by F^V on fuzzy covers coincides with the set of values taken (solely) on fuzzy clusterings.

Proposition 1. *For any set function v , the range of F^V is saturated by the values taken on fuzzy clusterings.*

Proof. In a fuzzy cover $\mathbf{q} = \{q^A : A \in 2^N\}$, for some (\supseteq -minimal) $A \in 2^N$, let $A_{\mathbf{q}}^+ = \{i : q_i^A > 0\}$ satisfy $\emptyset \subset A_{\mathbf{q}}^+ \subset A$, i.e. $0 < |A_{\mathbf{q}}^+| = \alpha < |A|$. Then,

$$F^V(\mathbf{q}) = \sum_{B \in 2^{A_{\mathbf{q}}^+}} f^v(q^B) + \sum_{A' \in 2^N \setminus 2^{A_{\mathbf{q}}^+}} f^v(q^{A'}),$$

where $2^A = \{B : B \subseteq A\}$ for all $A \in 2^N$. In particular,

$$f^v(q^A) = \sum_{B \in 2^{A_{\mathbf{q}}^+}} \left(\prod_{i \in A_{\mathbf{q}}^+} q_i^A \right) \mu^v(B).$$

Consider another fuzzy cover $\hat{\mathbf{q}}$ such that $\hat{q}^{A'} = q^{A'}$ for all $A' \in 2^N \setminus 2^{A_{\mathbf{q}}^+}$, while $\hat{q}_i^A = 0$ for all $i \in A$, with group membership $q_A^A = \sum_{i \in A} q_i^A = \sum_{i \in A_{\mathbf{q}}^+} q_i^A$ redistributed over subsets $B \in 2^{A_{\mathbf{q}}^+}$ according to:

$$\sum_{B \in (2_i^N \cap 2^{A_{\mathbf{q}}^+})} \hat{q}_i^B = q_i^A + \sum_{B \in (2_i^N \cap 2^{A_{\mathbf{q}}^+})} q_i^B \text{ for all } i \in A_{\mathbf{q}}^+,$$

$$\prod_{i \in B} \hat{q}_i^B = \prod_{i \in B} q_i^B + \prod_{i \in B} q_i^A \text{ for all } B \in 2^{A_{\mathbf{q}}^+}, |B| > 1.$$

These $2^\alpha - 1$ equations with $\sum_{1 \leq k \leq \alpha} k \binom{\alpha}{k} > 2^\alpha$ variables $\hat{q}_i^B, \emptyset \neq B \in 2^{A_{\mathbf{q}}^+}$ admit a continuum of solutions or fuzzy covers $\hat{\mathbf{q}}$ where $\sum_{B \in 2^{A_{\mathbf{q}}^+}} f^v(\hat{q}^B) = f^v(q^A) + \sum_{B \in 2^{A_{\mathbf{q}}^+}} f^v(q^B) \Rightarrow F^V(\mathbf{q}) = F^V(\hat{\mathbf{q}})$.

When reiterated for all (if any) $A' \in 2^N \setminus 2^{A_{\mathbf{q}}^+}$ where $0 < |\{i : q_i^{A'} > 0\}| < |A'|$, this procedure yields a final fuzzy clustering $\hat{\mathbf{q}}^*$ satisfying $F^V(\mathbf{q}) = F^V(\hat{\mathbf{q}}^*)$. \square

Example 1. Let $A = \{1, 2, \dots\} \supset A_{\mathbf{q}}^+ = \{1, 2\}$, hence

$$f^v(q^A) = q_1^A \mu^v(\{1\}) + q_2^A \mu^v(\{2\}) + q_1^A q_2^A \mu^v(\{1, 2\}),$$

with the following three conditions for $\hat{\mathbf{q}}$

$$\begin{aligned}
& - \hat{q}_1^{\{1,2\}} + \hat{q}_1^{\{1\}} = q_1^{\{1,2\}} + q_1^{\{1\}} + q_1^A, \\
& - \hat{q}_2^{\{1,2\}} + \hat{q}_2^{\{2\}} = q_2^{\{1,2\}} + q_2^{\{2\}} + q_2^A, \\
& - \hat{q}_1^{\{1,2\}} \hat{q}_2^{\{1,2\}} = q_1^{\{1,2\}} q_2^{\{1,2\}} + q_1^A q_2^A,
\end{aligned}$$

and four variables $\hat{q}_1^{\{1\}}, \hat{q}_1^{\{1,2\}}, \hat{q}_2^{\{2\}}, \hat{q}_2^{\{1,2\}}$. One solution is

$$\begin{aligned}
\hat{q}_1^{\{1,2\}} &= \hat{q}_2^{\{1,2\}} = \sqrt{q_1^{\{1,2\}} q_2^{\{1,2\}} + q_1^A q_2^A} > 0, \\
\hat{q}_1^{\{1\}} &= q_1^{\{1,2\}} + q_1^{\{1\}} + q_1^A - \sqrt{q_1^{\{1,2\}} q_2^{\{1,2\}} + q_1^A q_2^A} > 0, \\
\hat{q}_2^{\{2\}} &= q_2^{\{1,2\}} + q_2^{\{2\}} + q_2^A - \sqrt{q_1^{\{1,2\}} q_2^{\{1,2\}} + q_1^A q_2^A} > 0.
\end{aligned}$$

A main advantage of fuzzy clusters over hard ones is that they may display non-empty pair-wise intersections while also maintaining a unit (cumulative) membership that every $i \in N$ distributes over 2_i^N [30, 44, 46]. Yet, if fuzzy clusterings are evaluated via MLE as in expression (1), then they cannot score better than hard ones or partitions $P = \{A_1, \dots, A_{|P|}\}$, where these latter correspond to 2^n -collections $\mathbf{p} = \{p^A : A \in 2^N\}$ defined by $p^A = \begin{cases} \chi_A & \text{if } A \in P \\ \mathbf{0} & \text{if } A \in 2^N \setminus P \end{cases}$, with $\mathbf{0} \in \{0, 1\}^n$ denoting the all-zero n -vector. Apart from zero entries, \mathbf{p} coincides with the collection $(\chi_{A_1}, \dots, \chi_{A_{|P|}})$ of the characteristic functions of P 's blocks, which are pair-wise disjoint extreme points of the n -cube, i.e. $\langle \chi_{A_l}, \chi_{A_k} \rangle = 0$ for $1 \leq l < k \leq |P|$, satisfying $\chi_1 + \dots + \chi_{A_{|P|}} = \chi_N = \mathbf{1}$, where $\langle \cdot, \cdot \rangle$ denotes scalar product and $\mathbf{1} \in \{0, 1\}^n$ is the all-one n -vector. Expression (1) evaluates partitions $P \in \mathcal{P}^N$ as these collections $\mathbf{p} \subset \{0, 1\}^n$ of disjoint extreme points of the n -cube: $F^V(\mathbf{p}) = \sum_{A \in 2^N} f^v(p^A) = \sum_{A \in P} f^v(\chi_A) = \sum_{A \in P} v(A)$.

Proposition 2. *For any fuzzy clustering \mathbf{q} and set function v , some partitions P, P' satisfy $F^V(\mathbf{p}) \geq F^V(\mathbf{q}) \geq F^V(\mathbf{p}')$.*

Proof. By isolating the contribution of membership q_i to objective function $F^V(\mathbf{q}) = F^V(q_i | \mathbf{q}_{-i})$ when all other $n - 1$ memberships $q_j, j \neq i$ are given,

$$F^V(\mathbf{q}) = F_i^V(q_i | \mathbf{q}_{-i}) + F_{-i}^V(\mathbf{q}_{-i}), \quad (2)$$

where $F^V(\mathbf{q}) = \sum_{A \in 2_i^N} f^v(q^A) + \sum_{A' \in 2^N \setminus 2_i^N} f^v(q^{A'})$,

$$\begin{aligned}
F_i^V(q_i | \mathbf{q}_{-i}) &= \sum_{A \in 2_i^N} q_i^A \left[\sum_{B \subseteq A \setminus i} \left(\prod_{j \in B} q_j^A \right) \mu^v(B \cup i) \right] \\
\text{and } F_{-i}^V(\mathbf{q}_{-i}) &= \sum_{A \in 2_i^N} \left[\sum_{B \subseteq A \setminus i} \left(\prod_{j \in B} q_j^A \right) \mu^v(B) \right] + \\
&\quad + \sum_{A' \in 2^N \setminus 2_i^N} \left[\sum_{B' \subseteq A'} \left(\prod_{j' \in B'} q_{j'}^{A'} \right) \mu^v(B') \right].
\end{aligned}$$

Define $v_{\mathbf{q}_{-i}} : 2_i^N \rightarrow \mathbb{R}$ by

$$v_{\mathbf{q}_{-i}}(A) = \sum_{B \subseteq A \setminus i} \left(\prod_{j \in B} q_j^A \right) \mu^v(B \cup i). \quad (3)$$

Let $\mathbb{A}_{\mathbf{q}_{-i}}^+ = \arg \max v_{\mathbf{q}_{-i}}$ and $\mathbb{A}_{\mathbf{q}_{-i}}^- = \arg \min v_{\mathbf{q}_{-i}}$, where $\emptyset \subset \mathbb{A}_{\mathbf{q}_{-i}}^+, \mathbb{A}_{\mathbf{q}_{-i}}^- \subseteq 2_i^N$. Most importantly,

$$F_i^V(q_i | \mathbf{q}_{-i}) = \sum_{A \in 2_i^N} \left(q_i^A \cdot v_{\mathbf{q}_{-i}}(A) \right) = \langle q_i, v_{\mathbf{q}_{-i}} \rangle. \quad (4)$$

In words, for given membership distributions $q_j, j \neq i$, global score is affected by i 's membership distribution q_i through a scalar product. In order to maximize (or minimize) F^V by suitably choosing q_i for given \mathbf{q}_{-i} , the whole of i 's membership mass has to be placed over $\mathbb{A}_{\mathbf{q}_{-i}}^+$ (or $\mathbb{A}_{\mathbf{q}_{-i}}^-$), anyhow. Hence there are precisely $|\mathbb{A}_{\mathbf{q}_{-i}}^+| > 0$ (or $|\mathbb{A}_{\mathbf{q}_{-i}}^-| > 0$) available extreme points of Δ_i . After reiteration for all $i \in N$, the outcome shall generally consist of two fuzzy covers $\bar{\mathbf{q}}$ and $\underline{\mathbf{q}}$ such that $F^V(\bar{\mathbf{q}}) \geq F^V(\mathbf{q}) \geq F^V(\underline{\mathbf{q}})$ as well as $\bar{q}_i, \underline{q}_i \in ex(\Delta_i)$, where $ex(\Delta_i)$ is the 2^{n-1} -set of extreme points of simplex Δ_i . When this is combined with Proposition 3, the desired conclusion follows. \square

These findings suggest to search for optimal partitions through reiterated improvements of the objective function: $F^V(\mathbf{q}(t+1)) > F^V(\mathbf{q}(t)), t = 0, 1, \dots$, with a cluster score function v and an initial fuzzy clustering $\mathbf{q}(0)$ as inputs.

5 Weights and Common Neighbors

Toward comparison with modularity, a first quadratic cluster score function obtains by focusing on weighted networks, where it seems natural to set the score $v(\{i, j\})$ of pairs equal to edge weights w_{ij} . Since \emptyset scores 0, the quadratic form then only requires to define the score $v(\{i\})$ of singletons. Basically, the greater $w_i = \sum_{j \in N \setminus i} w_{ij}$, the smaller the score of i as a singleton cluster. To formalize this, attention is placed on the complement or dual edge set $\bar{W} \in [0, 1]^{n \choose 2}$, whose entries $\bar{w}_{ij} = 1 - w_{ij}$ measure a repulsion between i and j . A second quadratic cluster score function is defined by focusing on the clustering coefficient (of non-weighted networks). A distinctive feature of several complex networks (including social ones) is an high density of ‘triangles’, namely triples of nodes any two of which are linked. Such a density is measured in terms of a $[0, 1]$ -ranged ratio by the so-called clustering coefficient, which is in fact the probability that by picking at random a vertex and two of its neighbors these latter are also neighbors of each other [20, 28]. A cluster score function can thus be conceived to measure the density of triangles *locally*, namely in the subgraphs spanned by vertex subsets. To this end, the score of any pair of vertices shall be determined by counting their common neighbors [2, 44].

5.1 Score of Singletons and Dual Weights

Given a weighted network $G = (N, W)$, consider the cluster score function v defined by $v(\emptyset) = 0$ and $v(\{i, j\}) = w_{ij}$ for pairs, while for singletons

$$v(\{i\}) = \sum_{j \in N \setminus i} \frac{1 - w_{ij}}{2(n-1)} = \frac{n-1-w_i}{2(n-1)} \quad (5)$$

as well as $\mu^v(A) = 0$ for larger subsets $A, |A| > 2$. Weight w_{ij} and its dual $\bar{w}_{ij} = 1 - w_{ij}$ may measure respectively an attraction and a repulsion, while these quadratic cluster scores load the former on pairs and the latter on singletons. In particular, \bar{w}_{ij} is equally shared between i and j , and the score of each singleton is the arithmetic mean of its $n-1$ shares. Therefore, $v(\{i\}) \in [0, \frac{1}{2}]$ attains the upper bound on isolated vertices, namely those i such that $w_i = 0$, and the lower one on those i such that $w_i = n-1$. The values of Möbius inversion on pairs are

$$\begin{aligned} \mu^v(\{i, j\}) &= w_{ij} - 2 \frac{\bar{w}_{ij}}{2(n-1)} - \sum_{k \in N \setminus \{i, j\}} \frac{2 - w_{ik} - w_{jk}}{2(n-1)} \\ &= \frac{w_i + w_j}{2(n-1)} - \bar{w}_{ij} = w_{ij} - \left(1 - \frac{w_i + w_j}{2(n-1)}\right). \end{aligned} \quad (6)$$

Cluster score function v , with Möbius inversion μ^v taking non-zero values only on singletons and pairs according to expressions (5) and (6), may be compared with modularity-based v_Q for simple graphs $G = (N, \chi_E)$, $E \subseteq N_2$. With the notation of Sect. 2, $v(\{i\}) = \frac{n-1-d_i}{2(n-1)} \geq 0$ and $v_Q(\{i\}) = \frac{-d_i}{4|E|^2} \leq 0$, while on pairs $\mu^v(\{i, j\}) = a_{ij} - 1 + \frac{d_i+d_j}{2(n-1)}$ and $\mu^{v_Q}(\{i, j\}) = \frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2}$. Hence if $a_{ij} = 1$ then $\mu^v(\{i, j\}), \mu^{v_Q}(\{i, j\}) \geq 0$, while if $a_{ij} = 0$ then $\mu^v(\{i, j\}), \mu^{v_Q}(\{i, j\}) \leq 0$. It may be noted that for singletons both μ^v and μ^{v_Q} assign maximum/minimum cluster score to vertices i of lowest/highest degree d_i .

Denoting by $d_A = \sum_{i \in A} d_i$ the group degree for $A \in 2^N$, expressions (5) and (6) quantify the score of larger clusters as $v(A) = \frac{|A|}{2} + \frac{d_A(|A|-2)}{2(n-1)} - \left(\binom{|A|}{2} - |E(A)| \right)$, where $E(A) = \{\{i, j\} : E \ni \{i, j\} \subseteq A\}$ is the edge set of the subgraph $G(A) = (A, E(A))$ spanned or induced³ by A . The first two summands are evidently rather rough, as they entail that many vertices of high degree constitute a valuable cluster, independently from how many of them are adjacent. However, the third summand is precisely the number of edges that spanned subgraph $G(A)$ lacks with respect to the complete one K_A . In particular, if $G(A) = K_A$ is complete and also a component (or maximal connected subgraph) of G , then $|E(A)| = \binom{|A|}{2}$ and $d_A = |A|(|A|-1)$. Substituting, $v(A) = \frac{|A|}{2} + \binom{|A|}{2} \frac{|A|-2}{n-1}$ is thus the maximum that a $|A|$ -subset of vertices may score in any network. Coming to partitions, if $G = K_N$ is the complete graph, then the additive partition function V resulting from v attains its unique maximum on the coarsest or top partition P^\top , where $V(P^\top) = v(N) = \binom{n}{2}$. In the opposite case, if $G = (N, \emptyset)$ is

³ The terminology and notation used here are standard in graph theory [11].

the empty graph, then additive global score V attains its unique maximum on the finest or bottom partition P_\perp , where $V(P_\perp) = \sum_{i \in N} v(\{i\}) = \frac{n}{2}$. In combinatorial theory, partitions $P = \{A_1, \dots, A_{|P|}\}$ of N correspond to those graphs $G_P = K_{A_1} \cup \dots \cup K_{A_{|P|}}$ on N each of whose components is complete (partitions being elements of the so-called polygon matroid on the edges of the complete graph of order n [3]). In terms of graph clustering problems, these \mathcal{B}_n partition-like graphs clearly constitute the simplest conceivable instances, in that global score attains its unique maximum $V(P) = \sum_{A \in P} \left(\frac{|A|}{2} + \binom{|A|}{2} \frac{|A|-2}{n-1} \right)$ precisely on the corresponding partition.

5.2 Score of Pairs and Common Neighbors

The clustering coefficient of a network $G = (N, E)$ is

$$cc(G) = \frac{3 \times \text{number of included cycles on 3 vertices}}{\text{number of included trees on 3 vertices}},$$

cycles and trees on 3 vertices [11] being also termed respectively ‘triangles’ and ‘connected triples’ [28]. Three vertices i, j, k spanning a complete subgraph $G(\{i, j, k\}) = K_{\{i, j, k\}}$, i.e. a cycle (of length 3), provide 3 trees. If $G(\{i, j, k\})$ is connected but not complete then there is only one tree. A disconnected $G(\{i, j, k\})$ evidently provides no tree. In network analysis $cc(G)$ is a key indicator measuring ‘transitivity’, namely to what extent sharing some common neighbor entails being adjacent, for any two vertices. While in social networks the clustering coefficient is higher than in non-social ones [28], several complex networks display the same asymptotic clustering coefficient as certain strongly regular graphs, in contrast to small-world networks [6, 20].

Now consider the aim to assign scores $v(A)$ to clusters A in a way such that higher values of the clustering coefficient $cc(G(A))$ over spanned subgraphs provide greater scores. A natural way to achieve this is by means of a *cubic* pseudo-Boolean function, i.e. such that $\mu^v(A) = 0$ if $|A| > 3$, which also seems to best interpret the model of random graphs with clustering where edges within triangles are enumerated separately from other edges [26]. Maintaining the $\binom{n+1}{2}$ values on singletons and pairs given by expressions (5) and (6) for weights $w_{ij} \in \{0, 1\}$, on the $\binom{n}{3}$ 3-subsets $\{i, j, k\} \in 2^N$ Möbius inversion may be defined

$$\text{simply by } \mu^v(\{i, j, k\}) = \begin{cases} \beta & \text{if } G(\{i, j, k\}) = K_{\{i, j, k\}} \text{ is complete,} \\ 0 & \text{if } G(\{i, j, k\}) \text{ is connected but not complete, with} \\ -\beta & \text{if } G(\{i, j, k\}) \text{ is disconnected,} \end{cases}$$

$\beta \in (0, 1]$. The resulting cluster scores incorporate additional reward/penalty for proximity/distance to/from completeness of the spanned subgraph $G(A)$. For $G(A) = K_A$ complete and a component of G , $v(A) = \frac{|A|}{2} + \binom{|A|}{2} \frac{|A|-2}{n-1} + \beta \binom{|A|}{3}$.

However, the same target can be achieved by means of a quadratic f^v where the count of both common and non-common neighbors for any two vertices determines the values taken by Möbius inversion μ^v on pairs. The neighborhood of vertex i in the given network $G = (N, E)$ [2, 44] is $\mathcal{N}_i = \{j : \{i, j\} \in E\}$. Then,

$|\mathcal{N}_i \cap \mathcal{N}_j|$ is the number of neighbors common to both i and j , while symmetric difference $\mathcal{N}_i \Delta \mathcal{N}_j = (\mathcal{N}_i \setminus \mathcal{N}_j) \cup (\mathcal{N}_j \setminus \mathcal{N}_i)$ contains non-common neighbors. Quadratic scores can be defined by $\mu^v(A) = 0$ if $1 \neq |A| \neq 2$, while

$$\mu^v(\{i\}) = \frac{1}{1 + |\mathcal{N}_i|} \quad \left(= v(\{i\}) \right), \quad (7)$$

$$\mu^w(\{i, j\}) = a_{ij} + \frac{|\mathcal{N}_i \cap \mathcal{N}_j| - |\mathcal{N}_i \Delta \mathcal{N}_j|}{|\mathcal{N}_i \cup \mathcal{N}_j|} \quad (8)$$

on singletons and pairs. The resulting cluster score set function is

$$v(A) = |E(A)| + \sum_{i \in A} \frac{1}{1 + d_i} + \sum_{\{i, j\} \subseteq A} \frac{|\mathcal{N}_i \cap \mathcal{N}_j| - |\mathcal{N}_i \Delta \mathcal{N}_j|}{|\mathcal{N}_i \cup \mathcal{N}_j|},$$

hence a spanned subgraph $G(A) = K_A$ which is complete and a component of G scores $v(A) = (|A| - 1)(|A| - 2) + 1$.

The remainder of this work is concerned with the search for partitions, in form $\mathbf{p} \in \Delta_1 \times \cdots \times \Delta_n$, that locally maximize objective function $F^V(\mathbf{q})$ in expression (1), with inputs given by: (i) the $\binom{n+1}{2}$ values of Möbius inversion μ^v on singletons and pairs, and (ii) an initial fuzzy clustering $\mathbf{q}(0)$. Modularity-based v_Q together with these two cluster score functions defined respectively by expressions (5)–(8) shall be further compared as such inputs (i).

6 Greedy Search

How to employ the results of Sect. 4 for graph clustering may be detailed through comparison with the so-called greedy agglomerative approach to modularity maximization [8, 24]. Starting from the finest partition, this algorithm iteratively selects one union of two blocks that results in a maximal increase of global score, thereby yielding a sequence $P(t+1) > P(t)$ of partitions as the search path, where $P(0) = P_\perp$ and $>$ denotes the covering relation between partitions, i.e. $|P(t+1)| = |P(t)| - 1$ and $P(t+1) > P(t)$ (see above). If there are tails, meaning that alternative unions of two blocks of $P(t)$ yield the same maximal increase of global score, then the two blocks to be merged are randomly selected. The stopping criterion is the absence of any further improvement. The iterative procedure thus is the following.

GreedyMerging(v, P)

Initialize: Set $t = 0$ and $P(0) = P_\perp$.

Loop: While $v(A \cup B) - v(A) - v(B) > 0$ for some $A, B \in P(t)$, set $t = t + 1$ and

- [1] select (randomizing in case of tails) $A, B \in P(t-1)$ such that $v(A \cup B) - v(A) - v(B) \geq v(A' \cup B') - v(A') - v(B')$ for all $A', B' \in P(t-1)$,
- [2] define $P(t) = \{A \cup B\} \cup (P(t-1) \setminus \{A, B\})$, i.e. obtain $P(t)$ from $P(t-1)$ by merging A and B .

Output: Set $P^* = P(t)$.

This algorithm has been used to maximize modularity in different types of networks [24]. In terms of combinatorial optimization, it is an heuristic (for the NP-hard maximum-weight set partitioning problem), meaning that its worst-case output is not guaranteed to provide any bounded approximation of optimal global score. In fact, for the $\frac{n}{2}$ -regular graphs considered in [8, Theorem 5.1], GreedyMerging provides a worst-case solution \hat{P} with zero modularity score $\mathcal{Q}(\hat{P}) = 0$, while the optimal solution P^* provides a strictly positive score $\mathcal{Q}(P^*) > 0$. These graphs $G = (N, E)$ have an even number $n > 4$ of vertices, with two disjoint vertex subsets $N^1 = \{i_1, \dots, i_{\frac{n}{2}}\}, N^2 = \{j_1, \dots, j_{\frac{n}{2}}\}$ of equal size. The edge set is $E = E(K_{N^1}) \cup E(K_{N^2}) \cup \{\{i_k, j_k\} : 1 \leq k \leq \frac{n}{2}\}$, where $E(K_{N^1}) = \{\{i, i'\} : \{i, i'\} \subset N^1\}$ and the same for $E(K_{N^2})$; it includes the edges of complete graphs K_{N^1} and K_{N^2} together with all the $\frac{n}{2}$ edges with endpoints $i_k \in N^1$ and $j_k \in N^2$ for $1 \leq k \leq \frac{n}{2}$. Therefore, $|E| = 2\left(\frac{n}{2}\right) + \frac{n}{2} = \frac{n^2}{4}$, with degree $d_i = \frac{n}{2}$ for all $i \in N$. At $t = 0$, for each of the $\binom{n}{2}$ unions of two blocks $\{i\}, \{j\} \in P_\perp$, the corresponding variation $\mathcal{Q}(P(1)) - \mathcal{Q}(P(0))$ of modularity equals $v_Q(\{i, j\}) - v_Q(\{i\}) - v_Q(\{j\}) = \mu^{v_Q}(\{i, j\}) = \frac{a_{ij}}{|E|} - \frac{d_i d_j}{2|E|^2} =$

$$= \begin{cases} 2/n^2 & \text{if } \{i, j\} \in E, \\ -2/n^2 & \text{if } \{i, j\} \in E^c. \end{cases}$$

The worst-case output is $\hat{P} = \{\{i_1, j_1\}, \dots, \{i_{\frac{n}{2}}, j_{\frac{n}{2}}\}\}$, i.e. the partition obtained in $\frac{n}{2}$ iterations through unions $\{i_k\} \cup \{j_k\}, 1 \leq k \leq \frac{n}{2}$, where modularity scores $0 = \mathcal{Q}(\hat{P}) = \sum_{i \in N} v_Q(\{i\}) + \sum_{1 \leq k \leq \frac{n}{2}} \mu^{v_Q}(\{i_k, j_k\}) = -\sum_{i \in N} \frac{d_i^2}{4|E|^2} + \frac{n}{2} \frac{2}{n^2} = -\frac{1}{n} + \frac{1}{n}$. The unique maximizer is $P^* = \{N^1, N^2\}$, where $\mathcal{Q}(P^*) = \sum_{i \in N} v_Q(\{i\}) + 2 \sum_{\{i, i'\} \subset N^1} \mu^{v_Q}(\{i, i'\}) = -\frac{1}{n} + 2\left(\frac{n}{2}\right) \frac{2}{n^2} = \frac{n-4}{2n} > 0$.

For the other two quadratic scores defined in Sect. 5, GreedyMerging provides the same worst-case output even when the input cluster score function v is that defined by expressions (5) and (6), in which case the $\binom{n}{2}$ unions of two blocks of P_\perp result in a variation of global score equal to $v(\{i, j\}) - v(\{i\}) - v(\{j\}) = \mu^v(\{i, j\}) = a_{ij} - 1 + \frac{d_i + d_j}{2(n-1)} = \begin{cases} \frac{n}{2(n-1)} & \text{if } \{i, j\} \in E, \\ \frac{2-n}{2(n-1)} & \text{if } \{i, j\} \in E^c. \end{cases}$ But if the input is v

defined by expressions (7) and (8), then the algorithm surely finds the optimum P^* , as the $\binom{n}{2}$ unions of two blocks of P_\perp result in variation $\mu^v(\{i, j\}) = a_{ij} + \frac{|\mathcal{N}_i \cap \mathcal{N}_j| - |\mathcal{N}_i \Delta \mathcal{N}_j|}{|\mathcal{N}_i \cup \mathcal{N}_j|} = \begin{cases} \frac{2(n-2)}{n+4} & \text{if } \{i, j\} \subset N^1 \text{ or } \{i, j\} \subset N^2, \\ \frac{2}{n} & \text{if } \{i, j\} \in E, i \in N^1, j \in N^2, \\ \frac{6-n}{n-2} & \text{if } \{i, j\} \in E^c. \end{cases}$

Coming to fuzzy clusterings $\mathbf{q} \in \Delta_1 \times \dots \times \Delta_n$, a quadratic v reduces the objective function in expression (1) to

$$F^V(\mathbf{q}) = \sum_{i \in N} v(\{i\}) + \sum_{\{i, j\} \in N_2} \left(\sum_{A \supseteq \{i, j\}} q_i^A q_j^A \right) \mu^v(\{i, j\}). \quad (9)$$

The main advantage of this MLE of additive partition functions V is that it allows search paths to start and develop in the continuum $\Delta_1 \times \dots \times \Delta_n$, although

the output shall be a partition in view on Proposition 5. Designing a search strategy as a sequence $\mathbf{q}(t)$ such that $F^V(\mathbf{q}(t+1)) > F^V(\mathbf{q}(t))$ amounts to formalize: an initial $\mathbf{q}(0)$, how $\mathbf{q}(t+1)$ obtains from the reached $\mathbf{q}(t)$, and a stopping criterion. By starting from an arbitrary input $\mathbf{q}(0)$, the search is local, although the more the initial n membership distributions are each spread over 2_i^N , the more it becomes global. The stopping criterion is determined by the definition of local optimality: if $\mathcal{N}(\mathbf{q}) = \bigcup_{i \in N} \{\hat{q}_i | \mathbf{q}_{-i} : \hat{q}_i \in \Delta_i\}$ is the neighborhood of \mathbf{q} , then

\mathbf{q}^* is a local optimum when $F^V(\mathbf{q}^*) \geq F^V(\hat{\mathbf{q}})$ for all $\hat{\mathbf{q}} \in \mathcal{N}(\mathbf{q}^*)$. In words, the neighborhood of \mathbf{q} contains all n -tuples of membership distributions where $n-1$ distributions are as in \mathbf{q} while only one may vary, and \mathbf{q}^* is a local optimum if $F^V(\mathbf{q}^*)$ is the greatest value taken by F^V over $\mathcal{N}(\mathbf{q}^*)$. It is shown below that for any partition P a necessary and sufficient condition for this local optimality is $v(A) \geq v(A \setminus i) + v(\{i\})$ for all $i \in A$ and all $A \in P$. A typical greedy local search would thus progress through a sequence $\mathbf{q}(t)$ such that $\mathbf{q}(t+1) \in \mathcal{N}(\mathbf{q}(t))$ and $F^V(\mathbf{q}(t+1)) - F^V(\mathbf{q}(t))$ is maximal, but none of these two conditions is here maintained. In fact, $\mathbf{q}(t+1) \notin \mathcal{N}(\mathbf{q}(t))$ as more than one of the n membership distributions $(q_1(t), \dots, q_n(t)) = \mathbf{q}(t)$ shall vary within the same t -th iteration, and rather than being applied directly to the increase $F^V(\mathbf{q}(t+1)) - F^V(\mathbf{q}(t))$ of global score, greediness is applied to average derivatives, defined hereafter.

$$\begin{aligned} \text{The } i\text{-th derivative [7] of the MLE } f^v \text{ of } v \text{ at } x \text{ is } f_i^v(x) = \\ = f^v(x_1, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_n) - f^v(x_1, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_n) = \end{aligned}$$

$$= \sum_{A \in 2_i^N} \left(\prod_{j \in A \setminus i} x_j \right) \mu^v(A) \text{ for } x = (x_1, \dots, x_n) \in [0, 1]^n. \text{ At vertices } (\chi_B)_{B \in 2^N}$$

of the n -cube, $f_i^v(\chi_B) = \begin{cases} v(B) - v(B \setminus i) & \text{if } B \in 2_i^N, \\ v(B \cup i) - v(B) & \text{if } B \in 2^N \setminus 2_i^N. \end{cases}$ This derivative may be reproduced for the MLE F^V of additive partition functions V as follows. For $i \in N$ and $A \in 2_i^N$, define membership distribution q_{i_A} by $q_{i_A}^B = \begin{cases} 1 & \text{if } B = A, \\ 0 & \text{otherwise.} \end{cases}$

Also let $q_{i_\emptyset}^B = 0$ for all $B \in 2_i^N$, noting that q_{i_\emptyset} is *not* a membership distribution, as it places no membership over 2_i^N at all. Now define the i_A -derivative of F^V at \mathbf{q} by $F_{i_A}^V(\mathbf{q}) = F^V(q_{i_A} | \mathbf{q}_{-i}) - F^V(q_{i_\emptyset} | \mathbf{q}_{-i}) = F_i^V(q_{i_A} | \mathbf{q}_{-i}) = v_{\mathbf{q}_{-i}}(A)$, where the last two equalities obtain from expressions (2) and (3) in Sect. 4. Distributions $q_j^A = 1, j \in A \setminus i$ yield $F_{i_A}^V(\mathbf{q}) = v(A) - v(A \setminus i)$, and $F_{i_{\{i\}}}^V(\mathbf{q}) = v(\{i\})$ independently from \mathbf{q} . These $n2^{n-1}$ derivatives $(F_{i_A}^V(\mathbf{q}(t)))_{i \in N, A \in 2_i^N}$ inform about how to obtain $\mathbf{q}(t+1)$ from the reached $\mathbf{q}(t)$ in order to maximize the objective function. In particular, any greedy strategy requires first to make clear what maximum distance may separate $\mathbf{q}(t+1)$ from $\mathbf{q}(t)$. Instead of $\mathbf{q}(t+1) \in \mathcal{N}(\mathbf{q}(t))$, the rule maintained here is the same, *mutatis mutandis*, as for GreedyMerging, namely that precisely one block is formed when transforming $\mathbf{q}(t)$ into $\mathbf{q}(t+1)$. Hence $\sum_{i \in A} q_i^A(t) < |A| = \sum_{i \in A} q_i^A(t+1)$, or equivalently $q^A(t) \neq \chi_A = q^A(t+1)$, for

exactly one A at each t . Given this, greediness is applied to *average derivative*

$$\bar{F}_A^V(\mathbf{q}) = \sum_{i \in A} \frac{v_{\mathbf{q}_{-i}}(A)}{|A|} = \sum_{B \subseteq A} \left[\sum_{i \in B} \left(\prod_{j \in B \setminus i} q_j^A \right) \right] \frac{\mu^v(B)}{|A|}.$$

In view of expression (9), for quadratic f^v this reduces to

$$\bar{F}_A^V(\mathbf{q}) = \frac{1}{|A|} \left[\sum_{i \in A} v(\{i\}) + \sum_{\{i,j\} \subseteq A} (q_i^A + q_j^A) \mu^v(\{i,j\}) \right].$$

Thus the chosen A (at iteration t , to be a block of the output partition \mathbf{p}^* being constructed) is one where $q^A(t) \neq \chi_A$ and $\bar{F}_A^V(\mathbf{q}(t))$ is maximal (randomizing in case of tails). The same iteration t also specifies how to redistribute membership $\sum_{B \in 2^N : B \cap A \neq \emptyset} q_j^B(t)$ over those $B \in 2_j^N$ such that $B \cap A = \emptyset$, for $j \in A^c$.

As for the stopping criterion, a greedy loop stops when $q^A(t) \in \{\mathbf{0}, \chi_A\}$, i.e. $\sum_{i \in A} q_i^A(t) \in \{0, |A|\}$, for all $A \in 2^N$. Thus, ignoring zeros, $\mathbf{q}(t) = \mathbf{p}^*$ is a partition $P^* = \{A_1, \dots, A_{|P^*|}\}$, dealt with in its Boolean representation $\mathbf{p}^* = (\chi_{A_1}, \dots, \chi_{A_{|P^*|}})$. Next, a second loop checks local optimality for this P^* , which attains if $v(A) \geq v(A \setminus i) + v(\{i\})$ for all $i \in A$ and all $A \in P^*$. If the inequality is not satisfied, then the partition updates by splitting block A in the two (new) blocks $A \setminus i$ and $\{i\}$. Finally, as for the starting point $\mathbf{q}(0)$, there surely exist many options, including a simplest (but computationally most demanding) one given by the n -tuple of uniform distributions $q_i^A(0) = 2^{1-n}$ for all $A \in 2_i^N$ and all $i \in N$. Broadly speaking, input $\mathbf{q}(0)$ sets the terms of trade between computational burden and search width, as the more distributions $q_i(0), i \in N$ are each spread over 2_i^N , the more computationally demanding and wider becomes the search. Specifically, if a family $\mathcal{F} = \{A_1, \dots, A_k\} \subset 2^N$ satisfies $q_i^B(0) = 0$ for all $B \in 2^N \setminus (2^{A_1} \cup \dots \cup 2^{A_k})$ and all $i \in N$ as well as $q_i^{A_l}(0) \neq 0$ for all $i \in A_l, 1 \leq l \leq k$, then the algorithm proposed hereafter searches for optimal blocks only inside $2^{A_1} \cup \dots \cup 2^{A_k}$, hence the output cannot be any partition P such that $B \in P$ for some $B \in 2^N \setminus (2^{A_1} \cup \dots \cup 2^{A_k})$. In particular, if the input $\mathbf{q}(0) = \mathbf{p}$ is a partition $P = \{A_1, \dots, A_{|P|}\}$, then the algorithm only checks if \mathbf{p} is a local optimum, and updates if necessary. In the bottom case $\mathbf{q}(0) = \mathbf{p}_\perp$ the output $\mathbf{p}^* = \mathbf{p}_\perp$ coincides with such an input (independently from input μ^v). A seemingly general and flexible manner to choose the initial n membership distributions is the following. Let $\hat{v}(A) = \frac{v(A)}{|A|}$ and consider setting $\mathbf{q}(0)$ via an arbitrary threshold $\theta \geq 0$ by:

$$q_i^A(0) = \begin{cases} 0 & \text{if } \hat{v}(A) \leq \theta \\ \hat{v}(A) / \sum_{B \in 2_i^N : \hat{v}(B) > \theta} \hat{v}(B) & \text{otherwise} \end{cases} \quad (10)$$

for all $i \in A$ and all $A \in 2^N$, entailing $\frac{q_i^A(0)}{q_i^B(0)} = \frac{\hat{v}(A)}{\hat{v}(B)}$ for all $i \in N$ and all $A, B \in 2_i^N$ such that $\hat{v}(A) > \theta < \hat{v}(B)$.

6.1 Local Search

The greedy local-search strategy just described formally is:

GreedyClustering(w, \mathbf{q})

Initialize: Set $t = 0$ and $\mathbf{q}(0)$ as in expression (10).

GreedyLoop: While $0 < \sum_{i \in A} q_i^A(t) < |A|$ for some $A \in 2^N$, set $t = t + 1$ and
(a) select (randomizing in case of tails) one such $A = A^*(t)$ where for all⁴ B :
 $0 < \sum_{i \in B} q_i^B(t) < |B|$ average derivative \bar{F}_A^V is $\bar{F}_A^V(\mathbf{q}(t-1)) \geq \bar{F}_B^V(\mathbf{q}(t-1))$;

(b) set $q_i^A(t) = \begin{cases} 1 & \text{if } A = A^*(t) \\ 0 & \text{if } A \neq A^*(t) \end{cases}$ for all $i \in A^*(t), A \in 2_i^N$;

(c) for all $j \in N \setminus A^*(t)$ and all $A \in 2_j^N : A \cap A^*(t) = \emptyset$, set $q_j^A(t) =$

$$= q_j^A(t-1) + \left(\hat{v}(A) \sum_{\substack{B \in 2_j^N \\ B \cap A^*(t) \neq \emptyset}} q_j^B(t-1) \right) \left(\sum_{\substack{B' \in 2_j^N \\ B' \cap A^*(t) = \emptyset}} \hat{v}(B') \right)^{-1};$$

(d) set $q_j^A(t) = 0$ for all $j \in N \setminus A^*(t)$ and all $A \in 2_j^N : A \cap A^*(t) \neq \emptyset$.

CheckLoop: While $q^A(t) = \chi_A$ and $v(A) < v(\{i\}) + v(A \setminus i)$ for some $A \in 2^N$,
i $\in A$, set $t = t + 1$ and

$$q_i^{\hat{A}}(t) = \begin{cases} 1 & \text{if } |\hat{A}| = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } \hat{A} \in 2_i^N,$$

$$q_j^B(t) = \begin{cases} 1 & \text{if } B = A \setminus i \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } j \in A \setminus i, B \in 2_j^N,$$

$$q_{j'}^{\hat{B}}(t) = q_{j'}^{\hat{B}}(t-1) \quad \text{for all } j' \in A^c, \hat{B} \in 2_{j'}^N.$$

Output: Set $\mathbf{p}^* = \mathbf{q}(t)$.

Proposition 3. *The output \mathbf{p}^* of GreedyClustering satisfies $F^V(\mathbf{p}^*) \geq F^V(\mathbf{q})$ for all $\mathbf{q} \in \mathcal{N}(\mathbf{p}^*)$, i.e. is a local optimum.*

Proof. The case of singleton blocks $\{i\}$, if any, is trivial, in that $p_j^{*A} = 0$ for all $j \neq i$ and all $A \in (2_j^N \cap 2_i^N)$ entails $F^V(q_i|\mathbf{p}_{-i}^*) = F^V(\mathbf{p}^*)$ for any distribution $q_i \in \Delta_i$. Hence let $i \in A \in P^*$ with $|A| > 1$. By switching from p_i^* to $q_i \in \Delta_i$, global score variation is $F^V(q_i|\mathbf{p}_{-i}^*) - F^V(\mathbf{p}^*) = v(\{i\}) - v(A) +$

$$+ \left(q_i^A \sum_{B \in 2^A \setminus 2^{A \setminus i}: |B| > 1} \mu^v(B) + \sum_{B' \in 2^{A \setminus i}} \mu^v(B') \right) = (q_i^A - 1) \sum_{B \in 2^A \setminus 2^{A \setminus i}: |B| > 1} \mu^v(B)$$

with the last equality due to $v(A) - v(A \setminus i) = \sum_{B \in 2^A \setminus 2^{A \setminus i}} \mu^v(B)$. Now assume that $F^V(q_i|\mathbf{p}_{-i}^*) - F^V(\mathbf{p}^*) > 0$, i.e. \mathbf{p}^* is not a local optimum. Since $q_i^A - 1 < 0$ (because $q_i \neq p_i^*$), then $\sum_{B \in 2^A \setminus 2^{A \setminus i}: |B| > 1} \mu^v(B) = v(A) - v(A \setminus i) - v(\{i\}) < 0$, and *CheckLoop* is meant precisely to screen out this. \square

⁴ As usual colon ‘:’ stands for ‘such that’.

For the $\frac{n}{2}$ -regular graphs seen above where in the worst-case *GreedyMerging* provides a null modularity score, it may be observed how *GreedyClustering* surely (and immediately, in terms of the number of iterations) finds the unique optimum, given a reasonable input $\mathbf{q}(0)$. Consider for simplicity the initial n -tuple of uniform distributions, namely $q_i^A(0) = 2^{1-n}$ for all $A \in 2_i^N, i \in N$. Then on every edge $\{i_k, j_k\} \in E$ where $i_k \in N^1$ and $j_k \in N^2$ the average derivative takes value $\bar{F}_{\{i_k, j_k\}}^Q(\mathbf{q}(0)) = \frac{1}{2} \left[-\frac{2}{n^2} + \frac{2}{2^{n-1}} \frac{2}{n^2} \right] = -\frac{1}{n^2} \left(1 - \frac{1}{2^{n-2}} \right) < 0$, while on every subset $A \subseteq N^1$ or $A \subseteq N^2$ its value is $\bar{F}_A^Q(\mathbf{q}(0)) = \frac{1}{|A|} \left[-\frac{|A|}{n^2} + \binom{|A|}{2} \frac{4}{n^2 2^{n-1}} \right] = -\frac{1}{n^2} \left(1 - \frac{|A|-1}{2^{n-2}} \right) < 0$, where F^Q is the MLE of additive partition function Q (i.e. modularity). Hence for $|A| = 2$ there is no difference, but $\bar{F}_A^Q(\mathbf{q}(0))$ increases with $|A|$ up to

$$\bar{F}_{N^1}^Q(\mathbf{q}(0)) = -\frac{1}{n^2} \left(1 - \frac{\frac{n}{2}-1}{2^{n-2}} \right) = \bar{F}_{N^2}^Q(\mathbf{q}(0)),$$

and $\frac{1}{2^{n-2}} < \frac{n-2}{2^{n-1}}$ (as $n > 4$). Similar results obtain for the two quadratic scores defined by expressions (5)–(8).

When n is large, an input of uniform distributions clearly is not viable, in itself (consisting of $n2^{n-1}$ reals) and mostly because of the computational burden at each iteration. On the other hand, if the initial distributions place membership uniformly on pairs only, i.e. $q_i^A(0) = \begin{cases} (n-1)^{-1} & \text{if } |A| = 2 \\ 0 & \text{otherwise} \end{cases}$ for all $i \in N, A \in 2_i^N$, then the search cannot see that some triples of vertices (namely those included in N^1 or in N^2) provide greater score. In fact, not only any two adjacent vertices provide the same score, but with this input *GreedyClustering* is actually prevented from outputting any partition with (some) blocks larger than pairs. In other terms, the worst-case output of *GreedyMerging* becomes certain. Also note that if scores are assigned according to expressions (7) and (8), then those $2\binom{\frac{n}{2}}{2}$ pairs included in N^1 or in N^2 are more valuable than edges $\{i, j\}, i \in N^1, j \in N^2$, because of common neighbors. However, with initial memberships distributed only on pairs the algorithm remains constrained to partition the vertices into blocks no larger than pairs. These observations aim to highlight the crucial role played by locality in the proposed search: if the initial distributions are too dispersed then the search is computationally impossible, but if they are too concentrated, especially over small subsets, then the search space is too small. How to exploit this trade-off toward overlapping module detection is discussed hereafter.

6.2 Overlapping and Multiple Runs

Local-search algorithms are commonly employed by varying the initial candidate solution over multiple runs. The idea is simple: eventually, among the resulting multiple outputs a best one is chosen. Now, multiple outputs of *GreedyClustering*, with varying initial membership distributions, *collectively* constitute a family $\mathcal{F} \subset 2^N$ of overlapping vertex subsets. In other terms, the union of

$k > 1$ (optimal) partitions does provide the sought overlapping modular structure. Formally, if $\mathbf{q}_1(0), \dots, \mathbf{q}_k(0)$ are different initial fuzzy clusterings or inputs, with outputs $\mathbf{p}_1^*, \dots, \mathbf{p}_k^*$ corresponding to partitions $P_1^*, \dots, P_k^* \in \mathcal{P}^N$, then $\mathcal{F} = P_1^* \cup \dots \cup P_k^*$. In fact, the only case where \mathcal{F} displays no overlapping is when these outputs coincide, i.e. $P_l^* = P_{l+1}^*, 1 \leq l < k$. Otherwise, their scores $V(P_l^*) = \sum_{A \in P_l^*} v(A)$ shall be generally different, but for overlapping module detection those outputs with lower score may still be valuable, precisely because set function v measures the score of subsets. That is to say, optimal partitions scoring lower than others in terms of additive partition function V may have some blocks scoring very high in terms of v . Therefore, the union of only some optimal partitions, namely those scoring higher than some threshold, might exclude very important modules. These reasonings lead to see that \mathcal{F} is in fact a *weighted family*, with weights $v(A)$ on its members $A \in \mathcal{F}$ quantified by v .

Given its computational demand for generic initial membership distributions, *GreedyClustering* may run k times only if each input $\mathbf{q}_1(0), \dots, \mathbf{q}_k(0)$ is non-generic, i.e. with all n memberships distributed only on small subsets. This severely restricts the search space by constraining the output partitions to only have small blocks (see above). Then, family \mathcal{F} only contains such small blocks, and thus important large modules might be excluded. On the other hand, in an overlapping structure a large module seems likely to include some small ones. Accordingly, the search for large modules may be restricted by concentrating the initial n membership distributions on those vertex subsets given by the union of family members $A \in \mathcal{F}$. In particular, let $\Omega(\mathcal{F}) \subset 2^N$ be the set system⁵ $\Omega(\mathcal{F}) = \{B : B = A_1 \cup \dots \cup A_m, \mathcal{F} \ni A_1, \dots, A_m, m > 0\}$. This is the collection of all (non-empty) subsets of N resulting from the union of some (i.e. at least one) family members $A \in \mathcal{F}$. As already explained, the search performed by *GreedyClustering* is top-down, as optimal blocks can only be found among \supseteq -maximal subsets $A \in 2^N$ where vertices $i \in A$ initially place strictly positive membership. Weighted family \mathcal{F} consisting precisely of small modules, large ones may be detected by initially distributing memberships only on large subsets $B \in \Omega(\mathcal{F})$. A simplest way to do this is uniformly over those with size $|B| > \vartheta$ exceeding a threshold ϑ , i.e. $q_i^B(0) = \begin{cases} |2_{\Omega_i}^N|^{-1} & \text{if } |B| > \vartheta \\ 0 & \text{otherwise} \end{cases}$ for all $B \in (2_i^N \cap \Omega(\mathcal{F}))$ and all $i \in N$, where $2_{\Omega_i}^N = 2_i^N \cap \{B : B \in \Omega(\mathcal{F}), |B| > \vartheta\}$. However, in the spirit of expression (10), weights or scores $v(B)$ of these $B \in \Omega(N)$ with size exceeding ϑ may be used to determine more suitable non-uniform distributions. In any case, when initial memberships are (non-trivially) distributed over some large subsets the search is computationally more demanding.

⁵ $\Omega(\mathcal{F})$ is a generalization of the field 2^P of subsets generated by partitions P , where $2^{P^\perp} = 2^N$, while $2^{P^\top} = \{\emptyset, N\}$.

7 Conclusion

This work details how to search for network modules by means of a recent approach to objective function-based clustering and set partitioning [31, 32], which applies to any graph clustering problem whose optimal solutions are extremizers of an additive partition function, namely a function assigning to every partition of vertices the sum over blocks of their cluster score. This score of vertex subsets is quantified by a pseudo-Boolean (set) function, which in particular is quadratic when the score of any subset is determined solely by the scores of included singletons and pairs. Although network topology is interpreted mostly in terms of alternative quadratic cluster scores, still here the quadratic form is an option and not a constraint, as cubic cluster scores appear as well, aimed at incorporating the clustering coefficient of spanned subgraphs into the objective function. Modularity being indeed an additive partition function with quadratic cluster scores, the whole setting is thorough detailed for the well-known case of modularity maximization [23, 25, 27, 35].

The optimization-based search for network modules is conceived in the continuous space of fuzzy clusterings, because the objective function is in fact the polynomial multilinear extension of additive partition functions. The extremizers are shown to be partitions of nodes, hence the proposed greedy local-search procedure outputs a graph partition. In particular, a greedy loop generates blocks by maximizing the average derivative, where this latter parallels the standard derivative of pseudo-Boolean functions [7], but most importantly an input fuzzy clustering where to start from makes the search local. The choice of this initialization is crucial, as it balances between computational burden and search width. Suitable inputs might allow for multiple runs, firstly searching for small modules only, and secondly for large ones only. Then, outputs are partitions, whose union is a set system of vertex subsets, i.e. an overlapping modular structure.

7.1 Future Work

How to conceive benchmark graphs for testing overlapping module detection methods seems far from obvious. The comparison between probabilistic models and real-world complex systems is at the heart of network analysis, and such models generally rely on maintaining randomness insofar as possible, while fixing quantities such as the degree sequence and the clustering coefficient [10, 18, 23, 26]. For random networks with ‘fixed’ modular structure, a problem is that there is no precise quantity to fix, namely no measurement of modules in real-world networks. One way to deal with this is to fix a partition P of N and the corresponding partition-like graph $G_P = (N, E_P) = \cup_{A \in P} K_A$ (each of whose components is the complete graph K_A on a block $A \in P$, see Sect. 5), and next introduce ‘noise’ by randomly adding some edges $\{i', j'\} \in N_2 \setminus E_P$ while removing some edges $\{i, j\} \in E_P$. Since blocks are non-overlapping, a random network G with *overlapping* modules then may obtain by developing from the union of k *cliques*: $G = K_{A_1} \cup \dots \cup K_{A_k}$. This means focusing on the *clique-type* $\kappa = (\kappa_1, \dots, \kappa_n)$, i.e. the number κ_m of maximal complete subgraphs on m

vertices, $1 \leq m \leq n$. In fact, despite NP-hardness, recently such an information has become available even for very large networks [37, 45]. The $\frac{n}{2}$ -regular graph G considered in Sect. 6 is the union $G = K_{N^1} \cup K_{N^2} \cup K_{\{i_1, j_1\}} \cup \dots \cup K_{\{\frac{n}{2}, \frac{n}{2}\}}$ of $\frac{n}{2} + 2$ overlapping cliques, with clique-type $\kappa_2 = \frac{n}{2}$, $\kappa_{\frac{n}{2}} = 2$ and $\kappa_m = 0$ for $2 \neq m \neq \frac{n}{2}$. The clique-type κ of graphs is a generalization of the type (or class) of partitions [34], and if it is fixed, then randomness essentially concerns the sizes of pair-wise intersections, i.e. $|A \cap A'|$ for $K_A, K_{A'} \subset G$, which in turn determine vertex degrees (and also seemingly provide the needed noise). More generally, a flexible probabilistic model might employ the clique-type observed in real-world networks as a parameter, rather than maintaining it fixed.

References

1. Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I., Vicsek, T.: CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics* **22**(8), 1021–1023 (2006)
2. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010)
3. Aigner, M.: Combinatorial Theory. Springer, Berlin (1997)
4. Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K., Kanaya, S.: Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinform.* **7**(207) (2006)
5. Asur, S., Ucar, D., Parthasarathy, S.: An ensemble framework for clustering protein-protein interaction networks. *Bioinformatics* **23**, i29–i40 (2007)
6. Bollobás, B., Riordan, O.M.: Mathematical results on scale-free random graphs. In: Bornholdt, S., Schuster, H.G. (eds.) *Handbook of Graphs and Networks: from the Genome to the Internet*, pp. 1–34. Wiley, Berlin (2003)
7. Boros, E., Hammer, P.: Pseudo-Boolean optimization. *Discrete Appl. Math.* **123**, 155–225 (2002)
8. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. *IEEE Trans. Knowl. Data Eng.* **20**(2), 172–188 (2007)
9. Brower, A.E., Haemers, W.H.: Spectra of Graphs. Springer, New York (2011)
10. Chakrabarti, M., Heath, L., Ramakrishnan, N.: New methods to generate massive synthetic networks. *cs. SI*, [arXiv:1705.08473 v1](https://arxiv.org/abs/1705.08473) (2017)
11. Diestel, R.: Graph Theory. Springer, New York (2010)
12. Fortunato, S.: Community detection in graphs. *Phys. Rep.* **486**(3–5), 75–174 (2010)
13. Freeman, T.C., Goldovsky, L., Brosch, M., van Dongen, S., Mazire, P., Grocock, R.J., Freilich, S., Thornton, J., Enright, A.J.: Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLOS Comp. Biol.* **3**(10-e206), 2032–2042 (2007)
14. Gilboa, I., Lehrer, E.: Global games. *Int. J. Game Theory* **20**, 120–147 (1990)
15. Gilboa, I., Lehrer, E.: The value of information—an axiomatic approach. *J. Math. Econ.* **20**(5), 443–459 (1991)
16. Graham, R.L., Knuth, D.E., Patashnik, O.: Concrete Mathematics—A Foundation for Computer Science, 2nd edn. Addison-Wesley, Reading (1994)
17. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.* **11**(3), 033015 (2009)

18. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* **78**(4), 046110 (2008)
19. Lei, X., Wu, S., Ge, L., Zhang, A.: Clustering and overlapping modules detection in PPI network based on IBFO. *Proteomics* **13**(2), 278–290 (2013)
20. Li, Y., Shang, Y., Yang, Y.: Clustering coefficients of large networks. *Inf. Sci.* **382–383**, 350–358 (2017)
21. Miyamoto, S., Ichihashi, H., Honda, K.: Algorithms for Fuzzy Clustering. Springer, Berlin (2008)
22. Nepusz, T., Petróczi, A., Négyessy, L., Baszó, F.: Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E* **77**(1), 016107 (2008)
23. Newman, M.E.J.: The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003)
24. Newman, M.E.J.: Fast algorithm for detecting communities in networks. *Phys. Rev. E* **69**(6), 066133 (2004)
25. Newman, M.E.J.: Modularity and community structure in networks. *PNAS* **103**, 8577–8582 (2006)
26. Newman, M.E.J.: Random graphs with clustering. *Phys. Rev. Lett.* **103**(5), 058701(4) (2009)
27. Newman, M.E.J., Barabási, A.L., Watts, D.J.: The Structure and Dynamics of Networks. Princeton University Press, Princeton (2006)
28. Newman, M.E.J., Park, J.: Why social networks are different from other types of networks. *Phys. Rev. E* **68**(3), 036122 (2003)
29. Pereira-Leal, J.B., Enright, A.J., Ouzounis, C.A.: Detection of functional modules from protein interaction networks. *PROTEINS: Struct. Funct. Bioinform.* **54**, 49–57 (2004)
30. Reichardt, J., Bornholdt, S.: Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.* **93**(21), 218701 (2004)
31. Rossi, G.: Multilinear objective function-based clustering. In: Proceedings of 7th IJCCI, vol. 2. Fuzzy Computation Theory and Applications, pp. 141–149 (2015)
32. Rossi, G.: Near-Boolean optimization—a continuous approach to set packing and partitioning. In: LNCS 10163 Pattern Recognition Applications and Methods, pp. 60–87. Springer (2017)
33. Rota, G.C.: The number of partitions of a set. *Am. Math. Monthly* **71**, 499–504 (1964)
34. Rota, G.C.: On the foundations of combinatorial theory I: theory of Möbius functions. *Z. Wahrscheinlichkeitsrechnung u. verw. Geb.* **2**, 340–368 (1964)
35. Rotta, R., Noack, A.: Multilevel local search clustering algorithms for modularity clustering. *ACM J. Exp. Algorithmics* **16**(2), 2.3:1–27 (2011)
36. Schaeffer, S.E.: Graph clustering. *Comput. Sci. Rev.* **1**, 27–64 (2007)
37. Schmidt, M.C., Samatova, N.F., Thomas, K., Park, B.H.: A scalable, parallel algorithm for maximal clique enumeration. *J. Parallel Distrib. Comput.* **69**(4), 417–428 (2009)
38. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. *Mol. Syst. Biol.* **3**, 88 (2007)
39. Stanley, R.: Modular elements of geometric lattices. *Algebra Universalis* **1**, 214–217 (1971)
40. Szalay-Bekő, M., Palotai, R., Szappanos, B., Kovás, I.A., Papp, B., Csermely, P.: Hierarchical layers of overlapping network modules and community centrality. *Bioinformatics* **28**(16), 2202–2204 (2012)
41. Vlasblom, J., Wodak, S.J.: Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinform.* **10**, 99 (2009)

42. Wang, J., Run, J., Li, M., Wu, F.X.: Identification of hierarchical and overlapping functional modules in PPI networks. *IEEE Trans. Nanobiosci.* **11**(4), 386–393 (2012)
43. Wu, H., Gao, L., Dong, J., Jang, X.: Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein networks. *Plos ONE* **9**(3-e91856) (2014)
44. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state of the art and a comparative study. *ACM Comput. Surv.* **45**(43), 43:1–43:35 (2012)
45. Yu, T., Liu, M.: A linear time algorithm for maximal clique enumeration in large sparse graphs. *Inf. Process. Lett.* **125**, 35–40 (2017)
46. Zhang, S., Wang, R.S., Zhang, X.S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A* **374**, 483–490 (2007)



Routing Sets and Hint-Based Routing

Ivan Avramovic^(✉)

George Mason University, Fairfax, VA 22030, USA
iavramo2@gmu.edu

Abstract. The number of addresses on the Internet grows rapidly, and thus there may be a point at which the state requirements for routing become unwieldy. The intent of this research is twofold. First to draw on compact routing theory with landmark routing, thus reducing router state requirements, but also to make the implementation of theoretical routing protocols with low state requirements more feasible in a policy constrained network. To that end, conceptual organizational scheme called routing sets is presented, which would allow flexibility in the choice of routing policy. Furthermore, an IPv6 extension and algorithm is presented for routing using hints, which moves some of the routing responsibility onto the end hosts, potentially freeing routers of a great deal of the routing state burden.

Keywords: Compact routing · Stretch · Landmark routing · IPv6

1 Introduction

The problem which motivates this work is the problem of routing in a growing network. In a network such as the Internet, resources are not unlimited, but the network is expected to grow, as evidenced by the expansion of address space between IPv4 and IPv6. It is thus desirable to give consideration to how well Internet routing scales in such a growing network [15].

The current Internet favors shortest-path based routing approaches, such as distance-vector and link state protocols in autonomous systems (AS), or the path vector algorithm used in the Border Gateway Protocol (BGP) [17]. A limitation inherent to all such protocols is that the necessary state required per router must be at least on the order of the number of reachable network destinations. Therefore, if the reachable Internet contains n nodes, a typical router will be required to maintain $\Omega(n)$ state for ordinary routing. It should be noted that the use of AS does not necessarily reduce overall routing state, because different AS are required to exchange their full routing state in order for one AS to reach networks in the other.

The other extreme of routing methodologies is to organize the network into a hierarchical structure. Under a hierarchical structure, the routing state requirements given the address of a destination network are trivial, since one only needs to climb up the hierarchy, and descend back down in order to reach the destination node. A pure hierarchy within a large-scale network such as the Internet

carries the potential problem that it may severely limit flexibility in forming network topology. Furthermore, such an addressing structure limits the effectiveness of mobile devices, because they would need to change address, and thus identity, if they change location within the network. Possibly the most significant limitation of a hierarchy is that the actual routing path used between two hosts may be far from the shortest path between the two hosts, and in fact can be arbitrarily large relative to the actual shortest path length [16].

Thus, the concept of *stretch* is introduced. Stretch is defined as the ratio between the routing path to a network destination, and the shortest possible path length to the same destination, using whichever distance metric is applicable to the given network. In other words, stretch is a measure of the indirectness of a given routing path.

Theoretical work exists to show that scalable routing is possible with limited stretch. In particular, if one bounds stretch to a maximum below 5, it can be shown that there is a routing state requirement of $\Omega(\sqrt{n})$ at some node in the network [18]. This bound on routing state can be achieved when using a routing protocol which makes use of routing landmarks.

The goal of this paper is to address the issue of increasing routing scalability by decreasing the state requirements of routers. To that end, a scheme will be presented which will allow flexibility in the choice of routing protocol. In particular, it will not restrict the use of current routing protocols, but it will allow the use of other schemes, such as address hierarchies or embedded landmark schemes. Additionally, a routing protocol will be described which makes use of hints, which extend the effective range of existing routing protocols without increasing state requirements. Effective use of hints will allow the emulation of other routing schemes such as landmark-based routing.

The remainder of this paper will be divided as follows. In Sect. 2, a conceptual overview of the routing methodologies which this paper is presenting is introduced. In Sect. 3, other work in scalable routing is described, as together with how the algorithm in this paper relates to those efforts. In Sect. 4, a protocol for hint-based routing is presented in detail, as it would be implemented in IPv6. Section 5 discusses some of the limitations of the research and the proposed protocol. Section 6 presents the conclusions of the work along with possibilities for further expansion of the research.

2 Overview

In this section, the concept of routing sets (RS) and the associated assumptions will be introduced, and the way in which RS can be used to organize routing will be described (Sect. 2.1). Then, hinting is defined as used in the research, and the way in which it can be applied to achieve scalable routing is described (Sect. 2.2). Finally, the way in which RS with hints can be used to emulate a complex routing scheme such as landmark routing will be explained (Sect. 2.3).

2.1 Routing Sets

Assume that one is given a network of nodes which is participating in routing. Define a *routing set* S as a subset of the network such that for any $u, v \in S$, u and v both have sufficient information to route to each other. Furthermore, u is aware of all network interfaces present at v .

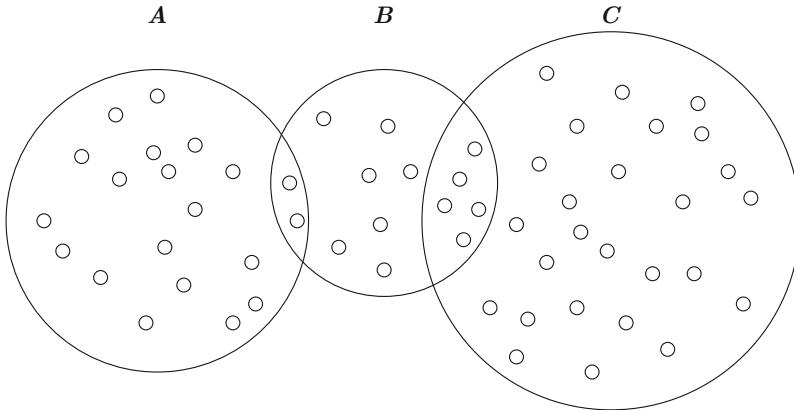


Fig. 1. A network with three routing sets defined, A , B , and C . RS B is familiar to both A and C due to shared nodes.

Any node in the network is permitted to be simultaneously part of more than one RS. It is not assumed that all nodes within a RS form a single connected component, and in fact, the possibility of a disjoint RS is necessary for implementing a landmark scheme for RS, as will be described later (Sect. 2.3). Thus, the RS is a scheme for network organization.

The requirement that a network on the Internet must have all the knowledge to reach any other network on the Internet is relaxed. The network may choose to collect extensive information, but it is not required to do so. With such a relaxation, one would expect that a typical node's routing table would contain more thorough routing information for local networks than for networks which are further away. Instead, routing range is extended by amending routes with hints, which will be described below (Sect. 2.2).

Without the guarantee of complete knowledge, it is no longer sufficient for a physical path to exist between two nodes in order for them to communicate. Suppose that a source node s is attempting to send a message to a destination node d . The message will only reach d if there is a *known path*, denoted $s \rightsquigarrow d$, between them. A known path exists when the following two conditions are satisfied:

1. a set $\{n_1, n_2, \dots, n_k\}$ of zero or more intermediary nodes exists, which are connected by direct links, $s \rightarrow n_1 \rightarrow n_2 \rightarrow \dots \rightarrow n_k \rightarrow d$, and

2. letting $n_0 = s$ and $n_{k+1} = d$, it follows that for any integer i , $0 \leq i \leq k$, if the message reaches node n_i , then there is sufficient information at n_i to route the message to d .

From the definition, if $a \rightsquigarrow b$ and $b \rightsquigarrow c$, then $a \rightsquigarrow c$ follows if a has sufficient information to determine that c can be reached by routing through b .

Since it is assumed that all nodes within a RS are mutually familiar, it will also be assumed that the RS has independent means to populate the routing tables of its nodes, with respect to routing information for other nodes in the RS. These means are likely to include collecting routing information for other nearby nodes, and is not restricted to information about nodes strictly within the RS. A disjoint RS, for example, would need to rely on means beyond local propagation of routing information in order to complete its routing state. However, it may make use of already-established routing state for adjoining network nodes in order to propagate state information.

A pair of RS S and T are said to be *familiar* if they contain a node u which is in both S and T . See Fig. 1. Call u a *mutual* node. Due to the fact that $u \in S$ and $u \in T$, for any $s \in S$ and any $t \in T$, s knows a path to t by virtue of knowing a path to u , and t knows a path to S by virtue of knowing a path to u . Formally, $s, u \in S$ implies $s \rightsquigarrow u$ and $u \rightsquigarrow s$, while $t, u \in T$ implies $t \rightsquigarrow u$ and $u \rightsquigarrow t$, although the nodes in the paths are not required to be completely contained within the corresponding RS. Thus, $s \rightsquigarrow u \rightsquigarrow t$ and $t \rightsquigarrow u \rightsquigarrow s$, so $s \rightsquigarrow t$ and $t \rightsquigarrow s$.

Assume that the RS of a node can be inferred from its address. If a node is part of multiple RS, then its interfaces will have multiple addresses to accommodate each RS. The RS can be inferred by using an RS-mask, the same way that network can be inferred from an address by using a netmask. The RS-mask would account for a smaller portion of the address than a netmask. In order to infer an RS from an address, one must possess the RS-mask of the RS.

If it is assumed that routers which know of an RS also store the RS-mask of the RS, it follows that each such router can identify any node that is within the RS as belonging to the RS. Therefore, not only does a router know how to route within its own RS, but it also knows how to route a packet to any node within any familiar RS.

2.2 Hinted Routing

Hints are defined as identifiers of network routing nodes, which are used to indicate a possible route in the path to some destination node. Hinted routing is comparable to source routing, except that the complete path is not specified, and hops along the path are optional.

If a router does not know the path to a destination node, one or more hints may be used to specify a path. Since nodes generally lie within some RS, it is not necessary to specify a full path for routing, but rather only a sufficient number of nodes to identify the intermediary routing sets. If a node knows a shorter path

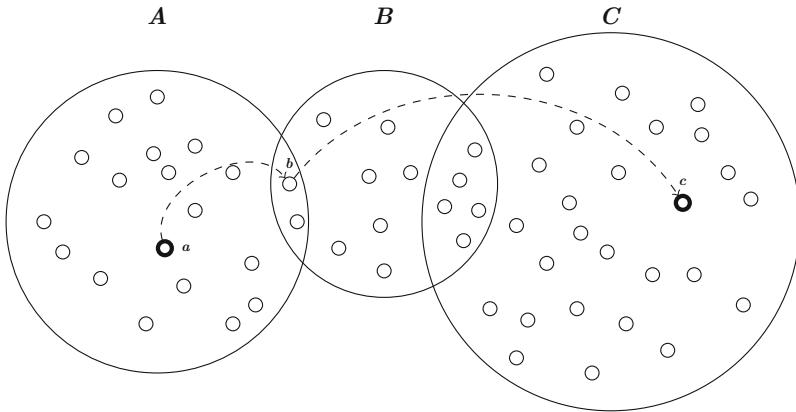


Fig. 2. Node a in A can route to node c in C by specifying node b as a hint. Node a knows how to reach b because they are both in A , and b knows how to reach c because b also lies in B , which is familiar with C .

than one which is hinted, it may ignore some of the hints and use the path it knows instead. See Fig. 2.

Routers may begin with a directory of path hints in order to extend their routing range. Additionally, they may learn of hints which lead to certain destinations from packets that pass through, and these hints may be saved for future use.

In a hint-based routing system, the burden of finding a route to exotic destinations is moved from the router to the end host. An originating router may keep tables of hints as reference, but ultimately an end host is expected to have the necessary prior knowledge to reach an exotic destination.

A specification for hint-based routing in IPv6 is detailed in Sect. 4.

2.3 Landmark Routing

A landmark routing scheme is an effective way to reduce routing state while keeping relatively low stretch. It is implemented by choosing a suitably spaced subset of routing nodes as landmarks, and by having the landmarks learn mutual shortest path lengths. Each landmark would also know about a neighborhood of network nodes, so that a node can be reached by routing to the landmark nearest to the destination, and then routing to the destination directly from the final landmark.

In order to emulate a landmark scheme using RS with hints, one can leverage the fact that a RS is permitted to be disjoint. Suppose that a source s , destination d , and set of landmarks T are given. Furthermore, suppose that $t_s \in T$ and $t_d \in T$ represent the closest landmarks to s and d , respectively. Assume that $s \rightsquigarrow t_s$, $t_s \rightsquigarrow t_d$, and $t_d \rightsquigarrow d$ are known. Routing to a destination can then be performed by including t_d as a hint at the source. Node s will know to route through t_s due

to the fact that t_s and t_d have matching RS. Thus, the path $s \rightsquigarrow t_s \rightsquigarrow t_d \rightsquigarrow d$ is formed. Another landmark algorithm, Disco [16], uses nearest landmark as part of the network address of a node, so the hint system would not be less efficient in that regard.

The advantage of implementing a landmark system with routing sets is that it allows the full range of a landmark system, but it does not impose any requirement on the overall structure of the network. Thus, it may be used where it is beneficial, and not extend elsewhere. This feature is in line with the Internet principles of protocol generality and stability [5].

Like every RS, a landmark RS node is assumed to have full knowledge of all other nodes within the RS. Therefore, the information must be shared between nodes somehow. How it is done is beyond the scope of the paper, although it the exchange of information between nodes should be feasible if the the landmark RS lie over top of some other connecting topology. See Fig. 3.

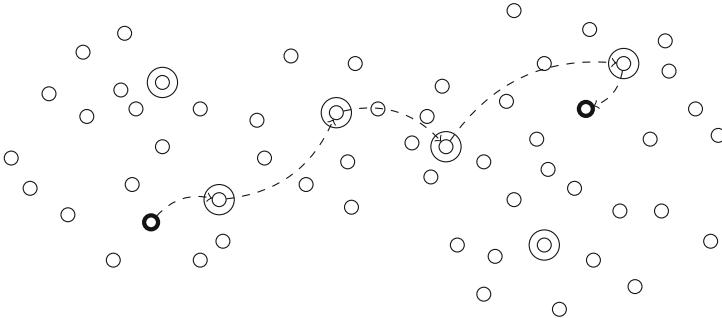


Fig. 3. Routing sets can be disjoint, as shown here where the circled nodes form a landmark RS. If the sending node knows the landmark closest to the destination, then it may use that landmark as a hint, and hop the shortest path from landmark to landmark to reach the destination.

3 Related Work

Theoretical approaches to low-stretch compact routing were presented by Thorup and Zwick [18], and by Abraham et al. [2]. Thorup and Zwick proposed a theoretical family of compact routing algorithms which attained the optimal state versus stretch bounds. The theoretical optimum for state given a maximum stretch of 3 is $\tilde{O}(\sqrt{n})$. Additionally, one can obtain progressive improvements on state requirements by relaxing the worst-case stretch bounds. Abraham et al. showed that the same level of compact routing can be achieved using a *name-independent* addressing model. Name-independent addressing implies that a hostname would not contain any addressing info, and thus the name lookup must be included in the stretch measurement.

A practical attempt at implementing a compact routing scheme using name-independent addressing was the Disco [16] algorithm, which achieves the same state bounds described by Thorup and Zwick. Disco uses landmarks, which are randomly chosen with network-size dependent frequency, to simplify routing. Packets can be transmitted to a destination by first transmitting to the nearest well-known landmark.

There are several drawbacks of Disco which may make it unsuitable as a universal routing algorithm for the Internet. Since landmarks are chosen at random based on node density, and the landmarks are required to be well-known, it would be undesirable for landmarks to be chosen from unreliable, mobile, or untrustworthy nodes. Furthermore, Disco relies on having efficient ways of generating reasonable estimates of network size. This may fail to be the case if a large portion of the network is comprised of mobile nodes, or if large sections of the Internet are not well known to other parts. Emulating a landmark scheme using hint-based routing was described in Sect. 2.3.

The recently proposed U-Sphere protocol [13] attempts to strengthen the security of Disco through the use of keys, validation, and name services which are not centralized on landmarks. It succeeds in making the protocol tolerant to *Sybil adversaries*, although many of the problems related to network disruptions remain.

The concept of routing sets introduced in this paper is independent of, but similar to, the concept of *routing substrates* presented by Drazic and Liebeherr [6] in the Landmark Domains Routing (LDR) scheme. The LDR scheme generalizes routing to include non-Internet networks, and applies the concept of landmark nodes to the network domains themselves. Routing within a domain is handled by the default routing mechanism for that domain.

DistRoute is a bounded stretch landmark scheme based on node colorings [8]. In addition to holding routing information about each landmark and the local vicinity, nodes would keep shortest paths to other nodes of the same color. The algorithm includes provable bounds on messaging requirements.

Another landmark scheme is the landmark hierarchy [19]. Under this scheme, local groupings of network nodes are represented by a single node that serves as a landmark. Communication between different sections of the network are done via these landmarks. Thus the landmarks form a simplified hierarchical scheme. The main difference between such a scheme and a landmark scheme such as Disco is that nodes inside of a group no longer take part in routing outside of the group, but rather use the landmark node for that purpose. Thus the landmark hierarchy scheme is more restrictive, and less capable of finding low-stretch paths. Such a scheme would actually be generalized by the use of routing sets, since as long as a node knows a path to any node in the destination RS, it is able to route to all nodes in that RS.

Mao et al. [14] propose Small State and Small Stretch Routing Protocol (S4), a compact routing protocol for wireless sensor networks. The primary drawback of S4 is that the upper bound on routing state is high.

Westphal and Pei [20] used their greedy embedding scheme to argue that if one can make certain locality assumptions about the topology of the network, then one can devise an efficient routing scheme that works with minimal state memory by forwarding greedily based on distance to the destination. One cannot make topology assumptions about the entire Internet, however it is very reasonable to devise a RS with well-understood topology. It is assumed that all nodes within a RS know how to route between one another, so it is possible for a RS to use greedy embedding when routing to other RS-local addresses.

Scalable Name-based Geometric Routing [21] is a geometric routing scheme intended for *information-centric networking* (ICN). It creates a local vicinity of unspecified size, and uses a greedy routing scheme outside of the local set.

One scheme which is similar to the hinting mechanism is pathlet routing [9]. Pathlet routing is done by storing a set of hop-by-hop partial routes. Full routes to a destination can be formed by combining several pathlets. Pathlets can be exchanged between routers to extend routing range through a broader pathlet vocabulary.

The forgetful routing [12] scheme for BGP observes that routing between AS is less common than routing within an AS, and thus not all addresses outside of an AS need to be remembered at all times. The suggested protocol will forget inter-AS routes after a period of time, and will refresh routes as necessary. The hint based routing scheme also endorses temporary storage of active routes, with expiration after a period of inactivity.

In [10], the effect of routing policies on compact routing is explored. The paper observes that some classes of routing policies have an incompressibility point, at which they cease to be compatible with compact routing schemes. This is an important observation with regards to this paper, because it suggests that a flexible scheme, such as hint-based routing with routing sets, is less likely to interfere with routing policies than one which imposes compact routing over the entire network.

Hint-based routing bears some general similarities to IETF protocols such as *distributed source routing* (DSR) [11], used for routing in mobile networks, and *segment routing* [7], used to provide instructions for steering packets.

4 Protocol

In this section, an implementation of hint routing for IPv6 is discussed. The discussion begins with the assumptions made in implementing the algorithm (Sect. 4.1). Then, the basic protocol is outlined (Sect. 4.2) and the process used to find the reverse route back to the source from the destination (Sect. 4.3). Finally, considerations for network size and security are discussed (Sect. 4.4). Pseudo-code for the algorithm is available in Sect. 4.5.

4.1 Assumptions

This hint based routing protocol is implemented in IPv6. Since an IPv6 address is 16 octets long, with the first 8 octets identifying the network portion, it is

assumed that 8 octets is sufficient to identify a routing node in the network. If a set of multiple routing nodes can be identified by the same 8 octet network portion, it is assumed that they are able to synchronize routing information efficiently so as to make it possible to treat them as a single entity from a routing perspective.

It is assumed that routers are able to identify familiar RS by IP address. It is also assumed that routers are capable of storing hints in order to learn new routes, at least temporarily.

Assume that use of the routing header to carry routing hints will not interfere with other forms of routing services.

A router may keep destinations with the associated hints to reach them in its routing table. If it does so, then it could append those hints to the list of hints in the outgoing packet.

4.2 Basic Protocol

The protocol that will be described amends an IP packet by adding hint information. The IPv6 packet header supports extensions through the use of extension headers. Thus, the set of hints will be carried in the routing extension header. See Table 1.

Table 1. IPv6 routing extension header for the hint-based routing protocol.

Octet	Bit	0	1	2	3
0	0	Next Hdr	Hdr Len	Routing Type	Path Len Est
4	32	Marker Hint Address			
8	64				
12	96	Hint Address 1			
16	128				
20	160	Hint Address 2			
24	192	...			

To this end, an unused routing type should be picked for the *routing type* field. For the purposes of this paper, the value 3 will be used. The extension header begins with four standard fields, the *next header* indicator, the *extension header length*, the *routing type* which was already mentioned, and a *segment length*. The *segment length* field will be used as a hop count estimate for the use of the hinted path. This hop count estimate is to be used only if the information is available, and if it not, then it should remain zero.

After the required fields, what follows is a series of 8 octet long hints. The first hint in the list is a special *stepping stone* hint, which will be explained in Sect. 4.3. All subsequent hints, not including the stepping stone, comprise the normal hint list. Hints closer to the top of the list represent nodes closer to the

destination, while the hints closer to the bottom of the list represent nodes closer to the source.

On receiving a packet with a hint header, routing will proceed as follows. If the router already knows a direct path to the destination, it may ignore the hints altogether, and direct the packet to the destination. If a direct route is not known, attempt to find a path using hints. Begin with the first hint in the normal hint list, and descend through the list until a hint is found for which a direct path is known. If such a hint is found, then the packet may be forwarded to the destination by way of the hinted node.

If none of the hints in the hint list are found, then attempt to identify an RS through which to forward the packet. Starting from the bottom hint in the list, identify its RS if it lies within a familiar RS. Proceed upward through the list until a familiar RS is found. If an RS is found, the packet can be forwarded through the RS, which would guarantee it a path to the hint that was contained in that RS, which would put the packet back on course to the destination.

The reason for checking for hint RS is performed only after failing to find any hints is twofold. First of all, if the hints represent a shortest path or a preferred path, then routing to a different point in a RS may lead to a path which is much longer than necessary. Second and more importantly, suppose that RS S is split by an intermediate set of nodes. If $u, v \in S$, and a packet leaves u to enter the intermediate space en route to v , then routing by RS may cause it to double back and return to u , as the closest member of S . Thus, the RS based routing strategy is held off until all node-based hints are known to have failed. See Fig. 4.

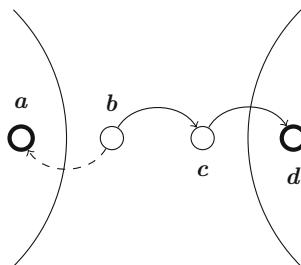


Fig. 4. Suppose that node a is trying to reach node d , which is part of the same (disjoint) RS as a . In order to cross the gap, a adds a hint to c . If b tries to route to d by using the RS of d rather than the hint of c , it may mistakenly return the packet to a due to shortest path considerations.

If no route is found, the router should respond to the source with an ICMP message, *Destination Unreachable, host network unknown*. Routers that see such a message should make note of the failure, and mark saved hinted routes as unreliable, possibly removing them.

If the sender sees such an ICMP message, it knows that the packet was not received. This is advantageous in that it allows a host to attempt sending with

a minimal set of hints. If the packet is not received, then it can send hints in greater detail in future packets. Once it has successfully sent packets, it can reduce the number of hints that it uses.

Whenever a router passes along a message with hints, some of the hints may become obsolete due to representing a node which is closer to the source than the router is. A router should cull obsolete hints from the message, to maintain minimal message size and to facilitate amending the list of hints during transit.

If a router sees a packet with a previously unknown destination or a previously unknown path to the destination, it may record the hinted path for future use, together with an expiration time. For as long as the route is being used, the expiration time will be reset.

One aspect of the protocol which remains to be discussed is how the destination host can find a path back to the source in order to respond to the message, assuming the path was built from hints. The procedure for reverse routing is explained in the following section, Sect. 4.3.

4.3 Reverse Route

Having sent a packet to a destination, it is likely that the destination node will want to reply at some point. If the path to the destination required hints, it is likely that the destination also does not know the full path back to the source. Thus, a mechanism to support reverse path routing is required.

It is possible for every single router along the path to remember the node which it received the message from, and remember the node as a hint path to the source node. This would effectively construct a step-by-step route to the source from the destination. The downside of this approach is that it may involve many more routers and reserve much more state than necessary, and will not necessarily choose the most efficient return path. Furthermore, the reliability of the path depends on the reliability of every single node along the path, since it would not route around a single node failure.

Instead, the return path will use a stepping stone marker algorithm that will be described below. The algorithm still depends on routers on the path to remember the return route via hints, although it attempts to involve only as many routers as are necessary.

The protocol proceeds as follows. Suppose that a source node s is sending to a destination node d . Additionally, a node is designated as the marker node m , with the assumption that $m \rightsquigarrow s$. The node m is stored as the top hint in the list of hints in a packet (Table 1). Initially, m is equal to the source node s .

When a router node n receives a packet from another node n' , if n knows the path to m , it forwards the packet normally, because $n \rightsquigarrow m \rightsquigarrow s$. If it does not know the path to n , then it assumes that $n' \rightsquigarrow m$, so it picks $m' = n'$ as the new value of the stepping stone marker. It replaces m with m' in the outgoing packet, and sends an ICMP message back to n' requesting that it remembers its saved route to source s .

For as long as routers store hinted routes that they have learned, and links can be assumed to be bidirectional, this process guarantees that a path exists

from the destination to the source. Also, since the route is only stored when necessary, the number of routers that are involved is kept minimal.

4.4 Considerations

Hint based routing is devised with the idea that there may be parts of the internet which are unknown to a router. The further two nodes are separated, the longer the list of hints needed to reach one node from the other. The list of hints is not a function of the number of nodes that lie between the source and destination, but of the number of routing sets that one needs to pass through to reach the destination.

Because the size of an IP packet is limited, there is a practical limitation on how far removed two nodes can be while still being reachable. Every hint which is added to a packet increases packet size. Thus, if one needs to reach all possible destinations in a wide network which is not well-connected, one may need to fall back to classical routing techniques to learn more complete routing information.

Due to the similarity of hint routing and the deprecated source routing mechanism [1] built into the IP protocol, a few words need to be said about security. Source routing is considered a security vulnerability due to the fact that it can be exploited to probe and override routing policy, initiate denial of service (DoS) attacks, bypass firewalls and spoof an attacker's source address [3, 4]. The author believes that hint routing is not susceptible to attack in the same way.

Source routing, where implemented and enabled, defines a mandatory path from the source to the destination. Thus, a boomerang packet can be exploited by an attacker in order to probe the internal structure of the network. Unlike source routing, hint routing is not mandatory, and a router may completely ignore the hints if they are not needed. This significantly reduces the effectiveness of the mechanism for probing and overriding network policy.

Source routing can be used in a DoS attack by sending packets which bounce back and forth between a pair of nodes by creating a source address list which includes the both nodes multiple times. In this way, a single packet from the attacker can be turned into multiple packets at the target. Hint routing does not share the same vulnerability, because the list of hints is only used in a minimal way to identify a path to the source. If a path is already known, then the hint list is not used at all, and if it is not known, then the node will only use a sufficient number of hints to identify a path to the destination.

An attacker may use source routing to select a trusted intermediary host in order to bypass a firewall. Hint routing does not explicitly prevent this form of exploit, but if needed, trust relationships may be strengthened by implementing some form of access control which restricts which sources may use which nodes as hints.

Source routing allows an attacker to spoof a trusted host on the target's local network. If the attacker uses the same address as the trusted host, it can pretend to be that host, while relying on the fact that a reverse source list will send responses back to the attacker rather than to the trusted host. Source routing

defeats this practice by routing directly back to the source when possible, and constructing a reverse route only when information is lacking.

A point which is not addressed by the hint routing algorithm is the handling of multicast addressing. A singlecast address can carry a chain of hints to reach a single address. However, this approach does not generalize to multicast addresses, because one may not know all members of a multicast group, and storing hints for multiple members in a single message would be impractical. On the other hand, simply ignoring hints for multicast addresses has the undesirable effect that multicast addresses become purely local in a network where hints are required in order to reach far-away hosts. Thus, multicast addressing is an area where future refinement is required.

A final issue to consider is how hints would be distributed. In cases where the source and destination are familiar with one another, they may exchange hint information beforehand. Alternately, a node may advertise itself on a well-known network center designated for such a purpose. If all parties knew the path to the network center, the advertising node may leave a signed message on a server containing its identity and a hint path to reach it. Other nodes could query the same server to discover the address. The disadvantage of such an approach is that it may tend to centralize network addressing, which was one of the things that the Internet was designed to avoid.

4.5 Algorithm Pseudo-code

The code below shows the actions performed when a hint-enabled packet is received.

```

Receive-and-preprocess(packet)
previous_node  $\leftarrow$  Request-sender()
rheader  $\leftarrow$  Read-routing-header(packet)
dest  $\leftarrow$  rheader.destination
addresses[]  $\leftarrow$  rheader.hints[]
{overwrite stepping stone with destination}
addresses[0]  $\leftarrow$  dest
dst_found  $\leftarrow$  FALSE
route  $\leftarrow$  Hint-lookup(dest, addresses[])
{if route found, send}
if route  $\neq$  NULL then
    step_stone  $\leftarrow$  rheader.hints[0]
    if not dst_found then
        Add-route(route)
    end if
    if Routing-table-lookup(step_stone) = NULL then
        Send-new-step-alert(previous_node, packet)
        step_stone  $\leftarrow$  previous_node
    end if
    addresses[0]  $\leftarrow$  step_stone

```

```

rheader.hints[] ← addresses[]
Update-header(packet, rheader)
Send(packet, route)
end if

```

The routine below performs a lookup to find a route to the destination using hints, by first checking the list of hints in order, and then checking for matching RS in reverse order. The *addresses* array is passed by reference, due to the fact that it may be amended with new hints.

```

function Hint-lookup(dest, addresses[]):
  route ← NULL
  {search hint list for routes}
  for i : 0.. Size(addresses[])–1 do
    entry ← Routing-table-lookup(addresses[i])
    if entry ≠ NULL then
      {check if destination found}
      if i = 0 then
        dst_found ← TRUE
      end if
      Truncate-list(addresses[], i + 1)
      route ← entry.route
      route.destination ← dest
      Append-to-list(entry.hints[], addresses[])
    end if
  end for
  if route = NULL then
    {no hints found, so check masks}
    for i : Size(addresses[])–1..0 do
      entry ← Rt-table-RS-mask-lookup(addresses[i])
      if entry ≠ NULL then
        route ← entry.route
        route.destination ← dest
        Append-to-list(entry.hints[], addresses[])
      end if
    end for
  end if
  return route

```

5 Limitations

The protocol described in this paper has not been tested in a live system, only in a network simulator. Thus, the conclusions presented in this paper should be considered theoretical rather than empirical.

Any gains in terms of routing scalability are dependent on the routing capabilities of the system of routing sets. Thus, an implementation of landmark rout-

ing would imply $\Omega(\sqrt{n})$ routing state for stretch below 5. That is a considerable improvement over linear state, but it would nevertheless remain large for large n .

The presented protocol depends on routers to maintain knowledge of hinted paths, and the reliability and security of the protocol depends on routers taking an active role in judging which hints to follow and which hints to ignore. If a network node knows the path to its destination but an intermediate router does not, then it can only send packets if the router chooses to follow the suggested path.

The protocol assumes that stepping stone nodes know the path back to a source. If there is a network failure, then there may be no way to send back replies or even to send control messages. The practicality of the protocol may depend on having reasonably short hinted paths.

6 Conclusion

There is no reason to expect that the Internet will not continue to grow, and thus it can be expected that routing scalability will continue to be an active area of research for as long as there is need for satisfactory solutions. Landmark based schemes are promising for their effectiveness in addressing scalability, although not all existing implementations are suitable for practical use.

The author feels that migration to a scalable solution is important, since the absence of such a solution would lead to increased stress on Internet infrastructure and threaten the existence of the Internet as a unified entity. However, a sweeping change to the fundamental nature of an established system is always difficult. It meets with resistance, raises questions of authority and control, and presents the non-trivial problem of how to incorporate the portion of machines which are slow to adopt newer networking conventions.

The approach presented in this paper is perhaps more suitable for practical deployment, since it envelops the use of landmark schemes, but at the same time it respects existing network protocols. More importantly, its use can be restricted only to places where it is needed. The author believes that adoption will be far more effective if it can be done without serious disruption, and if allows legacy systems to continue normally. In a way, this is comparable to the use of name-address translation devices, which already perform the task of simplifying Internet addressing while at the same time function without significant change to existing protocols.

The author believes that if the option of hint-based routing is introduced to several major network centers in the near future, then the practicality of the protocol will encourage its adoption among other network clients. Given a suitable rate of adoption, the chosen network centers can be treated as landmarks for landmark routing, which would lead to a gradual decline in the need to store routing information for newly emerging networks. It is assumed that modern technology is capable of handling current routing responsibilities, however if there are spikes in Internet growth (such as Internet-of-things devices), then the deployment of hint-based routing allows a degree of preparation for such a situation.

Future work will involve more rigorous consideration of security implications of a hint-based protocol. Additionally, serious thought will be invested into analyzing minimum-effort ways to ensure backwards compatibility for legacy systems. Most importantly, the protocol will be tried in live systems, because up to this point it has only been tested on network simulators.

Acknowledgements. The author would like to thank Dr. Robert Simon, who was instrumental in encouraging him to push this research through to completion.

References

1. Abley, J., Savola, P., Neville-Neil, G.: Deprecation of type 0 routing headers in ipv6. RFC 5095, Internet Engineering Task Force (2007)
2. Abraham, I., Gavoille, C., Malkhi, D., Nisan, N., Thorup, M.: Compact name-independent routing with minimum stretch. ACM Trans. Algorithms **4**(3), 37:1–37:12 (2008). <https://doi.org/10.1145/1367064.1367077>
3. Bellovin, S.M.: Security problems in the TCP/IP protocol suite. SIGCOMM Comput. Commun. Rev. **19**(2), 32–48 (1989). <https://doi.org/10.1145/378444.378449>
4. Biondi, A., Ebard, P.: Ipv6 routing header security. In: CanSecWest Security Conference (2007)
5. Clark, D.: The design philosophy of the DARPA internet protocols. In: Symposium Proceedings on Communications Architectures and Protocols, SIGCOMM ’88, pp. 106–114. ACM, New York, NY, USA (1988). <https://doi.org/10.1145/52324.52336>
6. Drazic, B., Liebeherr, J.: Improving routing scalability in networks with dynamic substrates. In: Teletraffic Congress (ITC), 2014 26th International, pp. 1–9 (2014). <https://doi.org/10.1109/ITC.2014.6932940>
7. Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., Shakir, R.: Segment routing architecture. RFC 8402, Internet Engineering Task Force (2018)
8. Gavoille, C., Glacet, C., Hanusse, N., Ilcinkas, D.: On the communication complexity of distributed name-independent routing schemes. In: Distributed Computing, pp. 418–432. Springer, Berlin (2013)
9. Godfrey, P.B., Ganichev, I., Shenker, S., Stoica, I.: Pathlet routing. In: Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication, SIGCOMM ’09, pp. 111–122. ACM, New York, NY, USA (2009). <https://doi.org/10.1145/1592568.1592583>
10. Gulyas, A., Retvari, G., Heszberger, Z., Agarwal, R.: On the scalability of routing with policies. IEEE/ACM Trans. Networking **PP**(99), 1–1 (2014). <https://doi.org/10.1109/TNET.2014.2345839>
11. Johnson, D., Hu, Y., Maltz, D.: The dynamic source routing protocol (DSR) for mobile ad hoc networks for IPv4. RFC 4728, Internet Engineering Task Force (2007)
12. Karpilovsky, E., Rexford, J.: Using forgetful routing to control BGP table size. In: Proceedings of the 2006 ACM CoNEXT Conference, CoNEXT ’06, pp. 2:1–2:12. ACM, New York, NY, USA (2006). <https://doi.org/10.1145/1368436.1368439>
13. Kos, J., Aiash, M., Loo, J., Trek, D.: U-sphere: strengthening scalable flat-name routing for decentralized networks. Comput. Netw. **89**, 14–31 (2015). <https://doi.org/10.1016/j.comnet.2015.07.006>
14. Mao, Y., Wang, F., Qiu, L., Lam, S., Smith, J.: S4: small state and small stretch compact routing protocol for large static wireless networks. IEEE/ACM Trans. Netw. **18**(3), 761–774 (2010). <https://doi.org/10.1109/TNET.2010.2046645>

15. Meyer, D., Zhang, L., Fall, K.: Report from the IAB workshop on routing and addressing. RFC 4984, Internet Engineering Task Force (2007)
16. Singla, A., Godfrey, P.B., Fall, K., Iannaccone, G., Ratnasamy, S.: Scalable routing on flat names. In: Proceedings of the 6th International Conference, Co-NEXT '10, pp. 20:1–20:12. ACM, New York, NY, USA (2010). <https://doi.org/10.1145/1921168.1921195>
17. Strowes, S.D., Mooney, G., Perkins, C.: Compact routing on the internet as-graph. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 852–857. IEEE (2011)
18. Thorup, M., Zwick, U.: Compact routing schemes. In: Proceedings of the Thirteenth Annual ACM Symposium on Parallel Algorithms and Architectures, SPAA '01, pp. 1–10. ACM, New York, NY, USA (2001). <https://doi.org/10.1145/378580.378581>
19. Tsuchiya, P.F.: The landmark hierarchy: a new hierarchy for routing in very large networks. In: Symposium Proceedings on Communications Architectures and Protocols, SIGCOMM '88, pp. 35–42. ACM, New York, NY, USA (1988). <https://doi.org/10.1145/52324.52329>
20. Westphal, C., Pei, G.: Scalable routing via greedy embedding. In: INFOCOM 2009, IEEE, pp. 2826–2830. IEEE (2009)
21. Yanbin, S., Yu, Z., Hongli, Z., Binxing, F., Jiantao, S.: An ICN-oriented name-based routing scheme. In: Wang, H., Qi, H., Che, W., Qiu, Z., Kong, L., Han, Z., Lin, J., Lu, Z. (eds.) Intelligent Computation in Big Data Era. Communications in Computer and Information Science, vol. 503, pp. 101–108. Springer, Berlin (2015). https://doi.org/10.1007/978-3-662-46248-5_13



Comparison between Maximal Independent Sets and Maximal Cliques Models to Calculate the Capacity of Multihop Wireless Networks

Maher Heal^(✉) and Jingpeng Li

Department of Computing Science and Mathematics,
University of Stirling, Stirling, UK
{maher.heal,jli}@cs.stir.ac.uk

Abstract. In this work we compare two models to calculate the capacity of multihop wireless networks. The first model utilizes the maximal independent sets of the conflict graph. The problem in that model is formulated as a linear program. The second model in our comparison utilizes the maximal cliques of the conflict graph using integer programming. We see the second model is much more efficient in calculating the capacity for larger networks. We make no assumption on the interference models and we only model it by assuming a conflict matrix. First, we prove there is a periodic schedule for the flow, by using that we formulate our integer programming model to attain maximum capacity for the network. We consider one source of data and one destination i.e. a single commodity network.

Keywords: Maximal independent set · Maximal clique · Linear programming · Binary programming · Maximum throughput

1 Introduction

Wireless multihop networks are networks that have no central entity to coordinate the communication between the network nodes such as wifi networks. The nodes are free to join and leave, and due to the limited wireless range they communicate in a multihop manner, which means that two far nodes may exchange data by forwarding the data to intermediate nodes and the data moves from hop to hop until reaching the destination without any central coordination. There are many realizations for such networks with wide range of applications. Such realizations include wireless mesh networks, wireless sensor networks and ad hoc networks. However, the models we deal with are for static wireless multihop networks, i.e. the nodes are fixed without any motion involved. Hence, they are more appropriate for mesh networks and static sensor networks, unlike ad hoc networks which may have moving nodes.

The capacity of multihop wireless networks has been the subject of intensive study by the research community. Indeed, as it was shown by Jain et al. [5] the general problem of finding the capacity of such networks for a general interference model characterized only by a conflict matrix is np-complete and accordingly no conclusive solution to the problem is possible unless P is equal to NP. Researchers have used information theoretic approaches and linear, integer and mixed-integer programming techniques to address the problem. In this work we propose a maximal cliques binary programming model which is far efficient than the independent sets linear programming model. Our paper is organized as follows: Sect. 1 is the introduction. Section 2 is the literature review shedding light on some of the research carried out on the problem of maximum capacity of multihop wireless networks. Section 3 is a summary of a maximal independent sets model to calculate the capacity, namely Jain et al. [5] model. In Sect. 4 we introduce our integer-programming model to calculate the single-commodity exact capacity of multihop wireless networks. A comparison between the maximal independent sets model and the maximal cliques model is in Sect. 5. Finally, we give our conclusion in Sect. 6.

2 Related Work

The capacity of multihop wireless networks is one of the fundamental questions for such networks. An ultimate answer of the question is not feasible unless P = NP because the problem is NP when interference is factor in the puzzle [5]. There have been two approaches to attack the problem. The first approach is information theoretic one, where bounds on the capacity are derived. The second approach is flow models approach. We will summarize some results of the first approach briefly as our main concern is the flow models approach. In the information theoretic approach, usually assumptions about the topology of the network, randomness and homogeneity of the nodes are assumed and only bounds are derived; while the flow models tend to make no restrictive assumptions apart from the interference models used. The seminal work of Gupta et al. [2] found that for a multihop wireless network of randomly placed identical nodes the throughput of each node is $\Theta(\frac{1}{\sqrt{n} \log n})$ assuming a random communication pattern. If an optimal communication pattern is used then each node throughput is $\Theta(\frac{1}{\sqrt{n}})$. They used two interference models: protocol interference model which is a binary model such that the nodes are either interfering or not based on nodes locations, and a signal-to-noise interference model which they called physical model. In this work we assume no restriction on the interference model, but only modeled by a conflict graph to be explained in Sect. 3. The capacity as derived by Gupta et al. is pessimistic and hence subsequent works searched for alternatives for better bounds. By using percolation theory and assuming pairwise coding and decoding at each hop, and a time-division multiple-access (TDMA) scheme a capacity of $\Theta(\frac{1}{\sqrt{n}})$ was able to be obtained even under random nodes locations assumption [1]. To optimize the bound some authors assumed using directional antennas, such as the work of Yi et al. [17] and

Peraki et al. [12]. Our work can be generalized for such scenarios by changing the conflict graph since our models are general for any interference models. A gain in the capacity is also possible by using multi-packet reception (MPR) as proved in [13]. However, in the models we compare, we made no such assumption, in spite of there are some flow models for the capacity studied the MPR scenario [16]. Some authors studied the effect of topology on the network capacity such as [4]. We are studying mainly lattice topologies and random topologies. The impact of traffic pattern was also a subject of studies by considering multicast and broadcast traffic and not only a unicast [6, 9, 10, 14]. we deal only with a unicast traffic, but extensions are possible for other kinds of traffic. A good survey paper of the information theoretic approach in calculating bounds on the capacity of multihop wireless networks is that by Lu et al. [11].

The other methods that were used to study the capacity are flow models. The first flow model that sparked off a whole research direction using these techniques to calculate the capacity of multihop wireless networks is that of Jain et al. [5]. We will summarize that model in Sect. 3 and we will use it as our base model for maximal independent sets models that calculate the capacity after listing maximal independent sets of the conflict graph. Although the authors discussed two interference models, i.e. the protocol interference model and the physical interference model, similar to those in [2], their model is quite general to any interference model since it is modeled by the adjacency matrix of a conflict graph. Kumar et al. [8] studied the problem of maximum capacity under different constraints, namely fairness and energy consumptions. However, their model is based on the geometric properties of three interference models, one of which is the protocol model. Their model is not applicable to the general case of interference characterized by a general conflict matrix. In [7] the authors suggested an algorithm that provides 68% of the optimal throughput in worst scenarios and up to 80% practically. However, their interference model is very limited by considering nodes that can transmit to and receive from one node at a time. They also suggested an extension to a limited version of IEEE 802.11 like interference protocol without specifying how close their found throughput to the optimal value. Here, we are interested in exact throughputs or network capacity. Some authors studied directional antennas and reconfigurable antennas such as [3]. Although we don't refer to that, we assume a general conflict graph which can accommodate for such scenarios. Moreover, the maximum throughput problem was studied under physical interference model as in [15].

3 Maximal Independent Sets Models for Capacity Calculation

We assume we have a network modeled by a graph of N vertices and L links, $G(N, L)$. The vertices represent the nodes and the links represent the communication channels between the nodes. Interference is modeled by a conflict graph H where each vertex in the graph corresponds to a link in the network graph, two vertices in the conflict graph are connected if the links in the network graph

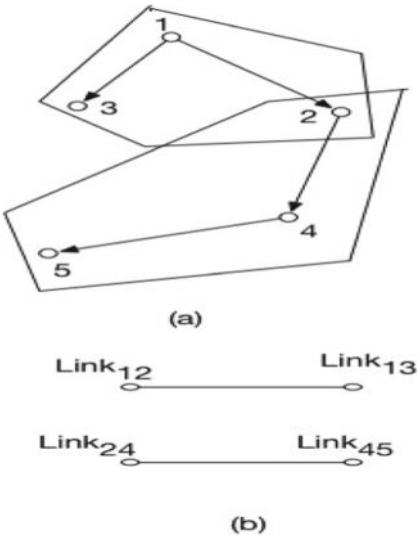


Fig. 1. Conflict graph. **a** A 5 node network with their interference zones, and **b** the conflict graph of the network.

are interfering, i.e. cannot be active at the same time. See Fig. 1 which shows a 5 node network with the interference zones. $Link_{12}$ is in the same interference zone of $Link_{13}$ and $Link_{24}$ is in the same interference zone of $Link_{45}$ hence we see them connected by an edge in the conflict graph in Fig. 1b. By assuming there is one flow from node n_s to node n_d , the maximum flow problem is given by:

$$\max \sum_{l_{si} \in L} f_{si} \quad (1)$$

Subject to:

$$\sum_{l_{ij} \in L} f_{ij} = \sum_{l_{ji} \in L} f_{ji} \quad n_i \in N \setminus \{n_s, n_d\} \quad (2)$$

$$\sum_{l_{is} \in L} f_{is} = 0 \quad (3)$$

$$\sum_{l_{di} \in L} f_{di} = 0 \quad (4)$$

$$f_{ij} \leq C_{ij} \quad \forall i, j | l_{ij} \in L \quad (5)$$

$$f_{ij} \geq 0 \quad \forall i, j | l_{ij} \in L \quad (6)$$

$$\sum_{i=1}^{K'} \lambda_i \leq 1 \quad (\text{because only one maximal independent set can be active at a time}) \quad (7)$$

$$f_{ij} \leq \sum_{l_{ij} \in I_i} \lambda_i \cdot C_{ij} \quad (8)$$

(because the fraction of time for which a link may be active is constrained by the sum of activity periods of the independent sets it is a member of)

where $\lambda_i \geq 0$ is the time allocated to maximal independent set I_i , K' is the total number of maximal independent sets and C_{ij} is the capacity of link ij . The objective, Eq. 1, is to maximize the outward flow from the source node n_s . Equation 2 is the flow conservation condition which means the inward flow equals to the outward flow for all nodes except the source or destination. The third Eq. 3 states that the inward flow at the source node is zero; and similarly Eq. 4 states that the outward flow from the destination node is zero. Equation 5 is a restriction on each link flow to be less than the link capacity, and Eq. 6 obviously states each flow is either positive or zero. This is a single commodity formulation since we have a single source - destination flow. Equations 7 and 8 are constraints due to interference. Please refer to [5] for details.

4 Maximal Cliques Models for Capacity Calculation

This section states our integer programming model to calculate the capacity of wireless multihop networks. We firstly state some opening definitions and prove there is a periodic schedule that attains the maximum capacity for the network which is crucial for our model correctness.

4.1 Preliminary Definitions

As before, we assume a wireless multihop network of N nodes and L links. The links are interfering according to any interference model, which is modeled by a conflict graph characterized by a conflict matrix (graph adjacency matrix). C is a column vector of links capacities. The network is a single commodity network with one source n_s and one destination n_d .

A feasible schedule of a link: it is a set of successive time periods such that in each time period the link is either active (transmitting data) or idle (not transmitting data). However when the link is active in a period, all other interfering links are idle.

A feasible schedule of the network: it is a schedule where all links schedules are feasible on the same time scale and the flow conversation rules are satisfied.

Maximum flow of the network: it is the maximum flow from n_s to n_d such that the network schedule is feasible. See Fig. 2 for illustration of these definitions.

Flow of a link. Let $\sum_0^t x^i$ is the sum of successful active time on link i in the period $[0, t]$ when schedule x is used, flow of link i (f_i) is defined as: $f_i = \lim_{t \rightarrow +\infty} \frac{\sum_0^t x^i}{t}$.

Feasible flow vector of the network is an assignment of flows $(f_i), i = 1, 2, \dots, l$ $i \in L$ where the schedule of the network is feasible.

4.2 Proof of the Existence of a Periodic Schedule

We prove here that there is always a periodic schedule that attains a maximum flow for the network from node n_s to node n_d .

Lemma 1. Let $g_i = \frac{\sum_{t_1}^{t_2} x^i}{t_2 - t_1}$ be the average of the sum of active periods on link i in the period t_1 to t_2 when feasible schedule x is used, and let (f_i^*) be the maximum flow of the link i , $i = 1, 2, \dots, l$ when the network flow is maximum, then g_i is less than or equal to (f_i^*) for any feasible schedule x and link $i = 1, 2, \dots, l$.

Proof. Let $f_i^* = \lim_{t \rightarrow +\infty} \frac{\sum_0^t x^i}{t}$ be the maximum link i flow when the network flow is maximum.

Now if $g_i > f_i^*$ then divide the time line into slots of size $t_2 - t_1$ and use schedule x in each of these slots, we have $f_i = \lim_{t \rightarrow +\infty} \frac{\sum_0^t x^i}{t} = \lim_{n \rightarrow +\infty} \frac{\sum_{j=1}^n g_i(t_2 - t_1)}{n(t_2 - t_1)} = \lim_{n \rightarrow +\infty} \frac{n g_i(t_2 - t_1)}{n(t_2 - t_1)} = g_i$. If $g_i > f_i^*$, we have a flow greater than f_i^* , which is clearly a contradiction since f_i^* is the maximum attainable flow. Accordingly $g_i \leq f_i^*$.

Theorem. There is always a periodic schedule to maximize the network flow in single commodity or multicommodity wireless multihop networks.

Proof. Let $f_i^* = \lim_{t \rightarrow +\infty} \frac{\sum_0^t x^i}{t}$ be the maximum feasible flow for link $i = 1, 2, \dots, l$. Now divide the time line into slots of size T , i.e. $[0, T], [T, 2T], [2T, 3T], \dots$ etc. we have $f_i^* = \lim_{n \rightarrow \infty} \frac{\sum_{j=1}^n \sum_{j=1}^{jT} x^i}{nT}$. Now if the schedule is periodic in T that is all what we need and $f_i^* = \frac{\sum_0^T x^i}{T}$. In case it is not periodic then based on the Lemma 1, we have the average flow in every T equals or less than f_i^* . Hence either $\sum_0^T x^i = \sum_T^{2T} x^i = \sum_{2T}^{3T} x^i = \dots = f_i^* T$ then replace the schedule of $[T, 2T], [2T, 3T], \dots$ by the schedule of $[0, T]$ and by that we have a periodic schedule; or if we have $\sum_0^T x^i, \sum_T^{2T} x^i, \sum_{2T}^{3T} x^i \dots$ all or some less than $f_i^* T$ then we have $\sum_{j=1}^n \sum_{j=1}^{jT} x^i < nT f_i^*$ dividing by nT and taking the limit as n tends to infinity we have $f_i^* < f_i^*$ which is clearly a contradiction. Accordingly, the schedule is periodic.

A remark on the period T It is clear T can be arbitrary as can be seen from the previous proof. For example if we take $T = 10$ time units, we can extend the time scale by 2 or shrink by 0.5 and the schedule used is extended or shrunken proportionally. See Fig. 3 for an example.

4.3 Integer Programming Model

Taking the period equals to 1, we can easily have maximum network flow is given by the solution of the following integer programming problem given that

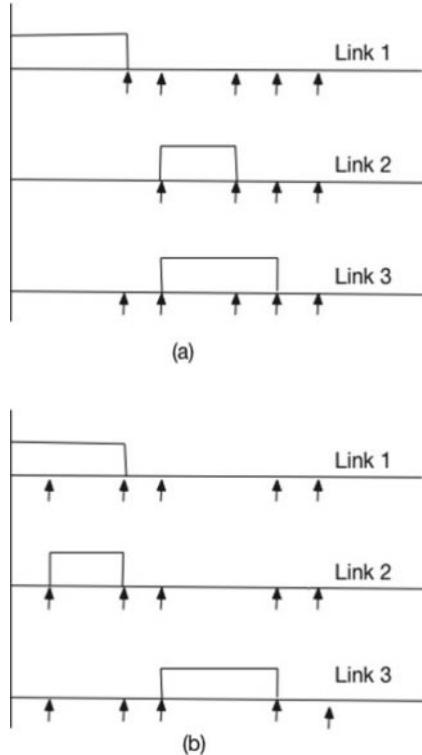


Fig. 2. Link 1 is interfering with link 2, but link 2 is not interfering with link 3. **a** A feasible schedule of 5 periods, and **b** infeasible schedule.

the period is divided into n equal slots and after dividing the solution by n and taking n tends to infinity.

$$\max_{l_{n_s i} \in L} \sum_{r=1}^n C_{n_s i} \theta_{n_s i}^r \quad (9)$$

$$\sum_{l_{ij} \in L} \sum_{r=1}^n C_{ij} \theta_{ij}^r = \sum_{l_{ji} \in L} \sum_{r=1}^n C_{ji} \theta_{ji}^r \quad (10)$$

$$\sum_{l_{in_s} \in L} \sum_{r=1}^n C_{in_s} \theta_{in_s}^r = 0 \quad (11)$$

$$\sum_{n_d i \in L} \sum_{r=1}^n C_{n_d i} \theta_{n_d i}^r = 0 \quad (12)$$

and at each maximal clique q

$$\sum_{l_{ij} \in q} \theta_{ij}^r \leq 1 \quad r = 1, 2, \dots, n \quad (13)$$

$$\theta_{ij}^r \in 0, 1 \quad (14)$$

where θ_{ij}^r is the time allocated in slot r for link ij , $r = 1, 2, 3, \dots, n$. The first equation is maximizing the outward flow from source node n_s and Eqs. 10, 11 and 12 are the flow conservation equations and Eq. 13 is a restriction on θ variables, allocated time, in order to have a feasible schedule free of conflicts. We illustrate that by a sample network of five links as shown in the Fig. 4. It is true we need large value of the slots number to confirm converging to the maximum flow of the network, but we can try smaller number of the slots starting by 1, 2, 3, 4, ... etc. until we hit the period of the network as we will see for many networks. This will be clear when we discuss the results in Sect. 5, and when we apply the algorithm in Procedure 1 to the network in Fig. 4 at end of this section. Indeed the calculated capacity for whatever number of slots, by Lemma 1, is less than the calculated capacity when the number of slots is the period of the schedule. Additionally when we double the period we have again the maximum attained flow. Hence we have the algorithm in Procedure 1.

In Table 1 we see the obtained throughput for different values of slots for the network in Fig. 4 when we use our integer programming model, for $n = 1-10$. It can be seen that the throughput is less than 0.4 in all values of slots expect at $n=5$ and $n=10$. Hence it is concluded the period is 5 and the maximum throughput is 0.4. To check we calculate the throughput for n assuming large values, for example when $n=99$ we found the throughput is equal to 0.3939

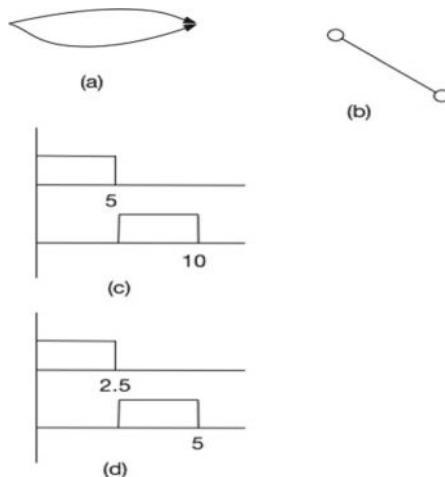


Fig. 3. A schedule that attains maximum flow for two nodes network. **a** The network, **b** the conflict graph, **c** schedule with period equal 10 time units, and **d** the period in c shrunken by factor of 2.

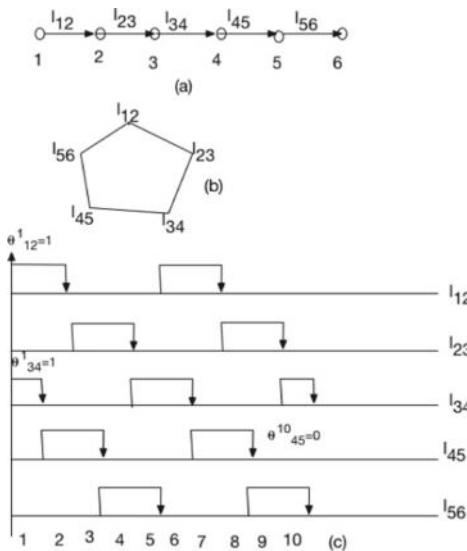


Fig. 4. Example Network. **a** The network topology, $n_s = 1$ and $n_d = 6$, **b** the conflict graph, and **c** a feasible schedule that attain maximum flow of 10 slots, with θ variables shown for some slots. The period is 5.

Table 1. Integer programming model throughput for different number of slots n

n	Throughput
1	0
2	0
3	0.333
4	0.25
5	0.4
6	0.333
7	0.2857
8	0.3750
9	0.333
10	0.4
99	0.3939
100	0.4

which is close to 0.4. Indeed when n tends to infinity we get a throughput equals to 0.4. When we take $n = 100$, the obtained throughput is 0.4 since 100 is a multiple of 5, i.e. we are repeating the period more than one time.

Procedure 1 Integer programming algorithm to calculate capacity

```

1: numberofslots :  $n \leftarrow 1, 2, 3, 4, 5, 6, 7, \dots$ 
2: if calculated flow changes and reaches maximum at  $M$  slots then
3:   check flow at slots number  $2M$  and less and more than  $2M$ 
4:   if flow is maximum at  $2M$  and less at number of slots less and more than  $2M$ 
      then
6:     the period is  $M$  and maximum capacity is flow at  $M$  or  $2M$ 
7:   else
8:     Keep trying for increasing value of number of slots
9:   end if
9: end if

```

5 Results

We run both the maximal independent sets model and the maximal cliques models using MacBook Pro, late 2012, 2.5 GHz Intel Core 5 processor and 8 GB RAM. The networks we run the models on to calculate the capacity are lattice networks with 802.11 MAC protocol, i.e. the interference is at the transmitter and the receiver of the packet and with one source that is laying at the lower left corner and the destination is at the upper right corner (see Fig. 5). Transmission range in all networks (d) is 1 and interference range (R) is the same, with a capacity of each link (C) equals to 1. In Table 2, m is the length of the side so $m = 32$ means 1024 nodes, ISMT1, ISMT2 and ISMT3 are the maximal independent sets listing time, linear solver time and total time respectively for the maximal independent sets model. S, CMT1, CMT2, CMT3 and T are the cliques model number of slots used, cliques listing time, binary solver time, total

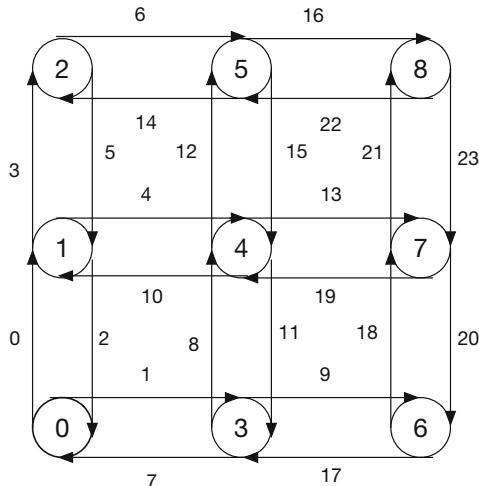


Fig. 5. Lattice networks of 9 nodes, $m = 3$

Table 2. Lattice networks of different sizes, $d = 1$, $R = 1$, $C = 1$

m	ISMT1	ISMT2	ISMT3	S	CMT1	CMT2	CMT3	T
3	0.0015	6.6887e-4	0.0024	1-8	6.796e-4	0.0103	0.0772	0.5
5	6.5969	0.0071	6.7217	1-12	0.0054	0.239	0.2462	0.6667
7	22 * 60	-	-	1	0.0201	0.005	0.0253	0
-	-	-	-	2	0.0148	0.0014	0.0162	0
-	-	-	-	3	0.0151	0.0025	0.0177	0.6667
-	-	-	-	4	0.0147	0.034	0.0183	0.5
-	-	-	-	5	0.0183	0.0109	0.0257	0.4
-	-	-	-	6	0.0141	0.0059	0.0201	0.6667
						Total	0.1233	0.6667
				100		0.1168	0.1374	0.6600
23	-	-	-	1-6	0.7834	85.082	88.9233	0.6667
-	-	-	-	100		4.5041	5.452	0.66
32	-	-	-	1	4.336	0.0108	4.3473	0
-	-	-	-	2	4.0076	0.04	4.055	0
-	-	-	-	4	4.0097	0.2587	4.2806	0.5
-	-	-	-	6	4.1316	0.3451	4.5193	0.6667
-	-	-	-	12	4.2209	0.707	4.9927	0.6667
						Total	22.195	0.6667
-	-	-	-	105	4.1301	9.3944	13.8378	0.6667

time and calculated throughput, respectively. All times are in minutes. As can be seen from the table the independent sets model can calculate the throughput when the number of nodes is maximum 25, $m = 5$. It completely fails when we increase the nodes for 49, 529 and 1024. This failure is due to the excessive time needed to list independent sets as can be seen when $m = 7$; after 22 hours of running the complete set of independent sets is still not complete. The clique model outperforms the independent set model due to the very short time in listing cliques and the bottleneck is the binary solver time; however, it is quite reasonable and when the solver takes excessive time for a slot number you may try a different slot number or tweak the binary solver. The solver we used is cplex 12.7.1 for matlab. In Table 2 we reported the time for some values of m in an aggregated manner due to space such as $m = 3$, but detailed for other values such as $m = 7$. The periods found for $m = 3, 5, 7, 23$ and 32 are 5, 6, 3, 3 and 3 respectively. Even when estimating the period is hard, the calculated throughput is quite close to the exact value when S is large such as S = 100 for $m = 23$ in a fairly short time. Our last example is a random network of 42 nodes and 188 links (see Fig. 6). The protocol used is 802.11 and hence interference is at transmitter and receiver. It is deployed in an area of 5×5 m and transmission and interference ranges are both 1m. capacity is 1 for each link. The source is

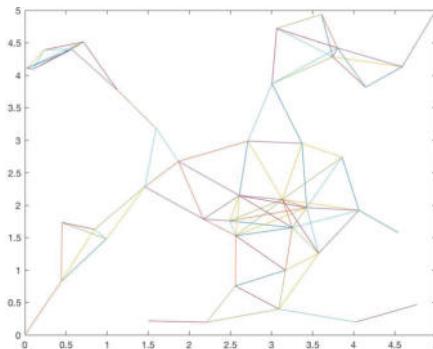


Fig. 6. Random network of 42 nodes and 188 links

the node at the lower left corner and the destination at the upper right corner. After running the maximal independent sets model for two hours we don't see a convergence to all maximal independent sets and hence we couldn't calculate the exact capacity. The time required is expected to be much more than 2 hours. With our clique model we found a period of 3 and were able to find an exact throughput of 0.3333 in 0.0626 min using slots from 1 to 6 and a throughput of 0.33 in 0.0573 min using 100 slots.

6 Conclusion and Future Work

We compared two models to calculate the maximum single commodity throughput in multihop wireless networks. The maximal independent sets model gives exact throughput but only for very small networks due to the excessive time required to list maximal independent sets; while the maximal cliques model calculates the exact throughput for far larger networks due to smaller time in listing maximal cliques. By taking a large number of slots, the maximal cliques model gives results very close to the exact value. We are planning to estimate the difference from the exact value for a large number of slots, by considering the estimated values of throughput for small values of slots when the period cannot be guessed.

References

- Franceschetti, M., Dousse, O., David, N., Thiran, P.: Closing the gap in the capacity of wireless networks via percolation theory. *IEEE Trans. Inf. Theory* **53**(3), 1009–1018 (2007)
- Gupta, P., Kumar, P.R.: The capacity of wireless networks. *IEEE Trans. Inf. theory* **46**(2), 388–404 (2000)
- Hou, Y., Li, M., Zeng, K.: Throughput optimization in multi-hop wireless networks with reconfigurable antennas. In: 2017 International Conference on Computing, Networking and Communications (ICNC), pp. 620–626. IEEE (2017)

4. Hu, C., Wang, X., Yang, Z., Zhang, J., Xu, Y., Gao, X.: A geometry study on the capacity of wireless networks via percolation. *IEEE Trans. Commun.* **58**(10), 2916–2925 (2010)
5. Jain, K., Padhye, J., Padmanabhan, V.N., Qiu, L.: Impact of interference on multi-hop wireless network performance. *Wirel. Netw.* **11**(4), 471–487 (2005)
6. Keshavarz-Haddad, A., Ribeiro, V., Riedi, R.: Broadcast capacity in multihop wireless networks. In: Proceedings of the 12th Annual International Conference on Mobile Computing and Networking, pp. 239–250. ACM (2006)
7. Kodialam, M., Nandagopal, T.: Characterizing achievable rates in multi-hop wireless networks: the joint routing and scheduling problem. In: Proceedings of the 9th Annual International Conference on Mobile Computing and Networking, pp. 42–54. ACM (2003)
8. Kumar, V., Marathe, M.V., Parthasarathy, S., Srinivasan, A.: Algorithmic aspects of capacity in wireless networks. In: ACM SIGMETRICS Performance Evaluation Review, vol. 33, pp. 133–144. ACM (2005)
9. Li, X.Y.: Multicast capacity of wireless ad hoc networks. *IEEE/ACM Trans. Networking (TON)* **17**(3), 950–961 (2009)
10. Li, X.Y., Zhao, J., Wu, Y.W., Tang, S.J., Xu, X.H., Mao, X.F.: Broadcast capacity for wireless ad hoc networks. In: 5th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, 2008. MASS 2008, pp. 114–123. IEEE (2008)
11. Lu, N., Shen, X.S.: Scaling laws for throughput capacity and delay in wireless networks—a survey. *IEEE Commun. Surv. Tutorials* **16**(2), 642–657 (2014)
12. Peraki, C., Servetto, S.D.: On the maximum stable throughput problem in random networks with directional antennas. In: Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking & Computing, pp. 76–87. ACM (2003)
13. Sadjadpour, H.R., Wang, Z., et al.: The capacity of wireless ad hoc networks with multi-packet reception. *IEEE Trans. Commun.* **58**(2) (2010)
14. Shakkottai, S., Liu, X., Srikant, R.: The multicast capacity of large multihop wireless networks. *IEEE/ACM Trans. Networking (TON)* **18**(6), 1691–1700 (2010)
15. Wan, P.J., Frieder, O., Jia, X., Yao, F., Xu, X., Tang, S.: Wireless Link Scheduling Under Physical Interference Model. IEEE (2011)
16. Wang, Z., Sadjadpour, H., Garcia-Luna-Aceves, J.J.: The capacity and energy efficiency of wireless ad hoc networks with multi-packet reception. In: Proceedings of the 9th ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 179–188. ACM (2008)
17. Yi, S., Pei, Y., Kalyanaraman, S.: On the capacity improvement of ad hoc wireless networks using directional antennas. In: Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking & Computing, pp. 108–116. ACM (2003)



A Priority and QoS-Aware Scheduler for LTE Based Public Safety Networks

Mahir Ayhan¹(✉) and Hyeong-Ah Choi²

¹ Department of Computer Science, American University,
Washington DC, USA
mayhan@american.edu

² Department of Computer Science,
George Washington University, Washington DC, USA
hchoi@gwu.edu

Abstract. 4G Long-Term Evolution (LTE) has been selected by the U.S. federal and EU authorities to be the access technology for public safety broadband networks (PSBNs) that would allow first responders to seamlessly communicate between agencies nationwide. From Release 11 on, 3rd Generation Partnership Project (3GPP) has been outlining the standards for the features that will enable LTE to be used as part of a PSBN. The requirements for scheduling user equipments (UEs) with appropriate quality of service (QoS) and priority has not been addressed yet. In this paper, we highlight the scheduling challenges in PSBNs and propose a solution which considers all the QoS parameters. Proposed algorithm minimizes packet losses while scheduling packets in a priority and QoS aware fashion. Simulations results illustrate superiority of the proposed scheduler.

Keywords: Scheduling · Public Safety Broadband Network · LTE · Priority · QoS · OFDMA · High throughput

1 Introduction

On Sept. 11, 2001, firefighters and police officers could not communicate to each other on their radios at the World Trade Center. Same problem was seen in 2005 Hurricanes Katrina and Rita. Public safety (PS) officers from different jurisdictions arrived at the scene only to find that they were unable to communicate with each other by radio [1]. Traditional Land Mobile Radio (LMR) communication networks (e.g. P25, and TETRA) no longer meet the needs of public safety agencies be it police, fire or EMS. Public safety agencies need wireless networks that provide more than push-to-talk. Some public safety wireless networks today are limited to voice communications only and are not capable of transmitting multimedia data . Besides being reliable, scalable, and secure, public safety networks have to provide high quality of service (QoS), so that it can meet first responders' (FRs) needs at all times. Long-Term Evolution (LTE) satisfies the growing need for broadband in public safety [2].

LTE will also provide an unprecedented opportunity for interoperability—even on a nationwide scale—which has always been an issue since most agencies use their own private radio systems, therefore they typically do not connect to other networks. To overcome the problems LMR is facing and to benefit from the advantages of the LTE, the Federal Communications Commission (FCC) has announced LTE to be the access network technology for the National Public Safety Broadband Network (NPSBN) [3]. This comes with a number of unprecedented challenges for the NPSBN. When disasters strike whether they are man-made such as terrorist attacks or natural disasters such as hurricanes, communication networks get congested. This saturation may result in a heavy load at a given cell, and it may be so severe that a responder is prevented from accessing the cell or receiving the QoS his/her applications require [3]. In order to prevent such situations, prioritization and QoS, both in access and core network, must be taken care of. In this paper, we provide a solution for prioritization and QoS in the access network.

Public Safety Broadband Networks (PSBN) are private networks with very strict performance constraints. Scheduling algorithms developed for commercial LTE networks mainly focus on increasing throughput, enhancing QoS, and providing fairness. Besides these requirements, PSBN needs to differentiate the bearers based on their priorities. To our knowledge, no scheduler has addressed priority along with other requirements yet. The contributions of this paper can be summarized as follows:

- Potential problems associated with scheduling packets in the PSBN are identified.
- An algorithm that meets the priority and QoS requirements of PSBN is proposed.
- The performance of the proposed algorithm is evaluated and its general validity is demonstrated by testing it under different network conditions.

The rest of the paper is organized as follows. Section 2 presents the related works. Section 3 provides the background information of the problem. Section 4 describes the problem formulation. Section 5 discusses the proposed algorithm. Simulation parameters, results and discussion are presented in Sect. 6. Finally, Section 7 concludes the paper.

2 Related Works

There are five scheduling strategies: (i) channel-unaware; (ii) channel-aware/QoS-unaware; (iii) channel and QoS-aware; (iv) semi-persistent for VoIP support; and (v) energy-aware [4]. Schedulers that are designed for LTE may focus on one or combine more of these strategies. RR is the most known channel-unaware scheduler. Resources are allocated to bearers in order. Therefore, it provides fairness in terms resource distribution, yet it is not an efficient scheduler in terms of throughput since it does not consider channel quality.

Works in [5–8] represent channel-aware/QoS-unaware schedulers. Best Channel Quality Indicator (BestCQI) [6] favors UEs with the largest CQI values to maximize system throughput; however, it starves UEs with unfavorable channel conditions [5]. Proportional Fair (PF) Scheduler [7] selects bearers with highest current throughput over average transmitted throughput. PF delivers reasonable overall cell throughput as well as fairness between bearers. Pokhariyal et al. [8] proposed a two step scheduler. In the first step, Time Domain Scheduler (TDS) selects a subset of active bearers in the current TTI; in the second step, Frequency Domain Scheduler (FDS) allocates RBs to selected bearers using PF or RR criteria. Channel-aware/QoS-unaware schedulers exploit radio resource management (RRM) features such as CQI and link adaption etc, to achieve a certain spectral efficiency; yet, they can not guarantee the QoS.

In [9–13] authors proposed schedulers that are both Channel-aware and QoS-aware. QoS is managed by QoS parameters to guarantee data rate, packet delay, or loss. A Guaranteed Data Rate (GDR) approach is proposed in [11, 13]. In [11] the Time Domain Priority Set Scheduler (TD-PSS) divides bearers into two sets. UEs below Target Bit Rate (TBR) comprise Set 1 and are given high priority. Bearers in the set-1 scheduled first, and if there are still RBs, bearers in the set-2 are selected according to PF criteria.

Schedulers that focus on delay are proposed in [12, 14]. Sandrasegaran et al. [12] proposed Delay Prioritized Scheduler (DPS) that selects UEs based on packets' head-of-line (HOL) proximity to delay threshold. Bearers whose packets are about to drop, are scheduled first. DPS achieves low packet loss rate. Bojovic et al. in [14] proposed an algorithm called Channel and QoS Aware (CQA) scheduler. The CQA scheduler is also based on decoupled time and frequency schedulers. In the TD, at each TTI, the scheduler groups bearers that have not met their target data rate based on HOLs. In FD, starting from most urgent flows, it allocates RBs to bearers while favoring GBR bearers over NGBR within each group. Fan et al. [15] proposed a semi-persistent scheduler to support maximum number of VoIP. The idea behind semi-persistent schedulers is that, radio resources are divided into several groups and each block is associated to a set of bearers. This approach minimizes signaling overhead by pre-configuring the bearers.

3 Scheduling Priority

LTE transports data between the UE and the Packet Data Network (PDN) Gateway (P-GW) using bearers, which can be considered as bi-directional data pipe with a specific requirements [16]. Bearers have a corresponding QoS description which should influence the behavior of the evolved Node-B (eNB) resource scheduling algorithm.

The 3GPP has studied the needs of applications on LTE and identified their attributes in [17]. Standardized combinations of these characteristics are called QoS Class identifiers (QCIs). It is a scalar that acts as a pointer into a look-up table (see Table 1) which describes four other quantities—bearer type, priority,

packet delay budget (PDB) and packet error loss rate (PELR). There are two types of bearers; bearers with a minimum guaranteed bit rate (GBR) are called GBR bearers and others are non-GBR (NGBR). The PELR is an upper bound for a rate of non-congestion related packets losses. The PDB defines an upper bound, with 98% confidence, for the time that a packet may be delayed between the UE and the P-GW. Priority determines when packets should be sent to or received from the UEs [3] and is handled in the access network by eNB. High priorities are associated with low numbers. A QCI assigned to a bearer may or may not be changed afterward.

Table 1. QCI table

QCI	Bearer type	Priority	PDB (ms)	PELR
1	GBR	2	100	10^{-2}
2		4	150	10^{-3}
3		3	50	10^{-3}
4		5	300	10^{-6}
65		0.7	75	10^{-2}
66		2	100	10^{-2}
5	NGBR	1	100	10^{-6}
6		6	300	10^{-6}
7		7	100	10^{-3}
8		8	300	10^{-6}
9		9	300	10^{-6}
69		0.5	60	10^{-6}
70		5.5	200	10^{-6}

Currently, PS agencies use different technologies for over-the-air Mission Critical Push-to-Talk (MCPTT) and data services to maintain a distinction between applications. MCPTT is provided with a pool of guaranteed resources by reserving certain bandwidth. With LTE, data, voice, bandwidth-intensive video and multimedia services and all other applications will share same resources, so this distinction will be removed. From the PS perspective, it is important to discern the most important applications [3].

According to the *UE Priority Model* which has been developed by National Public Safety Telecommunications Council's (NPSTC) Priority and QoS Working Group, one of the attributes that is closely related to the scheduling is *Type of Application* (see Fig. 1) [3]. The *Type of Application* feature is intended to demonstrate the relative default significance of all applications on the PSBN. The purpose is that mission critical applications (MCAs) receive elevated priority on the system. As congestion rises at a given cell, higher priority applications will receive more resources by default. Higher priority applications should be linked

to GBR bearers by core network so that when congestion arises in the cell, these applications can be guaranteed with proper QoS. Assigning such applications to GBR bearers may still not guarantee good service or continuity if priority is not considered. MCPTT is selected as the most important applications [3]. Thus, this application should always have the highest priority. In addition, in case of Responder Emergency situation (Fig. 1), where responder him/herself is in a life threatening situation, or Immediate Peril (IP), regardless of application type, any applications used under RE and IP situation must be assigned to highest priority bearer and even in the case of saturated network, every effort must be made to maintain these applications. The process of assigning application to bearers is beyond the scope of this work. Yet, scheduler in the air interface should consider the priorities of bearers when allocating RBs in order to assure proper QoS for MCAs.

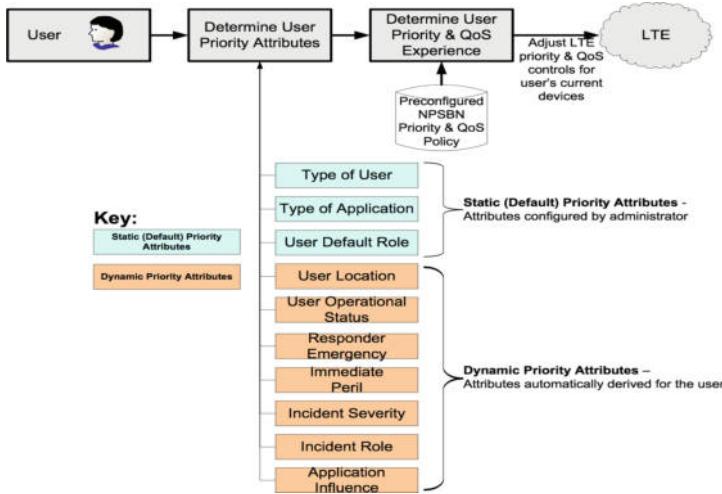


Fig. 1. UE prioritization model [3].

4 Problem Formulation

Table 2 shows the important notations used throughout the paper. There is a single cell and the total number of RBs available at each 1 s window is n . m is the number of bearers in the system. The constraints are given as:

- (1) The total number of RBs needed to satisfy GBR bearers' bit rate requirements and NGBR bearers' demands, which are calculated based on sizes and Inter-arrival times (IATs) of packets and UEs' Modulation and Coding Schemes (MCSs), at any moment, should be less than or equal to n

$$\sum \alpha(i) \leq n$$

Table 2. Notations

Notation	Definition
Bearer i	Bearer with QoS Class index of i
$BII(i)$	Importance index of bearer i
$PELR(i)$	PELR of bearer i
$PDB(i)$	PDB of bearer i
$GBR(i)$	GBR of bearer i
$HOL(i)$	Head of Line of bearer i
$RT(i)$	Remaining Time of bearer i
$\alpha(i)$	Number of RBs needed to satisfy bearer i 's GBR (demand in case of NGBR) within 1 sec window
$\theta(i, t)$	Throughput of bearer i at Transmission Time Interval (TTI) t
$\beta(i, t)$	Loss rate bearer i at TTI t
UE_j^i	UE j who employs bearer i
IAT	Inter-arrival time
MCS	Modulation and Coding Scheme

- (2) If (1) holds true, for each GBR bearer i , at any Transmission Time Interval (TTI) t , if throughput of the bearer is less than promised GBR than packet loss rate of the bearer must be less than or equal to threshold

$$\beta(i, t) \leq PELR(i), \text{ if } \theta(i, t) \leq GBR(i)$$

The objective is to minimize (3) while satisfying (1) and (2).

- (3) $\Sigma\{\beta(i, t) | (1 \leq i \leq m) \text{ and } 1 < t < \infty\}$

Simply put, the objective is to minimize the packet losses in the system while satisfying GBR and PELR requirements in importance order. We also assume that once throughput requirement for a GBR bearers is met, packet losses are allowed for that bearer.

5 The Design of Proposed Algorithm

The access network may use the QCI parameters to manage packet forwarding treatment because the goal of standardizing a QCI with corresponding parameters is to ensure that bearers mapped to that QCI receive the same minimum level of QoS in multi-vendor network deployments and in case of roaming [17]. We map every QCI to an appropriate Bearer Importance Index (BII) as shown in Table 3. The intention is to make sure that GBR bearers have higher importance than NGBR while the relative priority among GBR and NGBR bearers are preserved. These values suffice that. High numbers are associated with high importance. We proposed Priority and QoS Aware (PQA) scheduler (Algorithm

Table 3. QCI-BII mapping table

QCI	Bearer type	Priority	BII
1	GBR	2	9
2		4	7
3		3	8
4		5	6
65		0.7	10.3
66		2	9
5	NGBR	1	5
6		6	4
7		7	3
8		8	2
9		9	1
69		0.5	5.5
70		5.5	4.5

[1](#)), which takes all the QCI parameters into account. Utility function ψ is used to help determining the scheduling priority of bearers.

$$\psi(i) = \begin{cases} \min\left\{\frac{BII(i)}{1-\sqrt{\frac{HOL(i)}{PDB(i)}}}, 500\right\} & \text{if } PDB(i)\text{-HOL}(i) > 15 \\ 500 & \text{if } PDB(i)\text{-HOL}(i) \leq 15 \end{cases} \quad (1)$$

$HOL(i)$ refers to the time difference between the current time and the arrival time of the packet in the head of line of the bearer i . The rationale behind ψ is this: When the ratios of HOL to PDB for two or more different bearers with different QCIs are same the one with higher importance index must have higher utility except for the last 15 ms. Bearers share same utility value when they are closed to be dropped—last 15 ms. Based on our studies on this parameters, let's call it l , 15 ms is *an* optimal value as opposed to *the* optimal value since there are several values that achieve same results. We observed that when there are only GBR bearers in the cell, lower value for l produces better results. Yet, if there are both GBR and NGBR bearers in the system, higher l values result in better performances in terms of packet losses. The delay between P-GW and radio base station—20 ms—should be subtracted from a given PDB to derive the packet delay budget that applies to the radio interface. We add 20 ms to the HOL as soon as packets arrive to the queue. Figure 2 shows the outcome of ψ when all the parameters for each bearer with different requirement plugged into the Eq. (1). We pick 500 as the highest achievable utility value because it is one of high ψ values within the range of every bearers' utility values. In another word, ψ values for each QCI reaches to 500 or beyond within last 15–

1 ms. Higher value helps the algorithm to differentiate between the state of each bearer better. Lower value decreases diversity among ψ values.

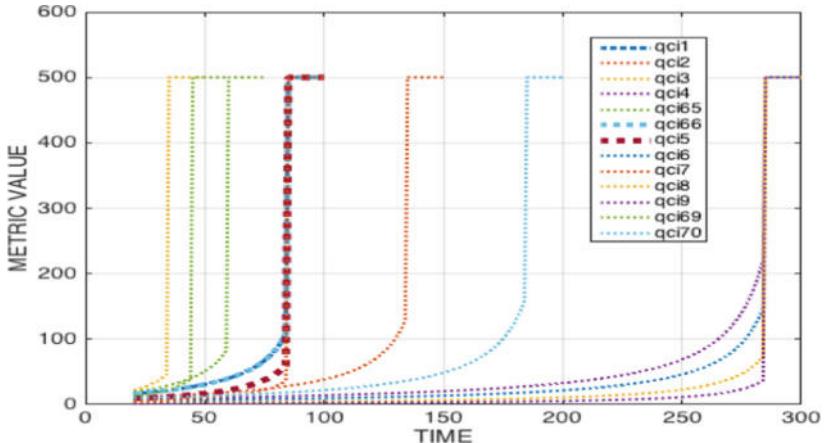


Fig. 2. Utility function ψ . Utility values of each QCI bearer when values plugged into the Eq. (1)

In every TTI, PQA (Algorithm 1) checks QoS parameters and computes ψ values for each bearer that has data to transmit and discard those whose HOLs exceed their PDBs. While there are RBs to allocate, it selects bearer with the highest value. If there is tie, it checks bearer types first. For the sake of simplicity, let's assume there are two bearers with highest ψ values. If bearers have same BIIs, it picks the one with lower RT. Else, if BIIs are different and both are GBR bearers; it picks the one with higher importance index, if it has not reached its required GBR. If it has, then it checks (1) if the one with lower importance has lower RT, (2) it has not achieved its GBR, (3) it has lower ratio of throughput to minimum bit rate in order, if any of these hold true in the order they are laid out, it picks the one with lower BII.

Else, if BIIs are different and both are NGGR bearers; if the one with higher importance index has a β value that is less than what it can tolerate and its RT is higher, then it picks the one with lower BII. Otherwise, it picks the one with higher BII.

Else, if BIIs are different and one is GBR and the other is NGGR bearer; it picks the one with higher importance index, if it has not reached its promised GBR, else it picks the one with lower RT. It assigns enough RBs to the selected bearer to transmit all of its packets stacked in the queue. If there is less than required assign all the remaining bearers and realize transmission. Finally, it discards the bearer from the list.

Algorithm 1 PQA Scheduler

1: Get the list of bearers that have data to transmit
 2: Obtain throughput, HOL and $\beta(i, t)$ values for each bearer
 3: Discard packets whose HOLs exceed their PDBs
 4: Compute ψ based on HOL
 5: **while** there are available RBs **do**
 6: Select the bearer with the highest metric
 7: **if** there is a tie (assume bearer k and l) **then**
 8: **if** $BII(k) == BII(l)$ **then**
 9: Pick the one with lower RT
 10: **else if** $BII(k) > BII(l)$ and Both are GBR **then**
 11: **if** $\theta(k, t) \geq GBR(k)$ and $RT(l) \leq RT(k)$ **then**
 12: Pick l
 13: **else if** $\theta(k, t) \geq GBR(k)$ and $\theta(l, t) < GBR(l)$ **then**
 14: Pick l
 15: **else if**

$$\frac{\theta(l, t)}{GBR(l)} \leq \frac{\theta(k, t)}{GBR(k)}$$
then
 16: Pick l
 17: **else**
 18: Pick k
 19: **end if**
 20: **else if** $BII(k) > BII(l)$ and Both are NGBR **then**
 21: **if** $\beta(k, t) \leq PELR(k)$ and $RT(l) < RT(k)$ **then**
 22: Pick l
 23: **else**
 24: Pick k
 25: **end if**
 26: **else if** $BII(k) > BII(l)$ and k is a GBR bearer **then**
 27: **if** $\theta(k, t) < GBR(k)$ **then**
 28: Pick k
 29: **else**
 30: Pick the one with lower RT
 31: **end if**
 32: **end if**
 33: **end if**
 34: Assign enough RBs to transmit all the packets in the selected bearer's queue
 (or all the remaining RBs, if there is not enough)
 35: Discard it from list of bearers
 36: **end while**

6 Simulation Results and Discussion

6.1 Simulation Setup Parameters

After an initial study on different schedulers—BestCQI, PF, RR, CQA, and TD-PSS—tested in the ns-3 LTE module [18] conducted by National Institute of Standards (NIST) Communications Technology Laboratory (CTL) group, we selected RR and CQA schedulers as baselines of comparison, since they provided the better performance in the studied public safety scenario. CQA implements two channel awareness metrics, we implemented PF metric for this project and will refer to it as *CqaPf* from now on. We developed a simulator which contains bearer features described in QCI table (Table 1) in order to examine the schedulers. Three different deployment scenarios are considered to test performances of the schedulers with respect to GBR, PELR, PDB, priority and system throughput. Simulation runtime is 65 s and the bearers are active until the end of the 59 s. Table 4 summarizes the simulation’s parameters that are common for three scenarios presented. There are 50 RBs in every ms; thus, 50,000 RBs are available in every second. Pre-emption and Admission and Retention Priority (ARP) are not implemented. Every transmission is successful meaning there is no failure due to channel impairment. We set radio link control (RLC) queue size to 1 MB to observe losses only due to PDB.

Table 4. Common simulation parameters for all scenarios

Parameter	Value
Number of RB available at each TTI	50 RB
RLC queue size	1 MB
Schedulers	CqaPf, RR and PQA
Simulation runtime	65 s
ARP, pre-emption	Not Implemented
Applications initiated between	0.1–0.35 ms

6.2 First Scenario—All QCI Bearers

Table 5 presents application configurations for the first scenario where every types of bearers running in the cell. First column represents the instance names. The number at the end of instance name represents the QCI. For example, App-1 is assigned to QCI-1 bearer, App-2 is linked to QCI-2 and so on. Second and third columns indicate packet’s sizes and IATs respectively. Session intervals are set to 0 for all the applications. Each application is employed by 3 UEs who experience different channel conditions: bad, good and better. Column 4, 5 and 6 show the MCS values of these UEs. The number of RBs needed, in 1 s, to

provide QoS is 49,876 for this scenario. This is calculated based on MCS values of the UEs and IAT and sizes of packets. In every figure, the x-axis represents QCI of the bearers and they are ordered with respect to MCS values from low MCS to higher MCS. For example, consecutive 1-1-1 in the x-axis in Fig. 3 represents UE_1^1 with MCS of 5, UE_2^1 with MCS of 18, and UE_3^1 with MCS of 28 in Table 5 employing QCI-1 bearers respectively.

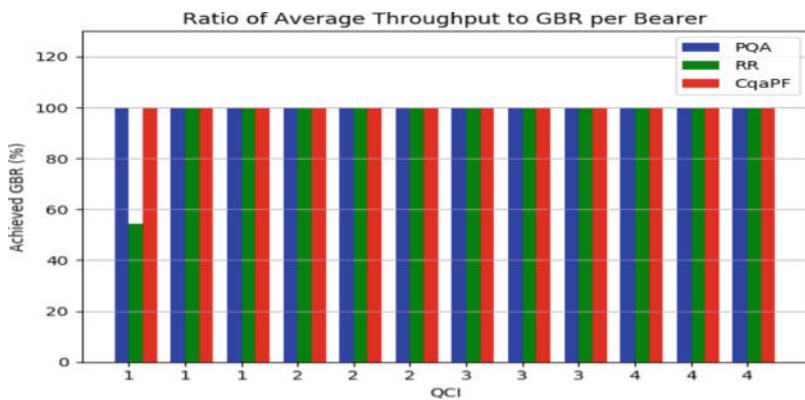
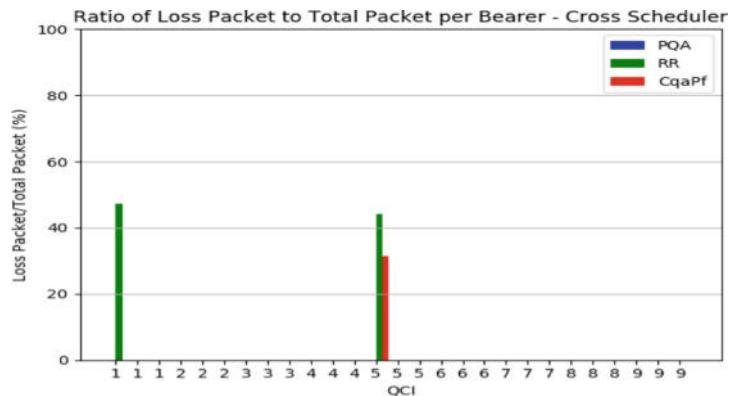
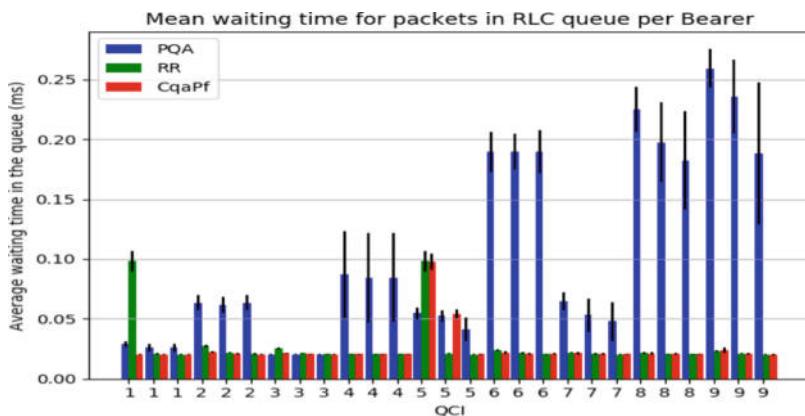
Table 5. Application configuration settings—first scenario

Instance	Size (B)	IAT (ms)	UE_1^j	UE_2^j	UE_3^j	GBR (Mbps)
App-1	338	2.47	5	18	28	136,375
App-2	320	85.3	2	9	17	3750
App-3	256	14.8	6	11	22	17,250
App-4	6	2.66	2	11	18	18,000
App-5	254	1.86	3	16	19	NA
App-6	384	236.3	6	13	23	NA
App-7	253	7.41	7	14	24	NA
App-8	256	14.7	8	16	20	NA
App-9	483	28	8	15	25	NA

Figure 3 illustrates ratio of average throughput of the bearers to set minimum bit rate per GBR bearer. PQA and CqaPf provide all GBR bearers with required minimum bit rate, whereas RR fails to provide GBR for UE_1^1 , which is the bearer with highest priority among GBR bearers. Figure 4 shows ratio of packet loss to total number of packet generated per bearer. CqaPf fails to meet PELR requirement for NGBR bearer used by UE_1^5 . Figure 5 illustrates average waiting time for packets to be transmitted in ms. Waiting time for packet for CQA and RR except UE_1^1 and UE_1^5 , are almost uniform. As for PQA, it is proportional to PDBs of bearers.

Discussion: The applications used by UE_1^1 and UE_1^5 have high demands with respect to MCSs and since it takes RR multiple TTIs to transmit a packet for these bearers, packets are dropped once they reach to PDB. CqaPf divides bearers into different groups in the TD. Because grouping parameter implemented in [14] is too high—300 ms—all the bearers fall into the same group. However, in the FD, bearers with GBR requirements are favored, that is why, UE_1^5 suffer losses but not UE_1^1 .

PQA exploits the differences between delay requirements. Once a packet gets closer to be dropped, the utility value for that packet grows asymptotically (Fig. 2) and thus, chances to be transmitted increases. One observation is that bearers with lower PDBs wait less, although they might have lower priority as in the case of QCI-3 and 1 and QCI-7 and 6. Another observation: Standard deviations for bearers with higher PDBs are higher. These bearers accumulate packets

**Fig. 3.** GBR requirements—first scenario**Fig. 4.** Packet loss ratio—scenario-1**Fig. 5.** Average time packets wait to be transmitted—scenario-1

with different arrival time in the queue until they are picked, once selected, all the packets in the RLC queue are transmitted, if there are enough RBs. Although high waiting time might seem as a trade-off, in fact, it is not because applications that are assigned to proper QCI should tolerate these delays. PQA satisfies all the constraints and objective function (3) of section-IV—zero loss. While CQA fulfills all the constraint, it losses 10,000 packets. RR fails to comply with constraint (2) and performs over 24.000 packets loss.

6.3 Second Scenario—GBR Bearers Only

In this scenario, there are only GBR bearers in the system which is a likely situation in PS environment because in case of emergency when congestion arises in the cell, just MCAs will be admitted to the system. Application configuration is not a realistic though, yet it serves to what we try to measure. We would like to observe how each schedulers behave when applications associated with different GBR bearers share same packets sizes and IATs (thus same Mbps). Table 6 summarizes the applications' configuration settings. 49,910 RBs are needed in 1 s to provide QoS. Each application is used by 6 different UEs and those experience same channel condition—bad, good, better—share same MCS values. Figure 6 represents ratio of packet loss to total packet generated per bearer. Loss above tolerable PELR also means unsatisfied GBR requirement in case of GBR bearer. Thus only this figure is plotted here.

Table 6. Application configuration settings—second scenario

Instance	UE_1^j	UE_2^j	UE_3^j	UE_4^j	UE_5^j	UE_6^j	GBR (Mbps)
App-1	2	5	13	17	20	27	34,125
App-2	2	5	13	17	20	27	34,125
App-3	2	5	13	17	20	27	34,125
App-4	2	5	13	17	20	27	34,125

Discussion: When applications and MCS values are kept constant, PDB is the distinctive property among QCI bearers for CqaPf, while channel quality determines what bearer will experience losses. QCI-3 has the shortest PDB among all QCI bearers and UE_3^1 owns the lowest MCS value among QCI-3 bearer users. For RR, MCS value decides who will suffer losses. For this scenario, PQA satisfies all the constraints and objective function of section-IV—zero loss. While CqaPf and RR fail to fulfill constraint (2) and perform over 4000 and 30.000 packets respectively.

6.4 Third Scenario—GBR Bearers Only

In the third scenario, there are only GBR bearers in the system. This time we play the devil's advocate and change the applications setting against the QCI-1

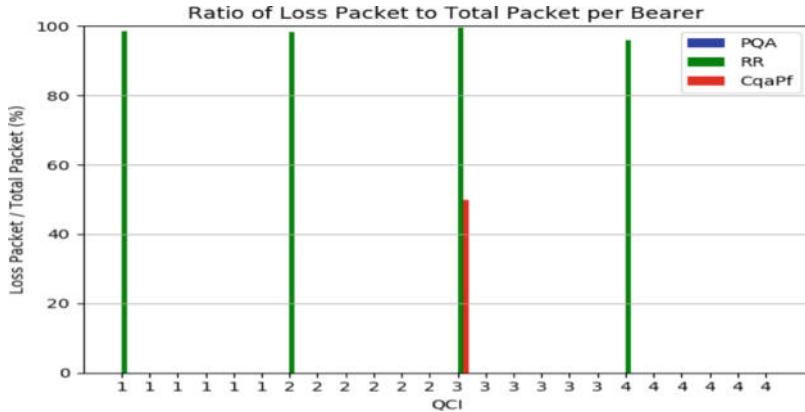


Fig. 6. Packet loss ratio—second scenario

(Table 7). This is a likely situation, for example, in a RE situation, FRs may be trapped in a building and have to utilize a data intensive applications with bad channel quality or in an IP situation, an EMS operator on the scene, attached to a congested cell, may want to use video to consult with doctors regarding a woman giving a birth. Packet sizes, IATs and GBRs of applications are same as Table 4 for this scenario. The number of RBs needed to provide QoS to the bearers is 49,881 in 1 s.

Table 7. Application configuration settings—third scenario

Instance	UE_1^j	UE_2^j	UE_3^j	UE_4^j	UE_5^j	UE_6^j	GBR (Mbps)
App-1	3	8	10	16	17	27	136,375
App-2	3	8	10	16	17	28	3750
App-3	3	8	10	16	17	28	17,250
App-4	3	8	10	16	17	28	18,000

Discussion: UE_1^j performs poorly (Fig. 7) because, CqaPf takes channel quality into account. Although all the UEs with worst channel conditions have same MCS, the one with highest demand losses packets. From a public safety perspective this is unacceptable. As we said earlier, *all the effort must be made to maintain the most important applications*, yet, CQA and RR fail to comply with this principal. PQA satisfies all the constraints and achieves no losses whereas CQA and RR lost over 10,000 and 13,000 packets respectively.

6.5 Fourth Scenario—All QCI Bearers with Higher Demand

With this scenario we would like to illustrate priority performances. This scenario is same as first scenario except the number of RBs needed to satisfy

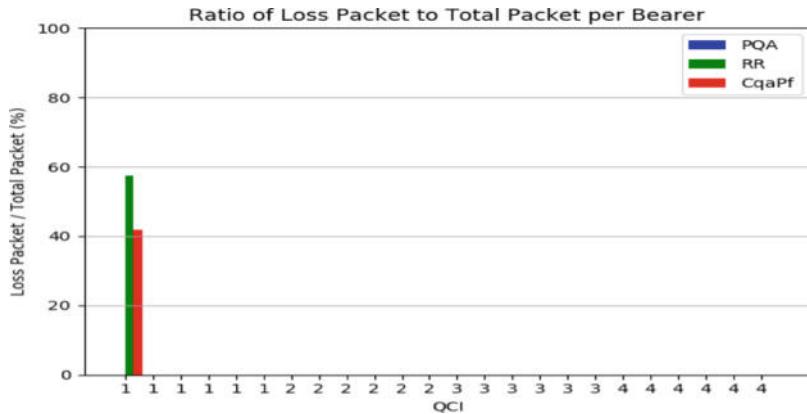


Fig. 7. Packet loss ratio—third scenario

GBR requirements—50,000 RBs. The MCS value of UE_1^8 reduced from 8 to 7 to increase RB demand so that PQA performs packet losses where it did not before. Priority performance is measured by observing in what bearers loss occur. Figure 8 shows aggregated packet losses for bearers with same QCI in every second.

Discussion: The PQA complies with constraints and achieves the objective function—67 packets loss. Losses starts at 37th second because system gets fully

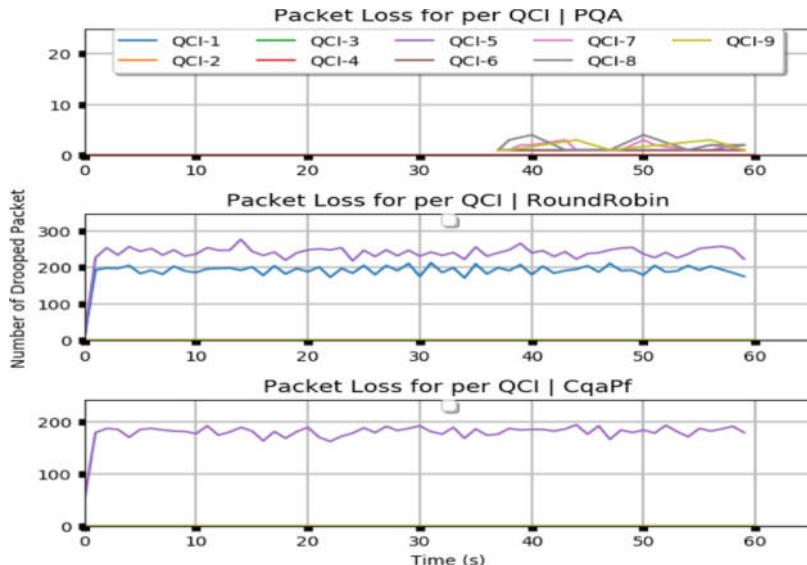


Fig. 8. Packet losses during simulation—scenario-4

congested by that time. Before that point, PQA could defer packets with longer RT further while satisfying urgent bearers. Losses, mostly, occur in bearers with lowest priorities—QCI-7, QCI-8 and QCI-9. If, at any TTIs, a packets has to be dropped PQA selects bearers with lowest BII or a bearer that has met its QoS requirements at that moment. As for CqaPf, packet losses happen immediately after simulations starts and occur only in CQI-5 bearers—over 10.000 packets. Similarly, in RR, packet losses—over 25.000 packets—only occur to QCI-1 and QCI-5 bearers.

Finally, Fig. 9 illustrates the system throughput achieved by each scheduler in four scenarios discussed. Cell throughput accomplished by each scheduler is inversely proportional with their packet losses. PQA achieves highest throughput in all scenarios.

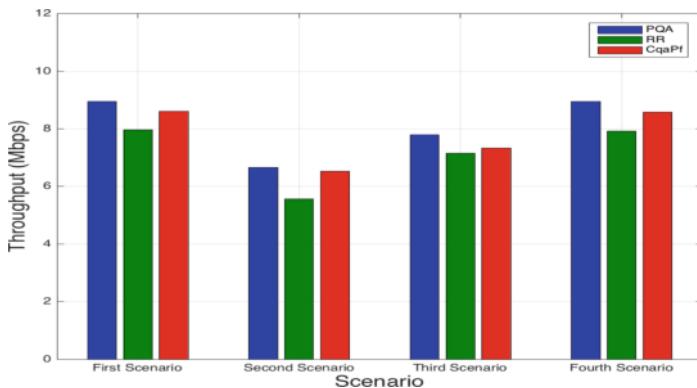


Fig. 9. Cell throughput—cross scenarios

7 Conclusion

We elaborate on a novel scheduler in the context of PSBN, describe scheduling priority and explained why it has to be addressed. Schedulers developed for commercial LTE networks may deprive FRs of critical communications. Channel and/or QoS aware schedulers cannot guarantee that the most important application will be maintained at all time. Therefore, bearer's priority must involve in the scheduling process. Proposed solution is the only algorithm, so far, that considers all the scheduling attributes—bearer type, priority, PDB and PELR—when scheduling packets. It decreases packet losses due to PDB; thus, increases cell throughput while handling QoS in a priority-aware manner. Future research will address the proof of optimality of proposed solution.

Acknowledgment. This work was in part supported by the National Institute of Standards and Technology (NIST) through Professional Research Experience Program—Communications Technology Laboratory (PREP-CTL) with award number 70NANB16H021.

References

1. Wyatt, E.: 9 years after 9/11, public safety radio not ready (2010)
2. Government Technology: A how-to guide for LTE in public safety. Technical report (2010)
3. NPSTC: Priority and quality of service in the nationwide public safety broadband network. Technical report, National Public Safety Telecommunications Council, Aug 2015
4. Capozzi, F., Piro, G., Grieco, L.A., Boggia, G., Camarda, P.: Downlink packet scheduling in LTE cellular networks: key design issues and a survey. *IEEE Commun. Surv. Tutorials* **15**(2), 678–700 (2013). (Second)
5. Ayhan, M., Zhao, Y., Choi, H.A.: Utilizing geometric mean in proportional fair scheduling: enhanced throughput and fairness in LTE DL. In: 2016 IEEE Global Communications Conference (GLOBECOM), pp. 1–6, Dec 2016
6. Dahlman, E., Parkvall, S., Skold, J., Beming, P.: 3G Evolution, 2nd edn. HSPA and LTE for Mobile Broadband. Academic Press (2008)
7. Jalali, A., Padovani, R., Pankaj, R.: Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. In: 2000 IEEE 51st Vehicular Technology Conference Proceedings, 2000. VTC 2000-Spring Tokyo, vol. 3, pp. 1854–1858 (2000)
8. Pokhriyal, A., Pedersen, K.I., Monghal, G., Kovacs, I.Z., Rosa, C., Kolding, T.E., Mogensen, P.E.: HARQ aware frequency domain packet scheduler with different degrees of fairness for the UTRAN long term evolution. In: 2007 IEEE 65th Vehicular Technology Conference—VTC2007-Spring, pp. 2761–2765, Apr 2007
9. Huang, J., Niu, Z.: Buffer-aware and traffic-dependent packet scheduling in wireless OFDM networks. In: Wireless Communications and Networking Conference, 2007. WCNC 2007, pp. 1554–1558. IEEE, Mar 2007
10. Kwan, R., Leung, C., Zhang, J.: Multiuser scheduling on the downlink of an LTE cellular system. *Rec. Lett. Commun.* **2008**, 3:1–3:4 (2008). (January)
11. Monghal, G., Pedersen, K.I., Kovacs, I.Z., Mogensen, P.E.: QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution. In: VTC Spring 2008—IEEE Vehicular Technology Conference, pp. 2532–2536, May 2008
12. Sandrasegaran, K., Mohd Ramli, H.A., Basukala, R.: Delay-prioritized scheduling (DPS) for real time traffic in 3GPP LTE system. In: 2010 IEEE Wireless Communication and Networking Conference, pp. 1–6, Apr 2010
13. Zaki, Y., Weerawardane, T., Gorg, C., Timm-Giel, A.: Multi-QoS-aware fair scheduling for LTE. In: 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), pp. 1–5, May 2011
14. Bojovic, B., Baldo, N.: A new channel and qos aware scheduler to enhance the capacity of voice over LTE systems. In: 2014 IEEE 11th International Multi-Conference on Systems, Signals Devices (SSD14), pp. 1–6, Feb 2014
15. Fan, Y., Lunden, P., Kuusela, M., Valkama, M.: Efficient semi-persistent scheduling for VoIP on EUTRA downlink. In: 2008 IEEE 68th Vehicular Technology Conference, pp. 1–5, Sept 2008

16. Cox, C.: An Introduction to LTE: LTE, LTE-Advanced, SAE, VoLTE and 4G Mobile Communications. Wiley, Chichester (2014)
17. 3GPP. Ts 23 203 (v13.6.0) (rel. 13). Technical report, 3GPP, Mar 2016
18. NS-3 Consortium. ns-3 network simulator (2018)



Efficient Mobile Base Station Placement for First Responders in Public Safety Networks

Chen Shen¹(✉), Mira Yun², Amrinder Arora¹, and Hyeong-Ah Choi¹

¹ Department of Computer Science, George Washington University,
Washington DC 20052, USA

{shenchens, amrinder, hchoi}@gwu.edu

² Department of Computer Science and Networking,
Wentworth Institute of Technology, Boston, MA, USA
yunm@wit.edu

Abstract. We consider the problem of mobile base station placement to meet the critical communication requirements of first responders in an ad hoc public safety network. By considering the class of first responders and UE applications, we provide an efficient base station placement algorithm to maximize critical communication needs according to priority levels. We present simulation results that compare two proposed algorithms with each other and with a baseline algorithm. Our results show that the algorithm of weighted priority and GBR significantly improves connectivity and coverage parameters compared to others.

Keywords: Mobile base station placement · Ad hoc public safety networks · 5G LTE

1 Introduction

Public Safety Networks (PSNs) aim to provide the most critical communication capabilities to the public safety community during both day-to-day operations and large scale events and emergencies [1]. Since disasters and emergencies can occur unexpectedly and exhibit various scales and classes of damage, PSNs may need to be deployed as an ad hoc mobile network. In order to support a wide spectrum of new user equipment (UE) applications of first responders in a timely manner, the PSN must be deployed promptly and efficiently [2,3].

Connectivity and coverage among UEs of some or all first responders are the most basic requirements in many PSNs [4,5]. When the first responders arrive at a disaster site, such as scene of a fire, volcanic eruption, terrorist attack, etc., a PSN must be dynamically deployed to meet the needs of different first responders. Many different deployment mechanisms exist for deploying the base stations. These include, but are not limited to, drones, truck bases, hot air balloons, and being manually established at a location in order to handle the

transportation and installation of the mobile base stations (mBSs). It is likely that these mechanisms will continue to evolve over time. For example, in a recent study from AT&T [6], the concept of the ‘Flying Cow’ or ‘Cell on Wings’ used in the extreme hazardous scenario serving the first responders is presented. Therefore, we study the mBSs placement problem from a deployment mechanism independent perspective and generalize these different mechanisms as various classes of transportation models that have their associated movement costs.

We design our performance metrics of priority based on the work in [1], where the features of QoS, priority and preemption in PSN are studied. For the mBS placement evaluation, we apply the model in [7] where the LTE structure is used for first responders. We extend the wireless network coverage in [8] by using the mBSs instead of flexible network configuration. We determine the optimal location to place mBSs in order to achieve maximum coverage for the various public safety scenarios where the priorities of first responders and the communication applications are emphasized.

The rest of this paper is organized as follows. Section 2 outlines the system model and problem statement. Section 3 describes the proposed method for finding optimal solution. Section 4 presents the empirical results regarding the performance of the algorithm and compares the different algorithms. Finally, Sect. 5 summarizes the paper and outlines ideas for future research.

2 System Model and Problem Statement

2.1 System Model

We are given a set of n UEs $\{U_1, \dots, U_n\}$ and their guaranteed demand bit rate (GBR) $D_U(i)$ and priority $P_U(i)$ for that UE and its application. UEs can move over time, and the location of the i th UE at t th time slot is given by $L_U(i)$. Further, we are given λ mobile base stations (mBSs), $\{B_1, \dots, B_\lambda\}$, which can be moved and configured to meet the needs of UEs. Depending on the location of the UEs and the mBSs, the UE will be affiliated to the mBS with the best signal-to-interference-plus-noise ratio (SINR).

2.2 Objective Function

The communication in PSN usually classifies first responders and the communication applications by different priorities. To represent the priority numerically in the simulation, we introduce the concept of a priority matrix where a priority value is selected for the first responder in a specific priority class and the communication application [1]. An illustrative example of priority matrix used in our simulation is shown in Table 1. The chief contribution of the construct of priority matrix is that it allows for the operational policy to be determined *at runtime* by the operator of the PSN. The algorithms proposed in this paper simply accept the priority matrix as an input and maximize the coverage based on the matrix provided.

Table 1. Table of priorities

UE's/application priority class	Immediate Peril	Responder emergency	Out of service
Mission critical voice	100	50	20
Audio	50	25	10
Video streaming	20	10	5
Periodic sensor data	10	5	1

Next, we design a metrics of performance considering UE's priority and its GBR and denote it as UE's satisfaction score (SS). With UE's location and its affiliated mBS's configuration, the satisfaction score is set to UE's priority if its GBR requirement is met. The satisfaction score is set to 0 if the GBR is not met. The goal is to maximize the total satisfaction score with specific weights on UE's priority and GBR in mBS placement. The process is illustrated as the following.

Algorithm 1: Objective Function Evaluation

```

With the mBSs' placement and UEs' affiliation;
Maximize  $\sum_{i=1}^n SS_i$ ;
if GBR is met then
    |  $SS_i = P_U(i)$  ;
else
    |  $SS_i = 0$  ;
end

```

3 Algorithms for Mobile Base Station Placement

We compare three algorithms of mBS deployment for dynamic coverage. The performance of the algorithm is measured by the total satisfaction score of all UEs. We define a square region of interest (ROI) and UEs can move inside it. Figure 1a illustrates a case of UE distribution where each dot represents a UE and the larger ones have higher priority over smaller ones and circular dots have higher priority over the stars.

3.1 Static Equal-Sized Blocks (SESB)

The first algorithm is a simple static algorithm that is used as a baseline for comparison. In this algorithm, the mBSs are deployed statically with equal-sized blocks. The mBSs serve the UEs that fall in their blocks, regardless of UEs' priority or their GBR. Figure 1b gives an illustrative case of 7 mBSs serving 50 UEs. The mBSs are represented by blue triangles and UEs with the same color

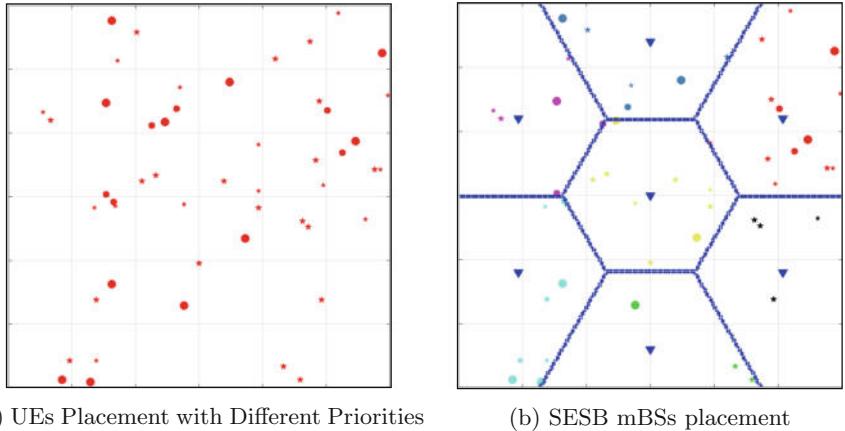


Fig. 1. UEs placement and constant mBS placement

are affiliated with the same mBS. While SESB may appear to be a very simple algorithm, as we observe in comparison results, SESB has the advantage of simple deployment and low costs associated with moving mBSs. This is especially true, since the UEs can move around in a manner that is not predictable and an algorithm that tries to follow the UEs can suffer from low performance if and when the UEs subsequently move away.

3.2 K-Means mBSs Clustering

This algorithm uses k -means clustering based on the UEs location. A random initial deployment of mBSs can fit the convergence of k -means clustering. As a practical matter though, we set our initial mBSs deployment as the above SESB and then apply k -means iteration for faster clustering. With the initial deployment of mBSs and UEs' affiliation, during each iteration, the new location of each mBS will be the geometric center or centroid of its current affiliated UEs. Algorithm 2 illustrates this process. We set the iteration number (MAX_ITER) of 15 where the locations of mBSs usually converge with no further change. Figure 2a shows the final clustering of 7 mBSs.

3.3 GBR and Priority Weighted K -means Clustering

In order to consider both UE's location and priority, we introduce the GBR and priority weighted k -means clustering algorithm. We calculate mBSs placement to maximize critical communication needs according to UE's priority and GBR. The algorithm also uses an iterative process to approach the best result. Similarly starting with the SESB initial placement of mBSs and UEs' affiliation, during each iteration, the location of each mBS will be updated with the combined

Algorithm 2: *k*-means mBSs clustering

```

Initial SESB mBSs placement and UE's affiliation;
iter = 1;
while iter < MAX_ITER do
    for i = 0; i < λ; i = i + 1 do
        num_UE = 0;
        X = 0;
        Y = 0;
        for j = 0; j < n; j = j + 1 do
            if  $U_j \in B_i$  then
                num_UE = num_UE + 1 ;
                X = X +  $U_j.x$  ;
                Y = Y +  $U_j.y$  ;
            end
        end
         $B_i.x = X \div num\_UE$  ;
         $B_i.y = Y \div num\_UE$  ;
    end
    affiliate UE to mBS ;
    iter = iter + 1 ;
end

```

weights as $D_U(i)^\alpha * P_U(i)^\beta$ for U_i . The new coordinate of U_i will be $C_{new} = \frac{\sum C_{current} * D_U(i)^\alpha * P_U(i)^\beta}{\sum D_U(i)^\alpha * P_U(i)^\beta}$, which guarantees the convergence of iteration.

Algorithm 3 demonstrates this process and the input value of α and β can be customized for different weights over priority and GBR. After 15 times of iteration, Figure 2b shows the UEs clustering with weighted GBR and priority on the mBSs' placement.

3.4 Placement Evaluation

In this section, we describe the overall process of how an entire placement is evaluated to receive a unified objective score for the placement. The overall process can be understood as follows. First, the UE is affiliated to the mBS with the best SINR. Then the SINR in dB is converted into channel quality indicator (CQI) value. CQI is an indicator carrying the information on current communication channel quality. According to CQI value, the modulation and coding schemes are selected and then the bit rate based on current radio condition can be calculated. Finally, if the UE's GBR is met, the satisfaction score of UE is set to the UE's priority. Otherwise, the satisfaction score is set to zero. Figure 3 illustrates this process. This process is repeated for all UEs to calculate an aggregate score.

Algorithm 3: GBR and Priority weighted k -means clustering

```

Initial SESB mBSs placement and UE's affiliation;
Initialize  $\alpha$  and  $\beta$  ;
iter = 1 ;
while iter < MAX_ITER do
    for  $i = 0; i < \lambda; i = i + 1$  do
        num_W = 0;
        X = 0;
        Y = 0;
        for  $j = 0; j < n; j = j + 1$  do
            if  $U_j \in B_i$  then
                num_W = num_W +  $D_U(i)^\alpha * P_U(i)^\beta$  ;
                X = X +  $U_j.x * D_U(i)^\alpha * P_U(i)^\beta$  ;
                Y = Y +  $U_j.y * D_U(i)^\alpha * P_U(i)^\beta$  ;
            end
        end
         $B_i.x = X \div num\_W$  ;
         $B_i.y = Y \div num\_W$  ;
    end
    affiliate UE to mBS ;
    iter = iter + 1 ;
end

```

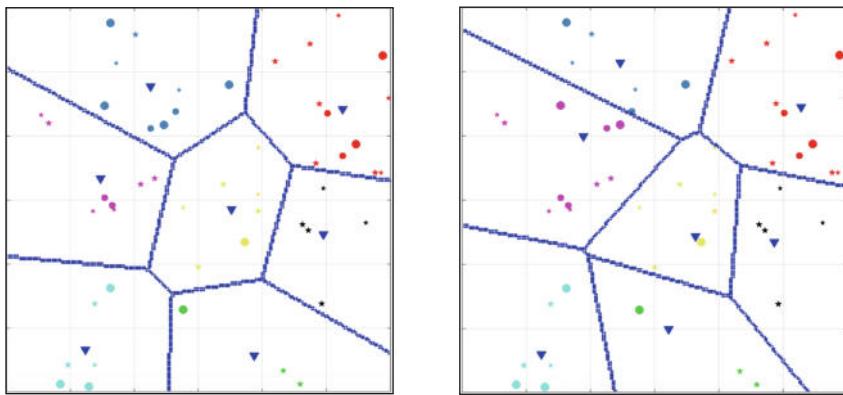


Fig. 2. K -means and GBR, priority weighted K -means clustering

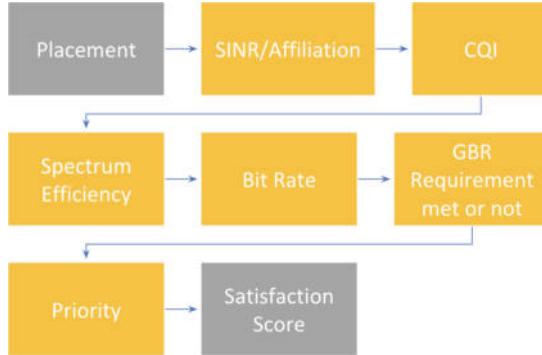


Fig. 3. Placement evaluation processes

4 Empirical Results

In this section, we compare three proposed algorithms in Sect. 3 of the total UEs' satisfaction score for selected simulation scenarios. Before we explain the specific scenarios, we also discuss the mobility model since that has a significant impact on the performance.

4.1 Mobility Model and Movement Dynamics

Due to the obvious and deep reliance of First Responders on mobile Internet-enabled devices, many mobility models for ad hoc networks and cellular networks have been proposed and analyzed [9, 10]. Since mobility models are application and scenario dependent, different mobility patterns are able to provide different impacts on overall network performance. Thus, researchers have repeatedly tried to understand the nature of mobility with respect to various mobility parameters. In this work, we use the well accepted random waypoint (RWP) mobility model [11] due to its simplicity and popularity. Our simulation starts with the UEs uniformly distributed in the rectangle as shown in Fig. 1a. Each UE chooses a random destination and a speed that is uniformly distributed between $[0, 4]$ m/s. Once UE arrives at the destination, it pauses for a random time uniformly distributed in $[0, 60]$ s.

Regardless of the choice of the model for this paper, we agree that the RWP model can not adequately represent all aspects and scenarios of a complete public safety network. In order to capture the movement patterns in disaster scenarios, a few disaster relief mobility models have been proposed. Event-driven and role-based (EDRB) mobility model [12] presented that environmental events and roles, such as civilians, police, firefighters, and ambulances, directly affect a node's movement patterns. Different set of mobility patterns are embedded into different object roles. In reference point group mobility (RPGM) model [13], mobile nodes are organized by groups according to their logical relationships. Each group acts seemingly independently of the other groups, and the random

motion of each user within the group are implemented via RWP model. Overlap mobility model of RPBM allows that different groups carry out different tasks over the same area. Since each group has a unique motion pattern, speed, and scope, the rescue team, medical assistant team, psychologist team, etc. can be modeled differently over the disaster recovery area. Thus, we acknowledge that there are many other mobility models that can be used as directions for our future work.

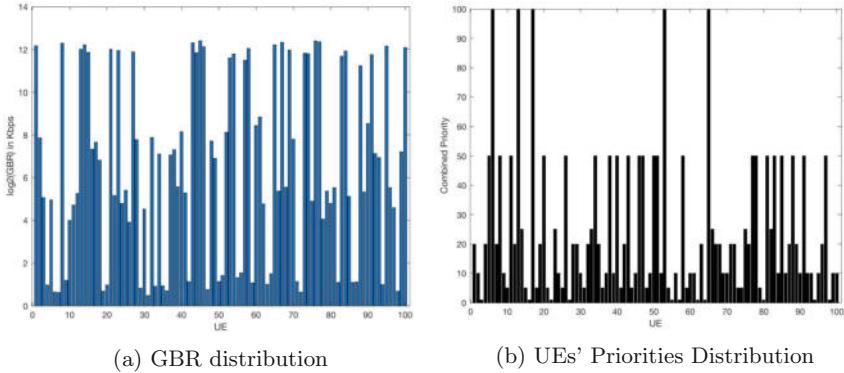


Fig. 4. UEs' GBR and priorities distribution

4.2 Simulation Scenarios

We implement our simulation in MATLAB for 2000 seconds of first responders' movement. The network consists of between 1 and 9 mBSs serving from 20 to 150 UEs in a grid size of 1 km^2 . The maximum communication range of each UE is $300m$. All the results are computed as the average of 100 repetitions. We use the free space radiation model where best SINR converts into the closest mBS.

Data rate between 200 and 4000 Kbps can be expected to support high definition (HD) video conferencing, 30 Kbps for voice communication and between 1 and 800 Kbps for sensory data including temperature, light, motion and chemical can be expected [14]. In the simulation, the UE's application data rate requirement (GBR) is generated as four normal distributions which represents four classes of application priorities where each follows the normal distribution. Figure 4a shows an example of GBR distribution. The UE's location in simulation is generated as uniform distribution in the ROI and the UE's priority class also follows the uniform distribution. Figure 4b shows an example of UEs' combined priority distribution.

4.3 Empirical Results for Static Case

For the static case, without considering UEs' mobility, the simulation has been run for 15 different experimental configurations where the satisfaction score is

Table 2. Empirical Results for Static Case

Satisfaction score averaged over 100 repetitions					
Exp index	Num of mBS	Num of UE	SESB	<i>k</i> -means	Weighted <i>k</i> -means
1	1	20	45.20	46.75	54.30
2	2	30	118.59	139.62	211.71
3	3	50	283.06	339.28	453.97
4	4	50	401.43	479.02	611.44
5	4	80	652.31	715.20	901.11
6	5	100	968.68	1169.7	1331.3
7	6	100	1189.8	1384.2	1562.8
8	7	100	1286.9	1511.7	1696.2
9	7	120	1597.5	1867.1	1999.6
10	7	150	1977.3	2274.4	2462.2
11	8	100	1452.4	1759.1	1917.1
12	8	120	1751.8	2074.9	2218.6
13	8	150	2161.4	2516.9	2632.0
14	9	120	2055.1	2277.3	2374.1
15	9	150	2549.1	2791.4	2912.3

averaged over 100 repetitions. The number of mBS is from 1 to 9 and the number of UE is accordingly from 20 to 150. The simulation results of the satisfaction scores show dominant benefit of Weighted *k*-means over the other two algorithms in the static case. Table 2 shows the satisfaction scores of the three proposed algorithms of different experimental configurations.

4.4 Empirical Results for Mobile Case

Our static simulation results clearly show that weighted priority and GBR significantly improves connectivity and coverage among different priority level requirements. We have four experiments with 7 or 9 mBSs serving 150 UEs. UEs apply either RWP model which is depicted in Sect. 4.1 or following the nearest leaders who have high priority. In our simulation, we define the leaders who have at least 50 priority value which is explained in Sect. 2.2. Table 3 shows the satisfaction scores of the three proposed algorithms in four different experimental configurations.

Since the mBS placement algorithms presented in this work are static and do not adequately take the mobility of UEs into consideration, the scores of SESB shows little fluctuation and the scores of the other two algorithms degrade gradually as the simulation time grows. In order to provide a dynamic mBSs placement algorithm for real world PSN scenario, both the first responders' mobility model and dynamic mBSs placement cost can be considered in future work.

Table 3. Empirical results for mobile case

Satisfaction score of 150 UEs				
Mobility model	Num of mBS	SESB	<i>k</i> -means	Weighted <i>k</i> -means
Random Waypoint	7	1455.9	1565.1	1728.7
Follow Leader	7	1599.0	1691.2	1570.1
Random Waypoint	9	1967.9	1874.0	1832.3
Follow Leader	9	1984.9	1887.9	1951.8

5 Conclusions

In this paper, we have studied the problem of mobile base station placement to meet the critical communication requirements of first responders in an ad hoc public safety network. By considering the class of first responders and the applications, we provide an efficient base station placement algorithm to maximize critical communication needs according to priority and application bit rate requirement. The simulation results have been presented with different network configurations of mBSs and the UEs. In static model, our results clearly show that the algorithm of weighted priority and GBR significantly improves connectivity and coverage parameters compared to two others. In order to provide prompt reaction to the dynamic environmental changes, we consider UEs' mobility models in mobile model simulation.

The future research can consist of studying different mBSs placement cost parameters and thus design the joint algorithm for dynamic coverage that also considers the cost of moving the different base stations. Also, as discussed in the empirical results, significantly more work can be done to validate the presented algorithm using a wider set of mobility models. Finally, as one narrow but specific item, in the GBR and priority weighted *k*-means clustering algorithm, future research can explore the suitable values of α and β .

References

1. Nationwide Public Safety Broadband Network (NPSBN) QoS Priority and Pre-emption (QPP) Framework, Nov 2015 (FirstNet CTO Whitepaper)
2. Li, X., Guo, D., Yin, H., Wei, G.: Drone-assisted public safety wireless broadband network. In: 2015 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), pp. 323–328, Mar 2015
3. Rouil, R., Izquierdo, A., Souryal, M., Gentile, C., Griffith, D., Golmie, N.: Nationwide safety: nationwide modeling for broadband network services. *IEEE Veh. Technol. Mag.* **8**(2), 83–91 (2013). (June)
4. Khakurel, S., Mehta, M., Karandikar, A.: Optimal relay placement for coverage extension in LTE—a cellular systems. In: 2012 National Conference on Communications (NCC), pp. 1–5, Feb 2012

5. Zolotukin, M., Sayenko, A., Hamalainen, T.: On optimal relay placement for improved performance in non-coverage limited scenarios. In: Proceedings of the 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWiM '14, pp. 127–135. ACM, New York, NY, USA (2014)
6. Art Pregler. Extreme Connections. <http://about.att.com/innovationblog/extreme-connections>, May 2018. (AT&T innovation blog)
7. Doumi, T., Dolan, M.F., Tatesh, S., Casati, A., Tsirtsis, G., Anchan, K., Flore, D.: LTE for public safety networks. *IEEE Commun. Mag.* **51**(2), 106–112 (2013). (February)
8. Bruno, R., Conti, M., Gregori, E.: Mesh networks: commodity multihop ad hoc networks. *IEEE Commun. Mag.* **43**(3), 123–131 (2005). (March)
9. Batabyal, S., Bhaumik, P.: Mobility models, traces and impact of mobility on opportunistic routing algorithms: a survey. *IEEE Commun. Surv. Tutorials* **17**(3), 1679–1707 (2015). (Thirdquarter)
10. Lin, X., Ganti, R.K., Fleming, P.J., Andrews, J.G.: Towards understanding the fundamentals of mobility in cellular networks. *IEEE Trans. Wirel. Commun.* **12**(4), 1686–1698 (2013). (April)
11. Johnson, D.B., Maltz, D.A.: Dynamic Source Routing in Ad Hoc Wireless Networks, pp. 153–181. Springer US, Boston, MA (1996)
12. Nelson, S.C., Harris III, A.F., Kravets, R.: Event-driven, role-based mobility in disaster recovery networks. In: Proceedings of the Second ACM Workshop on Challenged Networks, CHANTS '07, pp. 27–34. ACM, New York, NY, USA (2007)
13. Hong, X., Gerla, M., Pei, G., Chiang, C.-C.: A group mobility model for ad hoc wireless networks. In: Proceedings of the 2nd ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWiM '99, pp. 53–60. ACM, New York, NY, USA (1999)
14. Almalkawi, I.T., Guerrero Zapata, M., Al-Karaki, J.N., Morillo-Pozo, J.: Wireless multimedia sensor networks: current trends and future directions. *Sensors* **10**(7), 6662–6717 (2010)



Solutions of Partition Function-Based TU Games for Cooperative Communication Networking

Giovanni Rossi^(✉)

Department of Computer Science and Engineering DISI,
University of Bologna, 40126 Bologna, Italy
giovanni.rossi6@unibo.it

Abstract. In networked communications nodes choose among available actions and benefit from exchanging information through edges, while continuous technological progress fosters system functionings that increasingly often rely on cooperation. Growing attention is being placed on coalition formation, where each node chooses what coalition to join, while the surplus generated by cooperation is an amount of TU (transferable utility) quantified by a real-valued function defined on partitions -or even embedded coalitions- of nodes. A TU-sharing rule is thus essential, as how players are rewarded determines their behavior. This work offers a new option for distributing partition function-based surpluses, dealing with cooperative game theory in terms of both global games and games in partition function form, namely lattice functions, while the sharing rule is a point-valued solution or value. The novelty is grounded on the combinatorial definition of such solutions as lattice functions whose Möbius inversion lives only on atoms, i.e. on the first level of the lattice. While rephrasing the traditional solution concept for standard coalitional games, this leads to distribute the surplus generated by partitions across the edges of the network, as the atoms among partitions are unordered pairs of players. These shares of edges are further divided between nodes, but the corresponding Shapley value is very different from the traditional one and leads to two alternative forms, obtained by focusing either on marginal contributions along maximal chains, or else on the uniform division of Harsanyi dividends. The core is also addressed, and supermodularity is no longer sufficient for its non-emptiness.

Keywords: Networked communications · Cooperative game theory · Partition function · Shapley value · Lattice · Möbius inversion

1 Introduction

For the present purposes, a communication network may be looked at as a simple graph $G^t = (N, E^t)$ varying in time t . The set $N = \{1, \dots, n\}$ of nodes or vertices can be constant, while the set $E^t \subseteq N_2 = \{\{i, j\} : 1 \leq i < j \leq n\}$ of edges

in the network at any time t results from node behavior for given technological environment. In fact, in game-theoretical models of communication networking [1,2] nodes are players whose behavior generally realizes as a time-sequence of choices over available actions. Most importantly, these players attain a utility by exchanging information with each other through the network, while the functioning of this latter increasingly often relies on cooperation. In this view, “cooperative game theory provides a variety of tools useful in many applications”, allowing “to model a broad range of problems, including cooperative behavior, fairness in cooperation, network formation, cooperative strategies, and incentives for cooperation”, and with applications such as “information trust management in wireless networks, multi-hop cognitive radio, relay selection in cooperative communication, intrusion detection, peer-to-peer data transfer, multi-hop relaying, packet forwarding in sensor networks, and many other/s” [1, p. 220]. In particular, a range of settings seems to be fruitfully modeled by means of coalition formation games [3], which combine strategic and collaborative behavior, in that players choose what coalition to join, while behaving in a ‘fully cooperative’ manner within the chosen coalition. Such a modeling choice seems grounded on the evidence that “recently, there has been a significant increase of interest in designing autonomic communication systems. Autonomic systems are networks that are self-configuring, self-organizing, self-optimizing, and self-protecting. In such networks, the users should be able to learn and adapt to their environment (changes in topology, technologies, service demands, application context, etc), thus providing much needed flexibility and functional scalability” [2]. All of this sounds extremely hard -if not impossible- to achieve without the absolute and constant cooperation of nodes with each other according to an agreed protocol that best fits the needs and scope of the whole system. However, in order for choices to be rational (thereby allowing to discuss equilibria and related stability concerns), players have to be rewarded depending on their choices. How to reward them in partition function-based coalition formation games is precisely (and almost exclusively) the object dealt with in the present work, as the central role played by strategic equilibria in non-cooperative settings is replaced in cooperative ones with solutions or values, namely with mappings specifying how to share the surplus of cooperation between players. Solutions of cooperative games thus address the same stability issue as equilibria of non-cooperative games, since the idea is that for relevant generated surplus (i.e. given by, say, a supermodular lattice function) there are meaningful (i.e. fair, efficient, etc.) solutions leading everybody to cooperate.

1.1 Related Work

Insofar as TU (transferable utility) games are concerned, cooperative communication networking is frequently modeled by means of coalitional games (see Class I in [1, Sect. 7.2]), namely functions taking real values on the Boolean lattice [4] of subsets or coalitions of players. This is the traditional setting where TU games are mostly known, and the Shapley value is a fundamental solution [5]. However, recently attention has been placed also on more complex games involving the

geometric (indecomposable) lattice [4] of partitions of nodes [2]. A partition (or coalition structure) is a family of pairwise disjoint (nonempty) subsets of N , called blocks (or clusters, in the present framework), whose union is N . Many environments are characterized by space-time node dynamics and technological means resulting in a clustered wireless network G^t at each time t . That is, active nodes may be partitioned or clustered in conformity with the given communication technology [6–9]. For these settings, the proposed approach relies mostly on distributed P2P communication systems where all (active) nodes behave in a collaborative manner, and seems to best fit multi-hop scenarios, where it may also provide an additional perspective for identifying cluster heads [10–22]. The foundation is perhaps best summarized by the subtitle of [23], i.e. “*real egoistic behavior is to cooperate!*”. When the whole system itself is very worthy to collaborating users, how to share such a worth is precisely the purpose of point-valued solutions (possibly in conjunction with set-valued ones such as the core [24, 25], see below). The dual perspective applies as well: if network maintainance is (computationally) demanding, point-valued solutions also enable to fairly and efficiently share the corresponding costs. How a constant global cooperation in communication networking is quantifiable as a partition function and why this may be meaningfully modeled via coalition formation are both comprehensively explained in [1, 2], hence these topics are not discussed here. On the other hand, what emerges from the novel solutions proposed in the sequel is that the edges of the network may be looked at as the true players in partition function-based TU games. Accordingly, these proposed solutions are mappings that share the surplus generated by partitions *primarily* between such edges. Of course, since edges do not gain from receiving an amount of TU, their shares are going to be further divided between the corresponding endvertices [26]. While providing a reward criterion that enforces trust via automated reciprocal control because of the involvement of two players for each share, this is also consistent with a strictly game-theoretical perspective in view of the following combinatorial argument.

In TU cooperative game theory, partition functions are global games [27], meant to model cooperation over global issues such as environmental clean-up and preservation. Specifically, every coalition structure or partition P of players has an associated surplus or worth $f(P)$, to be interpreted as the level of satisfaction common to all players attained when cooperation operates through P . Partitions of players or nodes are elements of an atomic lattice [4] whose atoms are in fact the $\binom{n}{2}$ unordered pairs $\{i, j\} \in N_2$ of nodes, namely the edges of the network. In the same way, singleton nodes or players $\{i\}, i \in N$ are atoms in the Boolean lattice where coalitional games take their values. The solutions proposed in the sequel are defined in terms of lattice functions with Möbius inversion living only on atoms [28], and this means mapping any given coalitional game into n shares, one for each node, and any given global game into $\binom{n}{2}$ shares, one for each edge. Thus for traditional set functions or coalitional games the novel definition simply rephrases the existing one, but for global games it leads to crucial novelties. In this respect, communication networking is

also sometimes modeled by means of a further type of TU cooperative games, known as ‘games in partition function form’ PFF (see [1, p. 205] on [29]). These PFF games assign a worth to every embedded subset, namely to every coalition embedded into a partition as a block, and might appear quite puzzling, especially in terms of their solutions (see [30,31] among others). A recent approach [32] shows that embedded subsets may be dealt with as elements of a lattice isomorphic [4] to the partition lattice; specifically, PFF games on n players are combinatorially equivalent to global games on $n+1$ players. Hence the proposed solutions apply invariably to both global and PFF games, as explained below insofar as possible, especially through an example fully devoted to the parallelizing. These different TU games are formalized as lattice functions in Sect. 2 hereafter, while also briefly detailing the well-known Shapley value of coalitional games as well as the main combinatorial aspects of global and PFF games. Next Sect. 3 introduces the novel solution concept in terms of Möbius inversion and atoms, showing how partition function-based TU cooperative games allow for two distinct Shapley values because of the linear dependence [4,33] characterizing the partition lattice, and with Sect. 3.1 devoted to symmetric games (i.e. functions [34]), which seem possibly useful to model the surplus generated by cooperation in certain communication networking systems and whose solutions are determined in a straightforward manner. Section 4 shows how PFF games on n players are isomorphic to global games on $n+1$ players as long as the lattice of embedded subsets is taken to be geometric following [32], and the isomorphism is computationally detailed by means of an example. Section 5 translates the proposed point-valued solutions in terms of the core [35,36], which is the main set-valued solution concept, while showing that supermodularity is no longer sufficient for its non-emptiness and by also briefly considering the case of additive [37] or additively separable [27,38] partition functions. Section 6 contains the conclusion.

2 TU Cooperative Games as Lattice Functions

TU cooperative games are functions taking real values on some lattice (L^N, \wedge, \vee) grounded on player set N . Standard C (coalitional) games $v : 2^N \rightarrow \mathbb{R}$ are set functions defined on Boolean lattice $(2^N, \cap, \cup)$, where $2^N = \{A : A \subseteq N\}$ is the 2^n -set of coalitions ordered by inclusion \supseteq . The meet \wedge and join \vee of any two subsets $A, B \in 2^N$ respectively are intersection $A \cap B$ and union $A \cup B$. On the other hand, global G games are partition functions $f : \mathcal{P}^N \rightarrow \mathbb{R}$, i.e. defined on the geometric indecomposable [4] lattice $(\mathcal{P}^N, \wedge, \vee)$ whose elements $P = \{A_1, \dots, A_{|P|}\}, Q = \{B_1, \dots, B_{|Q|}\} \in \mathcal{P}^N$ consist of blocks $A \in P, B \in Q$, namely nonempty and pairwise disjoint subsets $A, B \in 2^N$ whose union equals N , hence $A \cap A' = \emptyset$ for $A, A' \in P$ while $A_1 \cup \dots \cup A_{|P|} = N$. Partitions P, Q are ordered by coarsening \geqslant , i.e. $P \geqslant Q$ (or P is coarser than Q) if every block $B \in Q$ is included in some block $A \in P$, i.e. $A \supseteq B$. Meet $P \wedge Q$ and join $P \vee Q$ respectively are the coarsest partition finer than both P, Q and the finest partition coarser than both P, Q . Lattices $(2^N, \cap, \cup), (\mathcal{P}^N, \wedge, \vee)$ are atomic, since

every element decomposes as the join of those elements immediately above the bottom in the covering graph (or Hasse diagram) of the lattice, i.e. the atoms [4]. Among subsets, \emptyset is the bottom and the n singletons $\{i\}, i \in N$ are the atoms.

2.1 The Shapley Value

The Shapley value $\phi^{Sh}(v)$ is a fundamental solution of C games v [5]. Geometrically, $\phi^{Sh} : \mathbb{R}^{2^n} \rightarrow \mathbb{R}^n$ is a mapping with $\phi^{Sh}(v) = (\phi_1^{Sh}(v), \dots, \phi_n^{Sh}(v)) \in \mathbb{R}^n$ defined by

$$\phi_i^{Sh}(v) = \sum_{A \subseteq N \setminus i} \frac{v(A \cup i) - v(A)}{n \binom{n-1}{|A|}} = \sum_{A \subseteq N \setminus i} \frac{\mu^v(A \cup i)}{|A| + 1} \quad (i \in N), \quad (1)$$

and $\mu^v : 2^N \rightarrow \mathbb{R}$ is the Möbius inversion of v , i.e. $\mu^v(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} v(B)$ [4, 39]. The values taken by μ^v are also sometimes referred to as ‘Harsanyi dividends’ [5]. The former expression obtains by placing the uniform (probability) distribution over the $n!$ maximal chains in lattice $(2^N, \cap, \cup)$, and considering marginal contributions $v(A \cup i) - v(A)$ of players $i \in N$ to coalitions $A \subseteq N \setminus i$. A maximal chain is a $n+1$ -set $\{A_0, A_1, \dots, A_n\} \subset 2^N$ such that $|A_k| = k, 0 \leq k \leq n$ and $A_{l+1} \supset A_l, 0 \leq l < n$, where \supset is proper inclusion, hence A_{l+1} covers A_l [4]. In this view, $\phi_i^{Sh}(v)$ is the expectation of the marginal contribution of i to a random coalition $A \subseteq N \setminus i$, as detailed in [40] in terms of probabilistic and random-order values. The latter expression regards $\mu^v(A)$ as the net added worth (possibly < 0) of cooperation within coalition A with respect to all its proper subcoalitions B . That is, $\mu^v(A) = v(A) - \sum_{B \subset A} \mu^v(B)$. This net added worth is equally shared between coalition members $i \in A$ according to the following well-known axiomatic characterization of ϕ^{Sh} (v, v' are C games, $\alpha, \beta \in \mathbb{R}$, $i, j \in N$).

L (linearity): $\phi(\alpha v + \beta v') = \alpha\phi(v) + \beta\phi(v')$.

S (symmetry): if $v(A \cup i) = v(A \cup j)$ for all $A \subseteq N \setminus \{i, j\}$, then $\phi_i(v) = \phi_j(v)$.

D (dummy): if $v(A \cup i) = v(A) + v(\{i\})$ for all $A \subseteq N \setminus i$, then $\phi_i(v) = v(\{i\})$.

E (efficiency): $\sum_{i \in N} \phi_i(v) = v(N)$.

Indexing axes by coalitions $A \in 2^N$, C games $v \in \mathbb{R}^{2^n}$ are points in a vector space. A main basis of this space is $\{\zeta_A : A \in 2^N\}$, where $\zeta_A(B) = \begin{cases} 1 & \text{if } B \supseteq A \\ 0 & \text{if } B \not\supseteq A \end{cases}$ is the zeta function ζ_A [4, 39] (also termed unanimity game u_A [5]). Any v is the following linear combination of basis elements: $v(\cdot) = \sum_{A \in 2^N} \zeta_A(\cdot) \mu^v(A)$ or $v(B) = \sum_{A \subseteq B} \mu^v(A)$. If ϕ satisfies *L*, then $\phi(v) = \sum_{A \in 2^N} \mu^v(A) \phi(\zeta_A)$. The focus can thus be placed on the implications of *S*, *D*, *E* for any ζ_A . Now, *D* entails $\phi_j(\zeta_A) = 0$ for all $j \in A^c = N \setminus A$, while $\phi_i(\zeta_A) = \phi_{i'}(\zeta_A)$ for all $i, i' \in A$ in view of *S*. Finally, *E* requires $\phi_i(\zeta_A) = \frac{1}{|A|}$ for all $i \in A$. Observe that the fixed points of the Shapley value mapping are those C games v such that $v(A) = \sum_{i \in A} v(\{i\})$ for all coalitions $A \in 2^N$. Such set functions v are valuations [4] of Boolean lattice $(2^N, \cap, \cup)$, satisfying $v(A \cup B) + v(A \cap B) = v(A) + v(B)$ for all $A, B \in 2^N$. Their

Möbius inversion satisfies $\mu^v(A) = 0$ for all $A \in 2^N$, $|A| > 1$. Furthermore, $\mu^{\zeta_A}(B) = \begin{cases} 1 & \text{if } B = A \\ 0 & \text{otherwise} \end{cases}$.

2.2 Global and PFF Games

The bottom partition is $P_\perp = \{\{1\}, \dots, \{n\}\}$, with rank $r(P_\perp) = n - |P_\perp| = 0$, while the $\binom{n}{2}$ atoms have rank 1 and consist each of $n - 1$ blocks, i.e. $n - 2$ singletons and one (unordered) pair. Denote by $[ij] \in \mathcal{P}^N$ the atom whose non-singleton block is pair $\{i, j\} \in N_2$. A main difference between $(2^N, \cap, \cup)$ and $(\mathcal{P}^N, \wedge, \vee)$ relies in the join-decomposition [4] of their elements: $A \in 2^N$ decomposes uniquely as $A = \cup_{i \in A} \{i\}$, while most partitions admit several decompositions as a join of atoms. In fact, $r(A) = |A|$ is precisely the number of atoms involved in the join-decomposition of subset A , while $r(P) = n - |P|$ is the minimum number of atoms involved in the join-decomposition of partition P . This is immediately seen for the top partition which consists of a single block, i.e. $P^\top = \{N\}$, in that $P^\top = [ij]_1 \vee \dots \vee [ij]_{n-1}$ for any $n - 1$ atoms satisfying $|\{ij\}_k \cap \{ij\}_m| = \begin{cases} 1 & \text{if } m = k + 1 \\ 0 & \text{otherwise} \end{cases}$, $1 \leq k < m < n$, whereas clearly $P^\top \geq [ij]$ for all the $\binom{n}{2}$ atoms, entailing $P^\top = [ij]_1 \vee \dots \vee [ij]_{\binom{n}{2}}$. Define the maximum number of atoms involved in the join-decomposition of partitions to be the size $s : \mathcal{P}^N \rightarrow \mathbb{Z}_+$ of these latter. That is to say, $s(P) = |\{[ij] : P \geq [ij]\}|$. Evidently, $s(P) \geq r(P)$, with equality if and only if P has no blocks larger than pairs.

As outlined above, G (global) games are partition functions $f : \mathcal{P}^N \rightarrow \mathbb{R}$, with $f(P)$ quantifying the surplus generated by cooperation when players are operating through coalition structure P . However, as detailed in the sequel, single players $i \in N$ factually may not provide any marginal contribution in G games. In terms of communication networks, they are isolated vertices, hence either inactive (at some time t under concern) or else not able to communicate with anybody because of their spatial position. Conversely, atoms essentially coincide with pairs of players, i.e. hops or edges, and they do contribute to the functioning of the network as vehicles for information exchange. In other terms, in order for an edge to be worthy to the collective, both endvertices must collaborate.

The zeta function $\zeta_P(Q) = \begin{cases} 1 & \text{if } Q \geq P \\ 0 & \text{otherwise} \end{cases}$ is the analog basis element as $\zeta_A, A \in 2^N$; again Möbius inversion $\mu^f(P) = f(P) - \sum_{Q < P} \mu^f(Q)$ provides the coefficients of the linear combination of these basis elements, where $<$ is proper coarsening and $\mu^{\zeta_P}(Q) = \begin{cases} 1 & \text{if } Q = P \\ 0 & \text{otherwise} \end{cases}$, i.e. $f(\cdot) = \sum_{P \in \mathcal{P}^N} \mu^f(P) \zeta_P(\cdot)$ or $f(Q) = \sum_{P \leq Q} \mu^f(P)$. Concerning marginal contributions of atoms, $P \not\geq [ij]$ entails that $P \vee [ij]$ covers P , usually denoted [4] by $P \vee [ij] > P$, and $P \vee [ij]$ obtains by merging those two blocks $A, A' \in P$ such that $A \cap \{i, j\} = \{i\}$ and $A' \cap \{i, j\} = \{j\}$. Maximal chains of partitions are collections $\{P_0, P_1, \dots, P_{n-1}\}$ where $P_k > P_{k-1}$, hence $r(P_k) = k$, $0 \leq k < n$, with $P_0 = P_\perp$, $P_{n-1} = P^\top$. There are $\frac{n!(n-1)!}{2^{n-1}}$ (distinct) maximal chains of partitions.

PFF games, here denoted by h , are more complex lattice functions, taking their values on embedded subsets, namely (ordered) pairs $(A, P) \in 2^N \times \mathcal{P}^N$ such that $A \in P$, i.e. A is a block of P . Although \emptyset is not a block of any partition, still Möbius inversion μ^h needs a bottom element, and thus (\emptyset, P_\perp) is taken to be the bottom embedded subset [41]. Combinatorial congruence then requires (\emptyset, P) to be an embedded subset for all $P \in \mathcal{P}^N$, otherwise the resulting lattice is non-atomic. In this way, the family $\mathcal{E}^N \subset 2^N \times \mathcal{P}^N$ of embedded subsets (A, P) such that either $A \in P$, or else $A = \emptyset$, is a geometric lattice isomorphic [4] to \mathcal{P}^{N+} , where $N_+ = \{1, \dots, n, n+1\}$. The $\binom{n+1}{2} = n + \binom{n}{2}$ atoms of \mathcal{E}^N are the n pairs (i, P_\perp) together with the $\binom{n}{2}$ pairs $(\emptyset, [ij])$. Denote by $(\mathcal{E}^N, \sqcap, \sqcup)$ this lattice, with order relation \sqsupseteq ; it is comprehensively detailed in [32]. Given the isomorphism, there are $\frac{(n+1)!n!}{2^n}$ maximal chains in \mathcal{E}^N , along which atoms provide marginal contributions like in C and G games above.

3 Solutions

Denote by (L^N, \wedge, \vee) a lattice $L^N \in \{2^N, \mathcal{P}^N, \mathcal{E}^N\}$ grounded on player set N , with elements $x, y, z, \dots \in L$, order relation \geqslant and bottom element x_\perp . Let L_A^N be the set of atoms of L^N , while C, G and PFF TU cooperative games may now be dealt with at once as lattice functions $f : L^N \rightarrow \mathbb{R}$. Recall that $f : L^N \rightarrow \mathbb{R}$, with Möbius inversion $\mu^f : L^N \rightarrow \mathbb{R}$, is totally positive if $\mu^f(x) \geq 0$ for all $x \in L^N$, while if $f(x \wedge y) + f(x \vee y) \geq f(x) + f(y)$ for all $x, y \in L^N$, then f is supermodular. Total positivity is a sufficient (but not necessary) condition for supermodularity. Valuations of (L^N, \wedge, \vee) are those f satisfying supermodularity with equality, i.e. $f(x \wedge y) + f(x \vee y) = f(x) + f(y)$ for all $x, y \in L^N$. As already explained, for coalitional games v , a solution is a mapping ϕ associating with v a further coalitional game $\phi(v)$ which is a valuation of subset lattice $(2^N, \cap, \cup)$. Hence $\phi(v)$ is an *inessential* game, as every player is a dummy: $\phi(v)(A) = \sum_{i \in A} \phi_i(v)$. That is, $\phi(v)$ is a coalitional game formalizing a situation where there is no surplus generated by cooperation. A standard and most natural assumption is $v(\emptyset) = 0 = \phi(v)(\emptyset)$. Then, given general result [4, Theorem 4.63, p. 190] for valuations of distributive lattices, solutions $\phi(v)$ of coalitional games v may be equivalently defined to be coalitional games whose Möbius inversion $\mu^{\phi(v)}$ lives only on atoms: $\mu^{\phi(v)}(A) = 0$ for all $A \in 2^N$ such that $|A| \neq 1$.

Both \mathcal{P}^N and \mathcal{E}^N are geometric indecomposable [4, p. 61], and thus valuations of these lattices are constant functions [4, Exercise IV.4.12, p. 195]. Accordingly, solutions of G and PFF games cannot be defined in terms of valuations of \mathcal{P}^N and \mathcal{E}^N , respectively. In fact, solutions of G games were defined [27] in terms of valuations of subset lattice $(2^N, \cap, \cup)$, with the implication that only the worth $f(P_\perp^A)$ of those $2^n - n$ partitions $P_\perp^A = \{A, \{i_1\}, \dots, \{i_{n-|A|}\}\}$ with at most one non-singleton block is taken into account, where $A^c = \{i_1, \dots, i_{n-|A|}\}$. Thus $B_n - (2^n - n)$ values taken by f are disregarded, B_n being the (Bell) number of partitions of a n -set [4]. The same argument applies to existing solutions of PFF games [30, 31]. For these reasons, in the remainder of this work solutions of TU cooperative games are conceived in terms of Möbius inversion of lattice functions [28] as follows.

Definition 1. Solutions of cooperative games $f : L^N \rightarrow \mathbb{R}$ are mappings ϕ associating with f an analog game $\phi(f) : L^N \rightarrow \mathbb{R}$ whose Möbius inversion $\mu^{\phi(f)} : L^N \rightarrow \mathbb{R}$ lives only on atoms, i.e. $\mu^{\phi(f)}(x) = 0$ for all $x \in L^N \setminus L_A^N$.

Thus a solution of a game is a lattice function taking values on the same lattice where the game itself takes its values. In this way a solution is a game as well, and in particular one where cooperation no longer generates any surplus. This seems the only combinatorially consistent way to include G and PFF games within the standard framework of C games. Let $a \in L_A^N$ be the generic atom of L^N and $\phi_a(f) = \phi(f)(a)$. Then,

$$\phi(f)(x) = \sum_{a \leqslant x} \phi_a(f) \text{ for all } x \in L^N. \quad (2)$$

Example 1. Consider a simplest G game $f : \mathcal{P}^N \rightarrow \mathbb{R}$ defined by $f = \binom{n}{2} \zeta_{P^\top}$. That is, $f(P) = \begin{cases} \binom{n}{2} & \text{if } P = P^\top \\ 0 & \text{if } P < P^\top \end{cases}$. The size $s = \phi(f)$ defined in Sect. 2 seems a suitable solution, as $s(P_\perp) = 0$ while on atoms $s([ij]) = 1 = \phi_{[ij]}(f)$, and its Möbius inversion satisfies $\mu^s(Q) = \begin{cases} 1 & \text{if } s(Q) = 1 \\ 0 & \text{if } s(Q) \neq 1 \end{cases}$. Hence $\phi_{[ij]}(f) = \mu^s([ij])$ and $s(P) = \sum_{Q \leqslant P} \mu^s(Q) = \sum_{[ij] \leqslant P} \mu^s([ij]) = \sum_{A \in P} \binom{|A|}{2}$. \square

Following the above traditional axiomatic characterization of the Shapley value [5], firstly consider L (linearity).

Definition 2. A solution ϕ is linear if $\phi(\alpha f) = \alpha \phi(f)$ for $\alpha \in \mathbb{R}$, as well as $\phi(f + f') = \phi(f) + \phi(f')$ for $f, f' : L^N \rightarrow \mathbb{R}$.

Since $\{\zeta_x : x \in L^N\}$ is a basis of the so-called [4] free vector space $\mathbb{R}^{|L^N|}$ of lattice functions $f : L^N \rightarrow \mathbb{R}$, with coefficients given by Möbius inversion, i.e. $f(\cdot) = \sum_{x \in L^N} \mu^f(x) \zeta_x(\cdot)$ or $f(y) = \sum_{x \leqslant y} \mu^f(x)$, a solution satisfying L has form

$$\phi(f) = \sum_{x \in L^N} \mu^f(x) \phi(\zeta_x). \quad (3)$$

Such solutions are univocally defined by specifying how to distribute the unit of TU given by ‘zeta games’ $\zeta_x, x \in L^N$.

Definition 3. Denoting by x^\top the top element of L^N , a solution ϕ is efficient if $\sum_{a \in L_A^N} \phi_a(f) = f(x^\top)$.

E (efficiency) was conceived [5] to deal with monotone (and superadditive [42]) C games, in which case it seems a most natural assumption. But for G and PFF games it requires some caution. In fact, a lattice function f is monotone if for any $x, y \in L^N$ such that $x \geqslant y$, inequality $f(x) \geq f(y)$ holds, entailing $f(x^\top) \geq f(x)$ for all $x \in L^N$ (while superadditivity is neither straightforwardly translated nor interesting for G and PFF games). When modeling the surplus generated by cooperation in clustered (multi-hop mobile) wireless networks as

a G or PFF game [1, 2], monotonicity would basically mean that by putting all nodes into a unique grand cluster the surplus of cooperation attains its maximum. Evidently, this is not the case, as the network is clustered (with an associated computational cost) precisely because partitioning the nodes enables for a better communication in view of the available technological infrastructure. One way to deal with this, while maintaining E as a fundamental axiom for characterizing solutions, is by letting Möbius inversion μ^f take value 0 on all partitions *non-finer* than that identified via the given global clustering algorithm [9, 12, 16, 21–23]. In terms of lattice functions, this is not much different from the above definition of solutions, as in both cases the basic modeling tool is Möbius inversion: taking its value to be identically 0 on a certain suitable region of the lattice formalizes the fact that some issues are autonomously addressed either by the agreed sharing criterion or else by the given communication technology. Formally, if P_t^* is the node partition defined by the chosen clustering algorithm at a generic time t , then $\mu^f(Q) = 0$ for all $Q \not\leq P_t^*$ yields

$$f(P^\top) = \sum_{Q \in \mathcal{P}^N} \mu^f(Q) = \sum_{Q \leq P_t^*} \mu^f(Q) = f(P_t^*). \quad (4)$$

While L provides expression (3), E additionally entails $\sum_{a \in L_A^N} \phi_a(\zeta_x) = 1$ for all basis elements or zeta games $\zeta_x, x \in L^N$. Concerning D (dummy), G and PFF games are crucially different from C games in view of the linear dependence characterizing geometric lattices [4, 33]. In fact, there are no dummy atoms in partition function-based games. To see this, consider any basis element ζ_P of G games and any atom $[ij] \in \mathcal{P}_A^N$ (\mathcal{P}_A^N being the $\binom{n}{2}$ -set of atoms of \mathcal{P}^N). If $[ij] \leq P$, then $\zeta_P(Q \vee [ij]) - \zeta_P(Q) = \begin{cases} 1 & \text{if } Q \not\geq [ij] \text{ and } P = Q \vee [ij] \\ 0 & \text{otherwise} \end{cases}$, while if $[ij] \not\leq P$, then $\zeta_P(Q \vee [ij]) - \zeta_P(Q) = \begin{cases} 1 & \text{if } Q \not\geq P \text{ and } Q \vee [ij] > P \\ 0 & \text{otherwise} \end{cases}$. This also applies to those $\binom{n}{2}$ zeta games $\zeta_{[ij]}$ for $[ij] \in \mathcal{P}_A^N$, as detailed hereafter.

Example 2. Let $N = \{1, 2, 3\}$, hence there are three atoms $[12], [13], [23]$ and five partitions: the bottom $P_\perp = 1|2|3$ (with vertical bar | separating blocks), the top $P^\top = [12] \vee [13] = [12] \vee [23] = [13] \vee [23] = [12] \vee [13] \vee [23]$, and the three atoms. For $\zeta_{[12]}$, of course $\zeta_{[12]}([12]) = 1$. However, $\zeta_{[12]}([13] \vee [23]) = \zeta_{[12]}(P^\top) = 1$. Even if game $\zeta_{[12]}$ requires atom $[12]$ to cooperate in order to generate the unit of TU, still such a unit also obtains through the cooperation of ‘only’ the other two atoms $[13]$ and $[23]$, since if they cooperate then $[12]$ ‘must’ cooperate too. \square

As detailed in the sequel, if the Shapley value of G games is translated according to the former equality in expression (1), then the unit of TU generated by zeta games has to be distributed over *all* atoms. Conversely, if the latter equality in expression (1) is employed, then the resulting solution is very different. Maintaining the axiomatic characterization of the Shapley value outlined in Sect. 2, in order to formalize S (symmetry) for G and PFF games recall that the class c^P (or type) of partitions $P \in \mathcal{P}^N$ [4, 39, 43] is $c^P = (c_1^P, \dots, c_n^P) \in \mathbb{Z}_+^n$, where

$c_k^P = |\{A : k = |A|, A \in P\}|$ is the number of k -cardinal blocks of P , $1 \leq k \leq n$. Analogously, the class of $(A, P) \in \mathcal{E}^N$ is $c^{A,P} = (c_0^{A,P}, c_1^{A,P}, \dots, c_n^{A,P}) \in \mathbb{Z}_+^{n+1}$, as embedded subset A may have cardinality $0 \leq |A| = c_0^{A,P} \leq n$ [32]. Hence $x \in L_A^N$ has class c^x , for $L_A^N \in \{\mathcal{P}^N, \mathcal{E}^N\}$. Also, $s(x) = |\{a : L_A^N \ni a \leq x\}|$ is the size of lattice elements x . The size of $(A, P) \in \mathcal{E}^N$ thus is $s(A, P) = |A| + s(P)$.

Definition 4. A solution ϕ satisfies S if $\phi_a(f) = \phi_{a'}(f)$ for any two atoms $a, a' \in L_A^N$ such that a bijection between $\{x : L_A^N \ni x \not\geq a\}$ and $\{y : L_A^N \ni y \not\geq a'\}$ satisfies (i) $f(x \vee a) = f(y \vee a')$ and (ii) $c^x = c^y$.

Since D cannot be employed to characterize solutions of G and PFF games in view of Example 5, these remaining axioms L, E and S do not yield uniqueness, but conversely define a whole class of solutions. In fact, given L and E, there is a continuum of alternative manners to also satisfy S, ranging from the following two extreme cases: (a) $\phi_a(\zeta_x) = \begin{cases} s(x)^{-1} & \text{if } a \leq x \\ 0 & \text{if } a \not\leq x \end{cases}$, and (b) $\phi_a(\zeta_x) = |L_A^N|^{-1}$ for all atoms $a \in L_A^N$ (and any ζ_x). Within this broad class of solutions satisfying L, E and S, a useful discriminant is the following FP *fixed-point* condition (which is a variation of the D axiom applying to C games).

Definition 5. A solution ϕ satisfies FP if $\mu^f(x) = 0$ for all $x \in L^N \setminus L_A^N$ entails $\phi(f) = f$.

Thus ϕ satisfies FP when it maps those games f whose Möbius inversion μ^f already lives only on atoms into themselves, i.e.

$$\phi_{a'}(\zeta_a) = \begin{cases} 1 & \text{if } a' = a \\ 0 & \text{if } a' \neq a \end{cases} \quad \text{for all } a, a' \in L_A^N. \quad (5)$$

In comparison with Example 5, this means that even if in G and PFF games every single atom a' may ‘swing’¹ in the zeta game ζ_a grounded on a fixed atom a , still FP requires a to exclusively get the whole unit of TU generated by ζ_a .

The remainder of this work is mostly concerned with the following two solutions ϕ^{CU}, ϕ^{SU} satisfying L, E and S.

Definition 6. The chain-uniform CU solution ϕ^{CU} is

$$\phi_a^{CU}(f) = \sum_{x \not\geq a} \frac{\kappa_x^a}{\kappa} \left(\frac{f(x \vee a) - f(x)}{s(x \vee a) - s(x)} \right), \quad (6)$$

where κ_x^a/κ is the ratio of the number κ_x^a of maximal chains (in L^N) meeting both x and $x \vee a$ to the total number κ of maximal chains, thus $\sum_{x \not\geq a} \frac{\kappa_x^a}{\kappa} = 1$.

The size-uniform SU solution ϕ^{SU} is

$$\phi_a^f(SU) = \sum_{x \geq a} \frac{\mu^f(x)}{s(x)}. \quad (7)$$

¹ See [5] on the Banzhaf value of C games.

For C games v , the CU and SU solutions coincide with the Shapley value, i.e. $\phi^{CU}(v) = \phi^{Sh}(v) = \phi^{SU}(v)$, where $\frac{1}{n^{\binom{n-1}{|A|}}} = \frac{|A|!(n-|A|-1)!}{n!}$ is the ratio of the number of maximal chains meeting both $A \subseteq N \setminus i$ and $A \cup i$ to the total number $n!$ of maximal chains, while $|A \cup i| - |A| = 1 = s(A \cup i) - s(A)$ is the size change. Conversely, for G and PFF games these two solutions are very different, and in particular ϕ^{SU} satisfies FP while ϕ^{CU} does not. Explicitely, for any $\zeta_y, y > x_{\perp}$,

$$\phi_a^{SU}(\zeta_y) = \begin{cases} \frac{1}{s(y)} & \text{if } a \leq y \\ 0 & \text{otherwise} \end{cases} \quad \text{for all } a \in L_{\mathcal{A}}^N. \quad (8)$$

Hence when $y = a$ expression (5) applies. On the other hand,

$$\begin{aligned} f(x \vee a) - f(x) &= \sum_{x \not\geq y \leq x \vee a} \mu^f(y) \text{ yields } \phi_a^{CU}(f) = \\ &= \sum_{x \not\geq a} \frac{\kappa_x^a}{\kappa} \left(\frac{\sum_{x \not\geq y \leq x \vee a} \mu^f(y)}{s(x \vee a) - s(x)} \right) = \sum_{y \in L^N} \mu^f(y) \left[\sum_{\substack{x \not\geq a \\ x \not\geq y \leq x \vee a}} \frac{\kappa_x^a}{\kappa[s(x \vee a) - s(x)]} \right]. \end{aligned}$$

Thus for any zeta game $\zeta_y, y > x_{\perp}$,

$$\phi_a^{CU}(\zeta_y) = \sum_{\substack{x \not\geq a \\ x \not\geq y \leq x \vee a}} \frac{\kappa_x^a}{\kappa[s(x \vee a) - s(x)]} \text{ for all } a \in L_{\mathcal{A}}, \text{ entailing}$$

$$\begin{aligned} \phi_a^{CU}(\zeta_a) &= \sum_{x \not\geq a} \frac{\kappa_x^a}{\kappa[s(x \vee a) - s(x)]} \leq \sum_{x \not\geq a} \frac{\kappa_x^a}{\kappa} = 1, \\ \phi_{a'}^{CU}(\zeta_a) &= \sum_{\substack{x \not\geq a, a' \\ x \vee a = x \vee a'}} \frac{\kappa_x^a}{\kappa[s(x \vee a) - s(x)]} \geq 0. \end{aligned}$$

For subset lattice $L^N = 2^N$, these two inequalities are satisfied as equalities, in that $\{x : x \not\geq a, a', x \vee a = x \vee a'\} = \emptyset$ for any two distinct atoms a, a' as well as $s(x \vee a) - s(x) = 1$, thus the SU and CU solutions coincide on these basis elements $\zeta_{\{i\}}, \{i\} \in 2^N$ of C games. Conversely, for G and PFF games the above inequalities are strict, i.e. $s(x \vee a) - s(x) > 1$ for most $x \not\geq a$ and $\{x : x \not\geq a, a', x \vee a = x \vee a'\} \neq \emptyset$ for any two distinct atoms a, a' . Hence the SU and CU solutions are different and the latter does not satisfy FP: $\phi_a^{CU}(\zeta_a) < 1$ and $\phi_{a'}^{CU}(\zeta_a) > 0$ for all $a, a' \in L_{\mathcal{A}}^N$. In practice, the only fixed points of the CU solution are also fixed points of the SU solution and take the form of linear functions $f = \alpha s, \alpha > 0$ of the size s . That is,

$$\phi_a^{CU}(\alpha s) = \sum_{x \not\geq a} \frac{\kappa_x^a}{\kappa} \left(\frac{\alpha[s(x \vee a) - s(x)]}{s(x \vee a) - s(x)} \right) = \alpha \sum_{x \not\geq a} \frac{\kappa_x^a}{\kappa} = \alpha \text{ for all } a \in L_{\mathcal{A}}^N.$$

3.1 Symmetric Games

An important class of games that may be useful for modeling certain communication networking systems consists of symmetric ones. In fact, ‘*partitions are of central importance in the study of symmetric functions, a class of functions that pervades mathematics in general*’ [44, p. 39] (see also [45, Ch. 5], [34], [43, Ch. 7, Vol. 2]). As detailed hereafter, for G and PFF symmetric games the CU and SU solutions coincide.

Definition 7. C games v , G games f and PFF games h are symmetric if

- $|A| = |B|$ entails $v(A) = v(B)$,
- $c_P^P = c_Q^Q$ entails $f(P) = f(Q)$,
- $c^{A,P} = c^{B,Q}$ entails $h(A, P) = h(B, Q)$.

Such v , f and h are indeed invariant under the action of the symmetric group S_n whose elements are the $n!$ permutations $\pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ of the indices or node identifiers $i \in N$. In particular, for every (number) partition $(\lambda_1, \dots, \lambda_n) \in \mathbb{Z}_+^n$ of (integer) n (i.e., $\sum_{1 \leq k \leq n} \lambda_k = n$), the number of distinct (set) partitions P of N with class $c_k^P = \lambda_k$, $1 \leq k \leq n$ is $n! \left(\prod_{1 \leq k \leq n} k!^{c_k^P} c_k^P \right)^{-1}$ [43, Vol. 1, p. 319]. In this view, if $f : L^N \rightarrow \mathbb{R}$ is a symmetric lattice function (where $L^N \in \{\mathcal{P}^N, \mathcal{E}^N\}$), then

$$\phi_a^{SU}(f) = \phi_{a'}^{SU}(f) \text{ or } \sum_{x \geq a} \frac{\mu^f(x)}{s(x)} = \sum_{y \geq a'} \frac{\mu^f(y)}{s(y)}$$

for any two atoms $a, a' \in L_A^N$. Accordingly, in view of E,

$$\sum_{a \in L_A^N} \phi_a^{SU}(f) = |L_A^N| \phi_a^{SU}(f) = f(x^\top) \Rightarrow \phi_a^{SU}(f) = \frac{f(x^\top)}{|L_A^N|}.$$

Analogously, a symmetric f yields $\phi_a^{CU}(f) = \phi_{a'}^{CU}(f)$ as

$$\sum_{x \geq a} \frac{\kappa_x^a}{\kappa} \left[\frac{f(x \vee a) - f(x)}{s(x \vee a) - s(x)} \right] = \sum_{y \geq a'} \frac{\kappa_y^{a'}}{\kappa} \left[\frac{f(y \vee a') - f(y)}{s(y \vee a') - s(y)} \right] \text{ for all } a, a' \in L_A,$$

$$\text{hence } \sum_{a \in L_A^N} \phi_a^{CU}(f) = |L_A^N| \phi_a^{CU}(f) = f(x^\top) \Rightarrow \phi_a^{CU}(f) = f(x^\top) / |L_A|.$$

Basic symmetric partition functions or G games (or PFF games, given the isomorphism exemplified in Sect. 4 below) are the size $s(P) = \sum_{A \in P} \binom{|A|}{2}$ and the rank $r(P) = n - |P| = \sum_{A \in P} (|A| - 1)$. Therefore,

$$\phi_a^{SU}(r) = \phi_a^{CU}(r) = \frac{r(x^\top)}{|L_A^N|} = \phi_a^{CU}(\zeta_a) \leq 1 = \phi_a^{SU}(\zeta_a)$$

for all atoms a , with strict inequality if $L^N \in \{\mathcal{P}^N, \mathcal{E}^N\}$. More generally, any f which is itself a function of the rank or a function of the size (see Example 2) is of course symmetric, and thus satisfies $\phi_a^{CU}(f) = f(x^\top) / |L_A^N| = \phi_a^{SU}(f)$.

4 Isomorphism Between G and PFF Games

This section details the computations for the SU and CU solutions of G and PFF games, while also providing a simple example of G games on player set $N = \{1, 2, 3\}$ together with the isomorphic PFF game on player set $N = \{1, 2\}$.

Solutions $\phi(f) = \{\phi_{[ij]}(f) : \{i, j\} \in N_2\} \in \mathbb{R}^{\binom{n}{2}}$ of G games $f : \mathcal{P}^N \rightarrow \mathbb{R}$ are $\binom{n}{2}$ -vectors. Regarded as lattice functions themselves, these solutions have Möbius inversion $\mu^{\phi(f)}([ij]) = \phi_{[ij]}(f)$ living only on atoms $[ij]$, and thus satisfy $\phi(f)(P) = \sum_{[ij] \leq P} \phi_{[ij]}(f)$ for all $P \in \mathcal{P}^N$.

There are $\kappa = \binom{n}{2} \binom{n-1}{2} \binom{n-2}{2} \cdots \binom{2}{2} = \frac{n!(n-1)!}{2^{n-1}}$ maximal chains of partitions. The number of such maximal chains meeting any P , with class c^P , is

$$\left[\prod_{1 \leq k \leq n} \left(\frac{k!(k-1)!}{2^{k-1}} \right)^{c_k^P} \right] \frac{|P|!(|P|-1)!}{2^{|P|-1}},$$

where the product on the left is the number of maximal chains in $[P_\perp, P]$, this latter segment or interval [39] being isomorphic to $\times_{A \in P} \mathcal{P}^{|A|}$, while the fraction on the right is the number of maximal chains in segment $[P, P^\top]$, isomorphic to $\mathcal{P}^{|P|}$. Here \mathcal{P}^k denotes the lattice of partitions of a k -set. Thus for an atom $[ij]$ and a partition $P \not\ni [ij]$, the number $\kappa_P^{[ij]}$ of maximal chains meeting both P and $P \vee [ij]$ is

$$\begin{aligned} \kappa_P^{[ij]} &= \left[\prod_{1 \leq k \leq n} \left(\frac{k!(k-1)!}{2^{k-1}} \right)^{c_k^P} \right] \frac{|P \vee [ij]|!(|P \vee [ij]|-1)!}{2^{|P \vee [ij]|-1}} = \\ &= \left[\prod_{1 \leq k \leq n} (k!(k-1)!)^{c_k^P} \right] \frac{(|P|-1)!(|P|-2)!}{2^{|P|-2+\sum_{1 \leq k \leq n} c_k^P(k-1)}} = \\ &= \left[\prod_{1 \leq k \leq n} (k!(k-1)!)^{c_k^P} \right] \frac{(|P|-1)!(|P|-2)!}{2^{n-2}}, \text{ yielding} \end{aligned}$$

$$\frac{\kappa_P^{[ij]}}{\kappa} = 2 \left[\prod_{1 \leq k \leq n} (k!(k-1)!)^{c_k^P} \right] \frac{(|P|-1)!(|P|-2)!}{n!(n-1)!}.$$

Thus the CU solution is $\phi_{[ij]}^{CU}(f) = \sum_{P \not\ni [ij]} \frac{\kappa_P^{[ij]}}{\kappa} \left[\frac{f(P \vee [ij]) - f(P)}{s(P \vee [ij]) - s(P)} \right],$

while the SU solution is $\phi_{[ij]}^{SU}(f) = \sum_{P \geq [ij]} \frac{\mu^f(P)}{s(P)}.$

Now for $N = \{1, 2, 3\}$, consider the G game $f = \zeta_{[12]}$ given by basis element or zeta game $\zeta_{[12]}$ (see Example 5). That is $\zeta_{[12]}(P) = 1$ if $P \geq [12]$ and 0 otherwise, where $[12], [13]$ and $[23]$ are the three atoms. Then,

$$\phi_{[ij]}^{SU}(\zeta_{[12]}) = \begin{cases} 1 & \text{if } [ij] = [12] \\ 0 & \text{if } [ij] \neq [12] \end{cases} \quad \text{and} \quad \phi_{[ij]}^{CU}(\zeta_{[12]}) = \begin{cases} 2/3 & \text{if } [ij] = [12] \\ 1/6 & \text{if } [ij] \neq [12] \end{cases},$$

$$\text{while } \phi_{[ij]}^{CU}(r) = \phi_{[ij]}^{SU}(r) = 2/3 = \frac{r(P^\top)}{\binom{n}{2}} = \phi_{[ij]}^{CU}(\zeta_{[ij]}) \text{ for all } [ij].$$

The remainder of this section is based on the approach to the geometric lattice of embedded subsets $(\mathcal{E}^N, \sqcap, \sqcup)$ defined in [32]. In particular, join \sqcup obtains via a closure operator [4] $cl : 2^N \times \mathcal{P}^N \rightarrow 2^N \times \mathcal{P}^N$ that cannot be detailed here for reasons of space and mostly because it surely does not fit a contribution that is aimed to be useful in the area of game-theoretical models of cooperative communication networking. Clarified this, if PFF games h are looked at as functions defined on \mathcal{E}^N , with solutions $\phi(h) = \{\phi_a(h) : a \in \mathcal{E}_{\mathcal{A}}^N\} \in \mathbb{R}^{\binom{n+1}{2}}$ being lattice functions themselves, then these latter have Möbius inversion $\mu^{\phi(h)}(a) = \phi_a(h)$ living only on atoms $a \in \mathcal{E}_{\mathcal{A}}^N$, i.e. $\phi(h)(A, P) = \sum_{a \sqsubseteq (A, P)} \phi_a(h)$ for all $(A, P) \in \mathcal{E}^N$.

Lattices $(\mathcal{P}^N, \wedge, \vee)$ and $(\mathcal{E}^N, \sqcap, \sqcup)$ are characterized by these isomorphisms: $[(\emptyset, P_\perp), (N, P^\top)] = \mathcal{E}^n \cong \mathcal{P}^{n+1}$ and $[(\emptyset, P_\perp), (A, P)] \cong \mathcal{E}^{|A|} \times (\times_{B \in P^{A^c}} \mathcal{P}^{|B|})$ and $[(A, P), (N, P^\top)] \cong \mathcal{P}^{|P|}$ if $A \neq \emptyset$, while $[(A, P), (N, P^\top)] \cong \mathcal{E}^{|P|}$ if $A = \emptyset$, where $P^{A^c} = \{A^c \cap B : P \ni B, \emptyset \neq A^c \cap B\}$ is the partition of $A^c \neq \emptyset$ induced by $P = \{B_1, \dots, B_{|P|}\}$ and \mathcal{E}^k is the geometric lattice of embedded subsets of a k -set. Accordingly, the number of maximal chains in \mathcal{E}^N (i.e. from (\emptyset, P_\perp) to (N, P^\top)) is $\kappa = \binom{n+1}{2} \binom{n}{2} \binom{n-1}{2} \cdots \binom{2}{2} = \frac{(n+1)!n!}{2^n}$, and the number of those meeting $(A, P) \in \mathcal{E}^N$ is

$$\frac{(|A| + 1)!|A|!|P|!(|P| - 1)!}{2^{|A|}2^{|P|-1}} \prod_{k=1}^{n-|A|} \left(\frac{k!(k-1)!}{2^{k-1}} \right)^{c_k^{P^{A^c}}} \quad \text{if } A \neq \emptyset, \text{ and}$$

$$\frac{(|P| + 1)!|P|!}{2^{|P|}} \prod_{k=1}^n \left(\frac{k!(k-1)!}{2^{k-1}} \right)^{c_k^P} \quad \text{if } A = \emptyset.$$

For an atom $a \in \mathcal{E}_{\mathcal{A}}^N$ and an embedded subset $(A, P) \not\supseteq a$, the number of maximal chains meeting both (A, P) and the embedded subset $cl((A, P) \sqcup a)$ obtained as the (above-mentioned) join of (A, P) and a thus is

$$\kappa_{(A, P)}^a = \frac{(|A| + 1)!|A|!|P|!(|P| - 1)!(|P| - 2)!}{2^{|A|}2^{|P|-2}} \prod_{k=1}^{n-|A|} \left(\frac{k!(k-1)!}{2^{k-1}} \right)^{c_k^{P^{A^c}}} =$$

$$= \frac{(|A| + 1)!|A|!(|P| - 1)!(|P| - 2)!}{2^{n-1}} \prod_{k=1}^{n-|A|} ((k!(k-1)!)^{c_k^{P^{A^c}}}) \quad \text{if } A \neq \emptyset, \text{ and}$$

$$\kappa_{(A,P)}^a = \frac{|P|!(|P|-1)!}{2^{|P|-1}} \prod_{k=1}^n \left(\frac{k!(k-1)!}{2^{k-1}} \right)^{c_k^P} = \frac{|P|!(|P|-1)!}{2^{n-1}} \prod_{k=1}^n (k!(k-1)!)^{c_k^P}$$

$$\text{if } A = \emptyset. \text{ Or } \frac{\kappa_{(A,P)}^a}{\kappa} = \frac{2(|P|-1)!(|P|-2)!(|A|+1)!|A|!}{(n+1)!n!} \prod_{k=1}^{n-|A|} (k!(k-1)!)^{c_k^P}$$

$$\text{if } A \neq \emptyset \text{ and } \frac{\kappa_{(A,P)}^a}{\kappa} = \frac{2|P|!(|P|-1)!}{(n+1)!n!} \prod_{k=1}^n (k!(k-1)!)^{c_k^P} \text{ if } A = \emptyset.$$

Therefore, the CU and SU solutions of PFF game h are respectively

$$\begin{aligned} \phi_a^{CU}(h) &= \sum_{(A,P) \not\sqsupseteq a} \frac{\kappa_{(A,P)}^a}{\kappa} \left[\frac{h(cl((A,P) \sqcup a)) - h(A,P)}{s(cl((A,P) \sqcup a)) - s(A,P)} \right] \\ \phi_a^{SU}(h) &= \sum_{(A,P) \sqsupseteq a} \frac{\mu^h(A,P)}{s(A,P)}. \end{aligned}$$

For $N = \{1, 2\}$, consider the PFF game $h = \zeta_{(\emptyset, [12])}$ given by the basis element or zeta game $\zeta_{(\emptyset, [12])}$, i.e. $\zeta_{(\emptyset, [12])}(A, P) = 1$ if $(A, P) \sqsupseteq (\emptyset, [12])$ and 0 otherwise (for all $(A, P) \in \mathcal{E}^N$), and where $(1, 1|2)$ and $(2, 1|2)$ and $(\emptyset, [12]) = (\emptyset, 12)$ are the three atoms, with vertical bar $|$ separating blocks. That is, $1|2 = P_\perp$ while $12 = [12] = P^\top$, where $(\emptyset, 1|2) \sqsubset (1, 1|2), (2, 1|2), (\emptyset, 12) \sqsubset (\{1, 2\}, 12)$, with bottom $(\emptyset, P_\perp) = (\emptyset, 1|2)$ and top $(\{1, 2\}, 12) = (N, P^\top)$. Then the corresponding SU and CU solutions are:

$$\begin{aligned} \phi_a^{SU}(\zeta_{(\emptyset, [12])}) &= \begin{cases} 1 & \text{if } a = (\emptyset, [12]) \\ 0 & \text{if } a = (1, 1|2) \text{ or } a = (2, 1|2) \end{cases}, \\ \phi_a^{CU}(\zeta_{(\emptyset, [12])}) &= \begin{cases} 2/3 & \text{if } a = (\emptyset, [12]) \\ 1/6 & \text{if } a = (1, 1|2) \text{ or } a = (2, 1|2) \end{cases}, \end{aligned}$$

and $\phi_a^{CU}(r) = \phi_a^{SU}(r) = 2/3 = \frac{r(N, P^\top)}{\binom{3}{2}} = \phi_a^{CU}(\zeta_a)$ for all atoms $a \in \mathcal{E}_A^N$, with $r(A, P) = r(P) + \min\{|A|, 1\}$ as the rank function $r : \mathcal{E}^N \rightarrow \mathbb{Z}_+$.

5 The Core and Additivity

The core $\mathcal{C}(v)$ of C games v is often introduced as the set of point-valued solutions $\phi(v)$ that no coalition $A \in 2^N$ can block, meaning that under sharing rule ϕ every A must receive an amount of TU $\phi(v)(A) = \sum_{i \in A} \phi_i(v) \geq v(A)$, with equality for the grand coalition $A = N$. In fact, if $\phi(v)(A) < v(A)$, then why should coalition members $i \in A$ cooperate with non-members $j \in A^c$? Looking at the n values

of Möbius inversion $\mu^{\phi(v)}$, namely at the n shares $(\phi_1(v), \dots, \phi_n(v)) \in \mathbb{R}^n$, the core is a convex and possibly empty subset $\mathcal{C}(v) \subset \mathbb{R}^n$. It is the main set-valued solution concept, and the necessary and sufficient conditions for non-emptiness $\mathcal{C}(v) \neq \emptyset$ are the well-known ‘Shapley-Bondareva conditions’ (see, for instance, [1, p. 210]). On the other hand, supermodularity (also termed ‘convexity’ [36]) of v , i.e. $v(A \cup B) + v(A \cap B) \geq v(A) + v(B)$ for all $A, B \in 2^N$, is a sufficient but not necessary condition for non-emptiness $\mathcal{C}(v) \neq \emptyset$, and total positivity $\mu^v(A) \geq 0$ for all $A \in 2^N$ entails supermodularity (see Sect. 2). If v is supermodular, then the extreme points of (convex polyhedron) $\mathcal{C}(v)$ are those solutions obtained by rewarding each $i \in N$ with marginal contribution $v(A_k \cup i) - v(A_k)$ where $A_k, 0 \leq k \leq n$ is a maximal chain, i.e. $A_{k+1} = A_k \cup i$. In such a case, the Shapley value $\phi^{Sh}(v)$ defined by expression (1) is the center of the core, in that ϕ^{Sh} is the ‘uniform’ convex combination of the $n!$ solutions associated with maximal chains as just specified (see [36]). In multi-agent systems, $\mathcal{C}(v)$ is important precisely because coalitions have no incentive to oppose any worth/cost-sharing rule that is known to be in the core [24, 25, 35, 46].

When all of this is to be translated in terms G and PFF games, a first conceptual observation is that every partition involves all players, thus saying that ‘a partition of players can block a sharing criterion’ (or, more generally, an outcome) does not seem to allow for a straightforward interpretation. In fact, partition function-based TU games model situations where all players cooperate, in some way. Specifically, every partition formalizes a distinct form of global cooperation, although the bottom P_\perp seems to unambiguously represent the case where everyone stands alone.² Furthermore, as already explained in Sect. 3 in terms of monotonicity, the coarsest partition P^\top appears to hardly correspond to the ‘best form of cooperation’, especially insofar as technology leads cooperative communication networking to realize through clusters of nodes. On the other hand, if the possibilities given by the technological infrastructure are suitably translated by means of a Möbius inversion living only on a proper interval (or segment) of the lattice (see expression (4)), then P^\top does formalize the required overall agreement on how to distribute the cost of network maintainance (or equivalently the worth of its existence), toward optimal global functioning. Despite these premises, still from a geometric perspective the cores $\mathcal{C}(f) \subset \mathbb{R}^{\binom{n}{2}}, \mathcal{C}(h) \subset \mathbb{R}^{\binom{n+1}{2}}$ of G and PFF games f, h are well-defined in terms of the novel solution concept proposed here. Maintaining the above notation (L^N, \wedge, \vee) for a lattice $L^N \in \{\mathcal{P}^N, \mathcal{E}^N\}$ in order to deal with both these games at once, Definition 1 in Sect. 3 leads to obtain the core $\mathcal{C}(f) \subset \mathbb{R}^{|L_A^N|}$ of games $f : L^N \rightarrow \mathbb{R}$ as the possibly empty convex polyhedron consisting of those solutions $\phi(f)$ such that $\phi(f)(x) = \sum_{a \leqslant x} \phi_a(f) \geq f(x)$ for all elements $x \in L^N$, with equality for the top x^\top . Now, concerning both (i) the analog of the necessary and sufficient Shapley-Bondareva conditions for non-emptiness, and (ii) supermodularity as a sufficient but not necessary such condition, the main difference with respect to C games is due, once again, to linear dependence, which characterizes

² See [27] on how to deal with a bottom partition whose worth is $\neq 0$.

geometric (non-distributive) lattices [4, 33]. In fact, while noticing that in the CU and SU solutions, respectively, marginal contributions and the values of Möbius inversion appear divided by the size of lattice elements, it must be considered that the size is a totally positive lattice function (as its Möbius inversion takes value 1 on atoms and 0 elsewhere). Therefore, in order for $\mathcal{C}(f)$ to be non-empty, it is not sufficient that f is supermodular, i.e. $f(x \vee y) + f(x \wedge y) \geq f(x) + f(y)$ for all $x, y \in L^N$. Conversely, f has to quantify synergies minimally as great as those quantified by the size itself.

Example 3. For $N = \{1, 2, 3\}$ consider the supermodular symmetric G game $f : \mathcal{P}^N \rightarrow \mathbb{R}$ defined by $f(P_\perp) = 0$, $f([ij]) = 1$ for $1 \leq i < j \leq 3$ and $f(P^\top) = 2$, where $f(P \vee Q) + f(P \wedge Q) \geq f(P) + f(Q)$ for all $P, Q \in \mathcal{P}^N$ is easily checked: if $P \geq Q$, then equality holds, while the only remaining case is when both $P, Q \in \mathcal{P}_A^N$ are atoms, but $f([12] \vee f([13])) + f([12] \wedge [13]) = f([12]) + f([13]) = 2$ (for instance), i.e. equality holds as well. However, f is *not* totally positive: $\mu^f(P^\top) = 2 - 1 - 1 - 1 = -1 < 0$, and in particular $f(P^\top) < s(P^\top) = 3$. Thus $\mathcal{C}(f) = \emptyset$, as $\phi_{[12]}(f), \phi_{[13]}(f), \phi_{[23]}(f)$ cannot satisfy both $\phi_{[ij]}(f) \geq 1$, $1 \leq i < j \leq 3$ and $\phi_{[12]}(f) + \phi_{[13]}(f) + \phi_{[23]}(f) = 2$. \square

The necessary and sufficient conditions for non-emptiness of the core of G and PFF games deserve separate treatment.

5.1 Additively Separable Partition Functions

Partition functions $f : \mathcal{P}^N \rightarrow \mathbb{R}$ are additively separable when a set function $v : 2^N \rightarrow \mathbb{R}$ satisfies $f(P) = \sum_{A \in P} v(A)$ for all $P \in \mathcal{P}^N$ [27, 38], and are also sometimes termed, more simply, ‘additive’. They appear in a variety of settings, ranging from combinatorial optimization problems [47] to community/module detection in complex networks [37]. These partition functions admit in fact a continuum of additively separating set functions, as exemplified hereafter.

Example 4. Additively separating the rank $r(P) = n - |P| = \sum_{A \in P} (|A| - 1)$ of partitions, which is immediately seen to be additively separated by the symmetric set function v (see Sect. 3.1) defined by $v(A) = |A| - 1$ for all $A \in 2^N$, whose Möbius inversion takes values $\mu^v(\emptyset) = -1$, $\mu^v(A) = 1$ if $|A| = 1$ and $\mu^v(A) = 0$ if $|A| > 1$. However, $r(\cdot)$ may be checked to be also additively separated by the (again symmetric) set function v' with Möbius inversion $\mu^{v'}(A) = 0$ if $|A| \leq 1$ and $\mu^{v'}(A) = (-1)^{|A|}$ if $|A| > 1$, thus $v'(A) = 0$ if $|A| \leq 1$ and $v'(A) = 1$ if $|A| = 2$, $v'(A) = 2$ if $|A| = 3$, $v'(A) = 3$ if $|A| = 4$, i.e. $\mu^{v'}(A) = |A| - 1 - \sum_{B \subset A} \mu^v(B)$. \square

Example 5. Additively separating the size $s(P) = \sum_{A \in P} \binom{|A|}{2}$ of partitions, which is immediately seen to be additively separated by the symmetric set function v defined by $v(A) = \binom{|A|}{2}$ for all $A \in 2^N$, with Möbius inversion $\mu^v(A) = 0$ if $|A| \neq 2$ and $\mu^v(A) = 1$ if $|A| = 2$. However, $s(\cdot)$ is also additively separated by (symmetric) v' with Möbius inversion $\mu^{v'}(A) = \begin{cases} (-1)^{|A|+1} & \text{if } |A| \neq 2 \\ 0 & \text{if } |A| = 2 \end{cases}$, hence

$\mu^{v'}(\emptyset) = -1$, $\mu^{v'}(A) = 1$ if $|A| = 1$, $\mu^{v'}(A) = 0$ if $|A| = 2$, $\mu^{v'}(A) = 1$ if $|A| = 3$, $\mu^{v'}(A) = -1$ if $|A| = 4$, $\mu^{v'}(A) = 1$ if $|A| = 5$, $\mu^{v'}(A) = -1$ if $|A| = 6$ and so on according to recursion $\mu^{v'}(A) = \binom{|A|}{2} - \sum_{B \subset A} \mu^{v'}(B)$. \square

In order to briefly generalize these examples, let $\mathbf{AS}(f) \subset \mathbb{R}^{2^n}$ denote the convex, possibly empty set³ of set functions v that additively separate a given partition function f , i.e. $v, v' \in \mathbf{AS}(f), \alpha \in [0, 1] \Rightarrow [\alpha v + (1 - \alpha)v'] \in \mathbf{AS}(f)$, as $\sum_{A \in P} [\alpha v(A) + (1 - \alpha)v'(A)] = \alpha \sum_{A \in P} v(A) + (1 - \alpha) \sum_{A \in P} v'(A)$, and by assumption $\sum_{A \in P} v(A) = \sum_{A \in P} v'(A)$ for all $P \in \mathcal{P}^N$. Emptiness $\mathbf{AS}(f) = \emptyset$ corresponds to a non-additively separable f . For $v \in \mathbf{AS}(f)$, any $v' \in \mathbf{AS}(f)$ obtains recursively as follows: (i) $n\mu^{v'}(\emptyset) + \sum_{i \in N} \mu^{v'}(\{i\}) = n\mu^v(\emptyset) + \sum_{i \in N} \mu^v(\{i\})$, entailing $\sum_{i \in N} v'(\{i\}) = f(P_\perp) = \sum_{i \in N} v(\{i\})$; (ii) $\mu^{v'}(A) = v(A) - \sum_{B \subset A} \mu^{v'}(B)$ for all $A, |A| > 1$, entailing $v'(A) = v(A)$ for all $A, |A| > 1$. A similar argument applies to the set $\mathbf{AS}(h)$ of set functions v additively separating PFF games h , i.e. such that $h(A, P) = v(A) + \sum_{B \in P} v(B)$ for all $(A, P) \in \mathcal{E}^N$ [32].

Solutions of additively separable G and PFF games can be approached as traditional solutions of C games (i.e. via the Shapley value in expression (1), see [27, 31]). But if $\mathbf{AS}(f) \neq \emptyset \neq \mathbf{AS}(h)$, then the CU and SU solutions proposed above provide novel criteria for distributing either the generated TU or else the costs of system maintenance.

6 Conclusion

In a time-varying communication network $G^t = (N, E^t)$, any pair $\{i, j\} \in N_2$ of nodes may be an edge, i.e. a vehicle for exchanging information, over some period $\Delta t > 0$, and if an edge has never appeared over the whole history (up to some ‘present’ t), then it is simply to be regarded as one that has made no contribution (thus far). In this view, each pair of nodes receives a (possibly null) share of the surplus generated by cooperation. Thus the idea is that exhaustive information about network topology and traffic is constantly collected, in a distributed and local manner, enabling to use such an information for each period $t - 1 \rightarrow t$ in order to reward nodes at time t . A constant node set $N = \{1, \dots, n\}$ is also recognized to be employed here for expositional convenience, as not only at the beginningning of each period a new game starts being played, but new nodes may also enter at any time t .

In order to exemplify what form a fixed point of the SU solution of G games may take, consider the case where the surplus generated by cooperation is simply assumed to be quantified by the volume of data traffic over the whole (clustered) network during each time period. In particular, if the protocol requires redundant transmissions, let all transmissions contribute to the generated surplus independently from redundancy. In other terms, the (periodical) worth of global cooperation is the sum over all edges of the traffic that occurs through

³ $v(\emptyset) = \mu^v(\emptyset)$ being chosen arbitrarily, any $\mathbf{AS}(f) \neq \emptyset$ is not bounded.

them. Then, the share assigned to each edge by the SU solution is precisely the volume of data that is transferred through that edge, and thus such a G game is a fixed point of the SU solution (mapping). In practice, this would mean that an arbitrary unit of data transfer (say 1 MB) is equivalent to an arbitrary unit of TU (say a token giving temporary privileges).

Concerning how to divide the share of each edge between its endnodes, in graph-restricted C games [48, 49] players must be rewarded by taking into account only their position in the network and their marginal contributions to coalitions. In other terms, such players do not have different roles for network functioning, and thus the shares of edges seem best equally divided between endnodes (as suggested in [26] for ‘communication situations’). However, in many networking systems different nodes may well play different roles depending on their (time-varying) energy constraints and/or computational capabilities. In the simplest case, some of them are cluster heads while the remaining ones are not. Similarly, in cognitive radio networks players may be either primary users or else secondary users. This suggests that in several communication networking systems the shares distributed over edges can be next divided between nodes in more sophisticated manners than just via equal splitting.

A final but seemingly very important remark concerns Möbius inversion, which was defined to provide “*the combinatorial analog of the fundamental theorem of calculus*” [39]. Roughly speaking, Möbius inversion may be regarded as the ‘derivative’ of lattice (or more generally poset) functions, and constitutes a very useful tool for game-theoretical modeling. For instance, apart from G and PFF games, several communication networking systems can be modeled by means of graph-restricted C games [1, 2], and in particular the Myerson value [48] appears to be an important solution for such settings. In fact, as outlined in [1, Sec. 7.5.2] and [2, p. 23], the Myerson value is well-known to be the Shapley value (given by expression (1)) of a novel (i.e. graph-restricted) C game. However, it may be worth emphasizing that such a novel game v/G is characterized as follows: (i) it coincides with the original (unrestricted) C game v on those coalitions spanning (or inducing [50]) a connected subgraph, and (ii) its Möbius inversion $\mu^{v/G}$ takes value 0 on the remaining coalitions [49]. Hence graph-restricted C game v/G (also termed ‘coalitional graph game’ in [1, 2]) has Möbius inversion $\mu^{v/G}$ living only on connected coalitions. Insofar as applications are concerned, this means that the Myerson value may be computed by means of the second equality in expression (1), once the non-zero values of $\mu^{v/G}$ are recursively determined. Hopefully, the present paper may thus contribute to fostering the use of Möbius inversion (of all types of TU cooperative games) for modeling collaborative communication systems.

6.1 Future Work

Definition 1 in Sect. 3 enables to reconsider the whole theory of C games in terms of G and PFF games. As for the core (see Sect. 5), the well-known Shapley-Bondareva non-emptiness conditions may be paralleled with focus on what changes must be introduced in order to take into account the size of lattice

elements. In this view, it seems that dividing the main expressions for the CU and SU solutions by the size (see above) is already in compliance with the comprehensive approach (to the core of C games) relying on ‘concavification’ [51]. From a more general perspective, C games are commonly looked at as pseudo-Boolean functions, thereby obtaining the Shapley, Banzhaf and other values through the gradient of the polynomial multilinear extension of such functions [52], [5, pp. 139–151]. A challenging task thus is to reproduce the same entire pseudo-Boolean framework for G and PFF games.

References

1. Han, Z., Niyato, D., Saad, W., Başar, T., Hjørungnes, A.: Game Theory for Wireless and Communication Networks. Cambridge University Press, Cambridge (2012)
2. Saad, W., Han, Z., Debbah, M., Hjørungnes, A., Başar, T.: Coalitional game theory for communication networks: a tutorial, pp. 1–26. [arXiv:0905.4057v1](https://arxiv.org/abs/0905.4057v1) (2009)
3. Slikker, M.: Coalition formation and potential games. *Games Econ. Behav.* **37**, 436–448 (2001)
4. Aigner, M.: Combinatorial Theory. Springer, Berlin (1997)
5. Roth, A. (ed.): The Shapley Value-Essays in Honor of Lloyd S. Shapley. Cambridge University Press, Cambridge (1988)
6. Akyildiz, I., Lo, B., Balakrishnan, R.: Cooperative spectrum sensing in cognitive radio networks: a survey. *Phys. Commun.* **4**, 40–62 (2011)
7. Laneman, J., Wornell, G.: Distributed space-time-coded protocols for exploiting cooperative diversity in wireless network. *IEEE Trans. Inf. Theory* **49**, 2415–2425 (2003)
8. Letaief, K., Zhang, W.: Cooperative spectrum sensing. In: Hossain, E., Bhargava, V. (eds.) Cognitive Wireless Communication Networks, pp. 115–138. Springer, Berlin (2007)
9. Zhou, Z., Zhou, S., Cui, S., Cui, J.H.: Energy-efficient cooperative communication in clustered wireless sensor networks. *IEEE Trans. Veh. Technol.* **57**, 3618–3628 (2008)
10. Al-Karaki, J.L., Kamal, A.E.: Routing techniques in wireless sensor networks: a survey. *IEEE Wirel. Commun.* **11**, 6–28 (2004)
11. Cai, J., Pooch, U.: Allocate fair payoff for cooperation in wireless ad hoc networks using Shapley value. In: Proceedings of IPDPS 04, pp. 219–226 (2004)
12. Cavalcanti, D., Agrawal, D., Cordero, C., Xie, B., Kumar, A.: Issues in integrating cellular networks, WLANs and MANETs: a futuristic heterogeneous wireless network. *IEEE Wirel. Commun.* **12**, 30–41 (2005)
13. Cho, J.H., Swami, A., Chen, I.R.: A survey on trust management for mobile ad hoc networks. *IEEE Commun. Surv. Tutorials* **13**(4), 562–583 (2011)
14. Chow, C.Y., Leong, H.V., Chan, A.T.S.: GroCoca: group-based peer-to-peer cooperative caching in mobile environment. *IEEE J. SAC* **25**(1), 179–191 (2007)
15. Huang, X., Zhai, H., Fang, Y.: Robust cooperative routing protocol in mobile wireless sensor networks. *IEEE Trans. Wirel. Commun.* **7**(12), 5278–5285 (2008)
16. Ye, M., Li, C., Chen, G., Wu, J.: An energy efficient clustering scheme in wireless sensor networks. *Ad Hoc Sens. Wirel. Netw.* **3**, 99–119 (2007)
17. Nandan, A., Das, S., Pau, G., Gerla, M., Sanadidi, M.: Co-operative downloading in vehicular ad-hoc wireless networks. In Proceedings of WONS 05, pp. 32–41 (2005)

18. Vardhe, K., Reynolds, D., Woerner, B.: Joint power allocation and relay selection for multiuser cooperative communication. *IEEE Trans. Wirel. Commun.* **9**(4), 1255–1260 (2010)
19. Wang, B., Han, Z., Liu, K.: Distributed relay selection and power control for multiuser cooperative communication networks using buyer/seller game. In: IEEE INFOCOM 2007 Proceedings, pp. 544–552 (2007)
20. Wang, T., Giannakis, G.B.: Complex field network coding for multiuser cooperative communications. *IEEE J. SAC* **26**(3), 561–571 (2008)
21. Younis, O., Fahmy, S.: HEED a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. *IEEE Trans. Mob. Comput.* **3**(4), 366–379 (2004)
22. Younis, O., Krunz, M., Ramasubramanian, S.: Node clustering in wireless sensor networks: recent developments and deployment challenges. *IEEE Netw.* **20**, 20–25 (2006)
23. Fitzek, F.H.P., Katz, M.D. (eds.): Cooperation in Wireless Networks: Principles and Applications—Real Egoistic Behavior is to Cooperate!. Springer, Berlin (2006)
24. Saad, W., Han, Z., Başar, T., Debbah, M., Hjørungnes, A.: Coalition formation games for collaborative spectrum sensing. *IEEE Trans. Veh. Technol.* **60**, 276–297 (2011)
25. Saad, W., Han, Z., Zheng, R., Hjørungnes, A., Başar, T., Poor, H.V.: Coalitional games in partition form for joint spectrum sensing and access in cognitive radio networks. *IEEE J. Sel. Top. Sig. Proc.* **6**, 195–209 (2012)
26. Borm, P., Owen, G., Tijs, S.: On the position value for communication situations. *SIAM J. Discrete Math.* **5**, 305–320 (1992)
27. Gilboa, I., Lehrer, E.: Global games. *Int. J. Game Theory* **20**, 120–147 (1990)
28. Rossi, G.: Worth-sharing through Möbius inversion. *Homo Economicus* **24**, 411–433 (2007)
29. Thrall, R.M., Lucas, W.F.: n -person games in partition function form. *Naval Res. Logistic Q.* **10**, 281–298 (1963)
30. Grabisch, M., Funaki, Y.: A coalition formation value for games in partition function form. *Eur. J. Oper. Res.* **221**(1), 175–185 (2012)
31. Myerson, R.: Values of games in partition function form. *Int. J. Game Theory* **6**, 23–31 (1977)
32. Rossi, G.: The geometric lattice of embedded subsets, pp. 1–17. [arXiv: 1612.05814](https://arxiv.org/abs/1612.05814) (2017)
33. Whitney, H.: On the abstract properties of linear dependence. *Am. J. Math.* **57**, 509–533 (1935)
34. Rosas, M.H., Sagan, B.E.: Symmetric functions in noncommuting variables. *Trans. AMS* **358**, 215–232 (2006)
35. Conitzer, V., Sandholm, T.: Complexity of constructing solutions in the core based on synergies among coalitions. *Artif. Intell.* **170**, 607–619 (2006)
36. Shapley, L.S.: Cores of convex games. *Int. J. Game Theory* **1**(1), 11–26 (1971)
37. Fortunato, S.: Community detection in graphs. *Phys. Rep.* **486**(3–5), 75–174 (2010)
38. Gilboa, I., Lehrer, E.: The value of information—an axiomatic approach. *J. Math. Econ.* **20**(5), 443–459 (1991)
39. Rota, G.C.: On the foundations of combinatorial theory I: theory of Möbius functions. *Z. Wahrscheinlichkeitsrechnung Verw. Geb.* **2**, 340–368 (1964)
40. Weber, R.J.: Probabilistic values for games. In: A.E. Roth (ed.) *The Shapley Value—Essays in Honor of Lloyd S. Shapley*, pp. 101–119. Cambridge University Press, Cambridge (1988)
41. Grabisch, M.: The lattice of embedded subsets. *Discrete Appl. Math.* **158**, 479–488 (2010)

42. Shapley, L.S.: A value for n -person games. In: Kuhn, H., Tucker, A.W. (eds.) Contributions to the Theory of Games, pp. 307–317. Princeton University Press, Princeton (1953)
43. Stanley, R.: Enumerative Combinatorics, 2nd edn. Cambridge University Press, Cambridge (2012)
44. Knuth, D.E.: Generating all combinations and partitions. The Art of Computer Programming, vol. 4, no. 3, pp. 1–150. Addison-Wesley (2005)
45. Kung, J.P.S., Rota, G.C., Yan, C.H. (eds.): Combinatorics: The Rota Way. Cambridge University Press, Cambridge (2009)
46. Rahwan, T., Jennings, N.: An algorithm for distributing coalitional value calculations among cooperating agents. *Artif. Intell.* **171**, 535–567 (2007)
47. Korte, B., Vygen, J.: Combinatorial Optimization—Theory and Algorithms. Springer, Berlin (2002)
48. Myerson, R.: Graphs and cooperation in games. *Math. Oper. Res.* **2**, 225–229 (1977)
49. Owen, G.: Values of graph-restricted games. *SIAM J. ADM* **7**(2), 210–220 (1986)
50. Diestel, R.: Graph Theory. Springer, Berlin (2010)
51. Azrieli, Y., Lehrer, E.: Concavification and convex games. Working Paper, Tel Aviv University, pp. 0–18 (2005)
52. Holzman, R., Lehrer, E., Linial, N.: Some bounds for the Banzhaf index and other semivalues. *Math. Oper. Res.* **13**, 358–363 (1988)



Interoperable Convergence of Storage, Networking, and Computation

Micah Beck^(✉), Terry Moore, Piotr Luszczek, and Anthony Danalis

Department of Electrical Engineering and Computer Science,
University of Tennessee, Knoxville, TN 37996, USA
mbeck@utk.edu, {tmoore,luszczek,adanalis}@icl.utk.edu

Abstract. In every form of digital store-and-forward communication, intermediate forwarding nodes are computers, with attendant memory and processing resources. This has inevitably stimulated efforts to create a wide-area infrastructure that goes beyond simple store-and-forward to create a platform that makes more general and varied use of the potential of this collection of increasingly powerful nodes. Historically, these efforts predate the advent of globally routed packet networking. The desire for a converged infrastructure of this kind has only intensified over the last 30 years, as memory, storage, and processing resources have increased in both density and speed while simultaneously decreasing in cost. Although there is a general consensus that it should be possible to define and deploy such a dramatically more capable wide-area platform, a great deal of investment in research prototypes has yet to produce a credible candidate architecture. Drawing on technical analysis, historical examples, and case studies, we present an argument for the hypothesis that in order to realize a distributed system with the kind of convergent generality and deployment scalability that might qualify as “future-defining,” we must build it from a small set of simple, generic, and limited abstractions of the low level resources (processing, storage and network) of its intermediate nodes.

Keywords: Networking · Distributed computing

1 Introduction

A variety of technological, economic, and social developments—most notably the general movement toward Smart Cities, the Internet of Things, and other forms of “intelligent infrastructure” [1]—are prompting calls from various quarters for something that the distributed systems community has long aspired to create:

Dr. Beck is an Associate Professor at University of Tennessee, Knoxville. he is currently on detail to the National Science Foundation in the Office of Advanced Cyberinfrastructure. The work discussed herein was completed prior to his government service and does not reflect the views, conclusions, or opinions of the National Science Foundation or of the U.S. Government.

A next-generation network computing platform. For example, the authors of a recent Computing Community Consortium white paper, writing with the US “Smart Cities” initiative [2] in view, express the research challenge as follows:

“What is lacking—and what is necessary to define in the future—is a common, open, underlying ‘platform’, analogous to (but much more complex than) the Internet or Web, allowing applications and services to be developed as modular, extensible, interoperable components. To achieve the level of **interoperation** and innovation in Smart Cities that we have seen in the Internet will require *federal investment in the basic research and development* of an analogous open platform for intelligent infrastructure, tested and evaluated openly through the same inclusive, open, consensus-driven approach that created [the] Internet.” [3] [Emphasis in source]

The experiences of the last two decades have made the distributed systems community acutely aware of how elusive the invention of such a future-defining platform is likely to be [4]. Achieving this vision has been the explicit or implicit ambition of a succession of well funded and energetically pursued research and development efforts within or around this community, including Active Networking [5], Grid Computing [6], PlanetLab [7], and GENI [8], to name a few. Although these broad efforts have produced both valuable research and useful software results, nothing delivered so far has achieved the *deployment scalability* necessary to initiate the kind of viral growth that everyone expects such an aspirational platform to exhibit. At the same time, chronic problems with network hotspots were an early and persistent sign that the Internet’s stateless, unicast datagram service had scalability limitations with respect to data volume and/or popularity. This fact has led to increasingly sophisticated and increasingly expensive technology “workarounds,” from the FTP mirror sites and Web cache hierarchies of the early years, to the content delivery networks (CDN) and commercial Clouds we see today.

The central idea of this paper is that the appropriate common service on which to base an interoperable platform to support distributed systems is an abstraction of the low layer resources and services of the intermediate node, i.e., a generalization of the Internet stack’s layer 2. The “Internet Convergence” of the 1990s developed the “hourglass” paradigm, with best-effort datagram delivery as the common service, or “spanning-layer,” at its narrow waist [9]; we believe that the paradigm required by the data saturated world now emerging in edge/fog environments is more accurately pictured as an “anvil” (Fig. 1), with a common service interface that exposes storage/buffer, network, and processor resources in a programmable way. Drawing on technical analysis and historical examples, we argue that in order to build distributed systems with the kind of interoperability, generality and deployment scalability that might qualify as “future-defining,” we must implement them using a small set of simple, generic, and limited abstractions of the data transfer, storage and processing services available at this layer. In our model, these abstractions all revolve around the fundamental common resource, the memory/storage buffer.

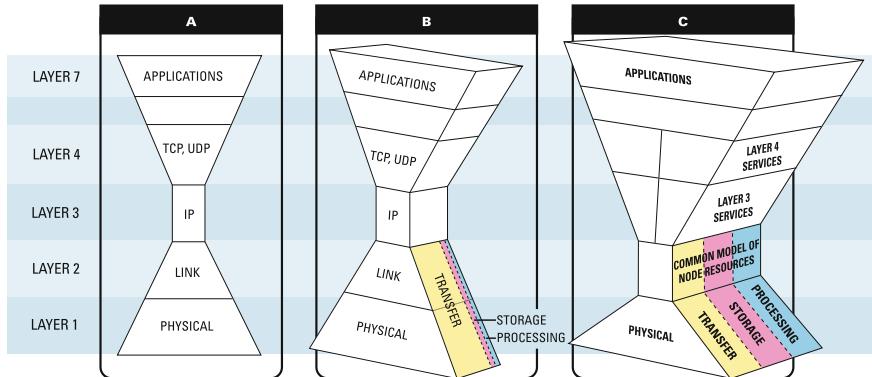


Fig. 1. The Hourglass versus The Anvil

2 Background

Given the inclination of computer scientists to add features, the fact that every form of digital store-and-forward communication (including the Internet) has intermediate forwarding nodes that are computers, with attendant memory and processing resources, makes attempts to create a wide area infrastructure with services beyond simple store-and-forward inevitable. Such efforts to make more general use of these increasingly powerful nodes—a *generalized converged network*, in our terminology—predate the advent of globally routed packet networking (e.g. uux [10]). The exponentially increasing density and speed, and rapidly decreasing cost of memory, storage and processing resources over the past 30 years has only intensified the desire to define and scalably deploy a converged infrastructure of this general description. Yet despite the general consensus that it should be possible to do so, this aspiration has remained unfulfilled.

One problem is that the goal of converged networking runs in the opposite direction of the traditional architectural approach of the Internet design community, which insists that services other than datagram delivery must be kept out of the Network Layer of the communication protocol stack. This community maintains that the ability of the Internet to function properly and to continue growing globally depends on keeping this common service layer “thin”, in the sense that it provides services that are simple, generic and limited. From this point of view, services other than datagram delivery should be implemented in systems connected to the Internet as communication endpoints. Various rationales supporting this point of view are collectively referred to as “End-to-End Arguments” [11].

Since a router that has substantial system storage (i.e. other than network buffers) and generalized computational resources (i.e. other than forwarding) is neither difficult nor expensive to build, there have been numerous efforts to resist this orthodox point of view. The simple fact that storage and computational resources can be provisioned and located throughout the network at reasonable

cost stimulates efforts in this direction. Moreover, the apparent opportunity to create such a powerful distributed infrastructure presents a temptation that is inherently difficult for computer scientists and engineers to resist. These facts, however, do not make it a good idea to add extensions to the fundamental service of the global Internet, nor do they ensure that if it is built, service creators and users will adopt it at a scale sufficient to enable economic sustainability beyond the prototype stage. Indeed, while a number of plausible network service architectures have been defined that can provide access to such distributed resources [5, 12], the widespread deployment of extended services on a converged wide area infrastructure has proved elusive.

Perhaps an even more compelling reason for the continued drive to create such a converged infrastructure is that some important distributed applications cannot be efficiently and effectively implemented through decomposition into two components, one implemented by a “thin” datagram delivery service in the core of the network, and the other implemented at “fat” endpoints. For example, some applications require an implementation that is sensitive to the location of storage and computation in the network topology. Point-to-multipoint communication was an early and obvious example. Using simple repetition of unicast datagram delivery was viewed as too inefficient by early Internet architects, but an efficient tree could be built only through the use of network topology information. Such low level information was seen as inappropriate for users of the “thin” and stable Network layer to access. Thus, multicast was added to Layer 3, fattening that thin layer with services that seemed to address this issue. However, IP multicast has proved difficult to standardize and has failed to achieve the universal deployment of “simple, generic and limited” unicast IP datagram delivery.

But problems with lack of generality in the intermediate nodes were manifest even in highly successful Internet applications. The early growth of the Internet was fueled by applications that seemed to fit the unicast datagram delivery model well enough: FTP and Telnet. Of these, the one-to-many nature of FTP, albeit asynchronous, created a problem in the distribution of popular and high-volume files. Ignoring the implications of topology led to ineffective use of the network, with hotspots at servers that attracted high volumes of traffic and unnecessary loads placed on expensive and overburdened wide area links. The result was the creation and management of collections of FTP mirror sites [13], and the ubiquitous invitation for users to “choose a mirror site near you”, which meant the use of approximate information about network topology by the end-user, at a level above even the Internet stack’s Application Layer.

The advent of the World Wide Web exacerbated the problem of indiscriminate access to servers with no reference to network topology or even geography. Mirror sites for file download proliferated, and redundancy in the storage of all high-traffic Web content became a necessity. A Network layer that hides topology from its clients is, after all, an inherently inadequate platform on which to build high traffic globally distributed systems. The need to work around this reality

gave rise to automated Web caching [14, 15] and server replication [16, 17], which were precursors to modern Content Delivery Networks [18, 19].

It should be noted that although both Web caching and server replication are obvious examples of the convergence of networking and storage, they also require computation in the implementation of policy and server-side processing; and so in fact they represent convergence of all three fundamental computational resources. We examine the approach to convergence that they represent in more detail in Sect. 5 below. Following a different strategy, Logistical Networking, discussed in Sect. 6.2, implements a convergence of networking and storage service that avoids the need for general computation by minimizing policy and other server-side processing [20], but was later extended to include limited server-side operations [21].

3 The Convergence Spectrum

The interplay between technological divergence and convergence is a dialectic with a long history. In the area of computing and communications, there was an early divergence in the conception and implementation of several different information technology resources. Because of the phenomenon of path dependence [22], such divergence has tended to be self-reinforcing, leading to a set of familiar technology *silos*, such as data transmission and broadcast using radio frequency signals, virtual circuits, switches and gates and magnetic or solid state storage cells. The success of the Internet in the 1990s provided the foundation for the substantial or partial convergence of various traditional telecommunication silos—telephony, broadcast television, etc.—in this century [23], but the three fundamental silos at the base of computing—storage, processing, networking—have remained as entrenched as ever.

The early divergence of basic computational resources has given rise to conceptual, technological and organizational *silos* corresponding to isolated communities. Formal models and methods of reasoning have been adapted to deal with the complexity and specific issues of each niche. For example Boolean logic is a useful model of solid state circuits, and “stateless” communication is a useful model of a wide area data network built out of switches and FIFO line buffers.

The development of silos has been an enabling strategy for modeling and optimization of these quickly evolving technological fields. However, they have also led to the creation of service stacks with highly specialized services at the top layers (see Fig. 2). But because the low level resources that these silos encapsulate can only be accessed through high level services, this inevitably tends to create barriers to the flexible and efficient use of constituent low level resources *in combination*.

The problem with silos as a strategy for dealing with the complexity and specialization of disparate underlying technologies has become more pronounced due to the evolution of low level systems toward general mechanisms that utilize processors or digital logic controlled by software, firmware or by hardware designed using computerized tools. Such generality in low level mechanisms holds out the

possibility of the implementation of highly efficient system architectures, with optimizations that span traditionally disparate resources. The challenge is to bridge or eliminate the existing silos, or, in other words, to implement *convergence*.

We say that a service interface (i.e., an API) is *converged* if it gives unified access to multiple low-level resources (or services) traditionally available only through isolated service silos. Historical examples of system design that leverage convergence include the auto-increment register, direct memory access I/O and vector processing.

When the goal is to achieve convergence for a service interface using previously non-interoperable resources, there are two fundamentally different ways to go about it: *overlay convergence* which combines silos at a layer above their high level services, and *interoperable convergence*, which strives to unify their foundations. These two strategies lie at the ends of a spectrum along which a variety of familiar examples can be arrayed.

Overlay convergence is the most common approach lying at one end of the spectrum and creating a high level interface that provides access to a number of traditionally separate service silos. We term this approach *overlay convergence* because it typically involves the creation of a service that provides unified access to the existing service silos from above, through their high level client interfaces (see Fig. 2). By contrast, at the other end of the convergence spectrum is what we call *interoperable convergence*. We say that a platform is interoperably converged if it minimizes the imposition of unnecessary high-level structure or performance costs when applying different low-level services, so that those underlying common resources can be accessed without incurring the overhead and restrictions that are associated with complex and specialized service silos.

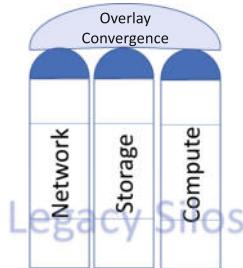


Fig. 2. Overlay convergence of legacy silos

Both overlay and interoperable forms of convergence seek to create a common service, or *spanning layer*, which supports a generalized set of applications requiring resources that were previously segregated. As exemplified in the classic case—the “narrow waist” of the Internet’s “hourglass” protocol architecture (see Fig. 1)—the purpose of such a spanning layer is to enable interoperability in the support of this rich category of applications [9].

Some examples that fall along this spectrum and illustrate these different approaches include the following:

- The BSD kernel created an overlay convergence of Unix process and local file management with local and wide area networking through the addition of the *socket* related system calls. While some calls that act on file descriptors such as `read()` and `write()` were extended to operate on sockets, the level of integration is mainly syntactic and does not extend deeply into integration of common functions such as buffer management.
- Following the implications of this example, in order to move data stored in a file to a TCP stream in UNIX, it was originally necessary to move it into a process' address space using the `read()` system call and then inject it into the TCP stream using `send()` (see Fig. 3). A more interoperable approach is a combined `sendfile()` system call which was added as an extension to Linux that allows data to be transferred from storage into a kernel memory buffer and from there directly to the network without moving it to process memory or using a dedicated network data buffer. However this buffer management solution is applicable only in quite specific scenarios. We thus characterize it as a *workaround*.
- A distributed file system converges storage and data movement in a more interoperable manner. These resources are otherwise available only through local file management and networked file transfer tools.
- A database system can store a set of tuples without order, but traditional data movement tools operate on files. Thus, it is necessary to serialize a set of tuples as a file in order to send it to a remote database system. The file is transferred serially, using TCP with retransmission to keep the serialized data in order. A somewhat interoperable approach would generate the serialized stream representing the tuple set on demand, rather than creating and storing it as a complete file. A more interoperable approach would be to implement a specialized protocol that takes advantage of the lack of natural sequentiality in the tuple set to perform retransmission out-of-order. This might require additional work to ensure that the new protocol was “TCP-friendly” when used in public shared networks.
- A data analysis system (such as MapReduce [24]) traditionally consists of a deep data store and a dedicated compute resource such as a cluster or a shared-memory parallel computer. Visualization typically requires data to be moved from the data store to the compute resource which then returns its results to the data store. User access then requires that the visualization output be moved to and interpreted by a human interaction system. A more interoperable approach would allow computations to be applied to the data in the data store (*in-situ*), and for the user to interact with the results of that computation directly as it occurs.

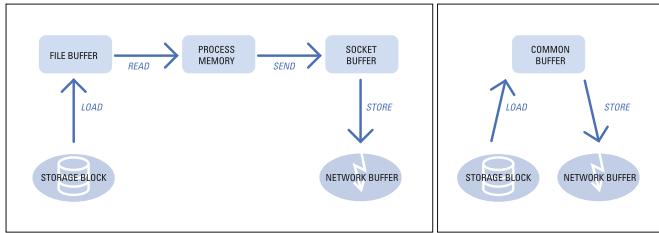


Fig. 3. Read-Send versus Sendfile

4 Deployment Scalability

When it comes to creating layered software stacks for large and diverse communities, the importance of a well designed spanning layer is difficult to overestimate. Most critically, a successful hourglass design connects directly with the diverse demands of *multi-stakeholder ecosystems*. As shown in Fig. 4, a well designed spanning layer can be implemented on many different substrates, yielding a wide “lower bell,” and yet also support an equally great variety of applications, yielding a wide “upper bell.” A wide lower bell means that the spanning layer can be implemented on heterogeneous hardware/software platforms, enabling applications and services above the spanning layer to access and utilize these diverse resources. The wider the lower bell, the stronger the assurance that both legacy and future platforms will support many different stakeholders. A wide upper bell means that the small set of primitive services that the spanning layer makes available on the system’s nodes can be composed, often using additional stakeholder-provisioned resources, to support a broad diversity of higher-level services and applications. The wider the upper bell, the stronger the assurance that more specialized application communities can build what they need atop the shared infrastructure. Thus, a successful hourglass architecture, with capacious upper and lower bells, will lower or eliminate barriers to adoption for a wide variety of stakeholders.

The spanning layer in such a successful hourglass-shaped stack is said to be “narrow” because “...it represents a minimal and carefully chosen set of global capabilities that allows both higher-level applications and lower-level communication technologies to coexist, share capabilities and evolve rapidly [25].” But deliberately creating a spanning layer with such a *minimally sufficient* specification has proven to be exceedingly difficult, as is shown by the list of major efforts over the last two decades that have attempted to do so (see Sect. 1).

Rationally evaluating alternative strategies requires a criterion for success. Accordingly, we introduce the concept of *deployment scalability* as the design goal of the foundational spanning layer of a converged infrastructure. We define *deployment scalability* as *widespread acceptance, implementation and use of a service specification*. Since spanning layers define communities of interoperability, the greater the deployment scalability of a given spanning layer, the larger the community of interoperability it can achieve. The workarounds we have described

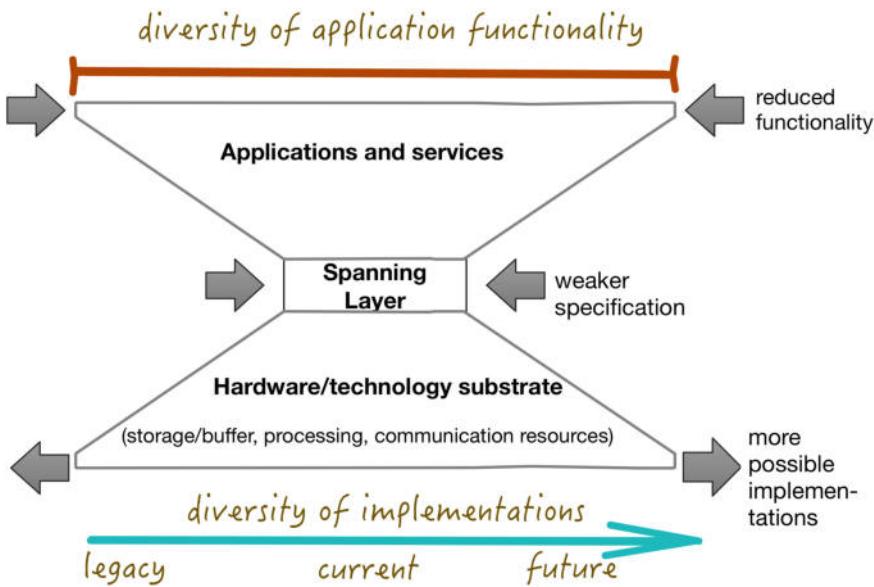


Fig. 4. The “hourglass model” for a system software stack. The goal is to achieve deployment scalability while maximizing the diversity of applications. The arrows indicate the design tensions involved in creating a spanning layer with a minimally sufficient service specification

build overlay converged systems, but they cannot achieve a sufficient level of deployment scalability, which constrains the size of the communities they can sustainably support.

In a recent paper [26], Beck makes an argument for a fundamental design principle of the common service underlying systems that exhibit deployment scalability:

The Deployment Scalability Tradeoff There is an inherent tradeoff between the deployment scalability of a specification and the degree to which that specification is weak, simple, general and resource limited.

The terms “simple, generic and resource limited” are derived from the classic paper “End-to-End Arguments in System Design” by Saltzer, Reed and Clark which discusses them in the context of Internet architecture. The term “weak” refers to logical weakness of the service specification as a theory of program logic, and is due to Beck’s partial formalization of the arguments in that paper. Stating this principle as a tradeoff is a further refinement of the usual interpretation of the original paper as an absolute rule (or principle) requiring or prohibiting particular design choices [27].

The classic example of the application of the End-to-End Principle, from which its name is derived, is the location of the detection of data corruption or packet loss or reordering in the TCP/IP stack [11]. The scalability argument for

end-to-end detection of faults is that removing such functions from the spanning layer makes it weaker, and therefore potentially admits more possible implementations. Because fault detection can be implemented above the spanning layer, the set of applications supported is not reduced.

The evolution of process creation in Unix teaches a similar lesson. In early operating systems it was common for the creation of a new process to be a privileged operation that could be invoked only from code running with supervisory privileges. There were multiple reasons for such caution, but one was that the power to allocate operating system resources that comprise a new process was seen as too great to be delegated to the application level. Another reason was that the power of process creation (for example changing the identity under which the newly created process would run) was seen as too dangerous. This led to a situation in which command line interpretation was a near-immutable function of the operating system that could only be changed by the installation of new supervisory code modules, often a privilege open only to the vendor or system administrator.

In Unix, process creation was reduced to the `fork()` operation, a logically much weaker operation that did not allow any of the attributes of the child process to be determined by the parent, but instead required that the child inherit such attributes from the parent [28]. Operations that changed sensitive properties of a process were factored out into orthogonal calls such as `chown()` and `nice()`, which were fully or partially restricted to operating in supervisory mode; and `exec()` which was not so restricted but which was later extended with properties such as the *setuid* bit that were implemented as authenticated or protected features of the file system. The decision was made to allow the allocation of kernel resources by applications, leaving open the possibility of dynamic management of such allocation by the kernel at runtime, and creating the possibility of “Denial of Service” type attacks that persists to this day.

These two classical examples of interoperable convergence point to a significant issue. Changing the low level services on which existing silos are built requires the redesign and reimplementations of complex higher level service stacks. The influence of path dependent thinking and the pain of abandoning “sunk investments” explain the natural tendency of service provider communities to develop *workarounds* that preserve widely deployed lower level services. In Sect. 5 below, we analyze some familiar overlay workarounds to the problems that can be traced to the tradeoffs that the designers of the Internet made.

5 Web Caching and CDNs: A Case Study in Overlay Workarounds

During what might be called the “Internet Convergence” in the 1990s, the generality and scalability of the Internet’s datagram delivery model gave rise to the idea of using it to implement the convergence of broadcast, telephony and data services [23]. The emergence of unicast datagram delivery as the only universal Internet service (discussed in Sect. 2) has meant that the underlying capabilities

of lower layer mechanisms to utilize electromagnetic broadcast and to guarantee quality of service through resource reservation are not accessible using Voice over IP and Streaming Media over IP protocols. In spite of such limitations, the convenience and cost benefits of convergence workarounds continue to dominate the commercial development of these services.

But the absence of a universal point-to-multipoint communication mechanism within the common Network layer of the Internet left a large class applications without native support, and this, in turn, has generated a whole series of overlay workarounds (see Fig. 5). For instance, the distribution of static Web pages (those that require only minimal rewriting of stored HTML pages) can be viewed as a form of point-to-multipoint communication. A browser cache uses moderate storage resources in the network endpoint to capture the delivered Web page and associated metadata and minimal processing to implement the cache policy and mechanism. A proxy cache uses larger scale storage and has a greater processing load, which is supplied by a substantially provisioned network intermediate node. The convergence of resources in Web caches led to an architectural development in which application-specific proxies are uploaded to a “middlebox” platforms which implements both caching and general processing.

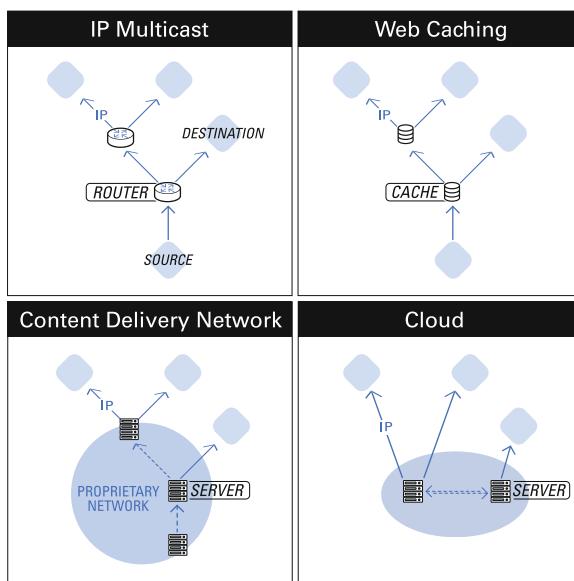


Fig. 5. Overlay workarounds addressing point-to-multipoint distribution

Web caching played a pivotal role in the expansion of the Web as a global data distribution service during the period when intercontinental data links were too expensive to allow unfettered access by academics. A hierarchical system of large scale caches was developed and deployed in US Research and Education

Networks [14, 15] and use of national caches to access Web data across intercontinental links was made mandatory in many countries.

In spite of its effectiveness in reducing the traffic loads due to delivery of static Web pages, the popularity of intermediate caches has waned dramatically in the past decade. There are several reasons for this trend including:

- The correctness of Web caching relies on lifetime metadata being provided by origin servers which is often missing or inaccurate.
- The growth of dynamic Web applications means that many Web objects are not cachable.
- The lack of accurate and universal mechanisms for reporting views interferes with the dominant business model of Web advertisers.
- Reliance on a complex cache infrastructure decreases the control by the implementer of a Web service over the Quality of Service experienced by customers.

Many of these factors stem from the implementation of Web caching on top of the HTTP application protocol, albeit with some modifications to increase control over intermediate and browser caches by origin Web servers. Cache networks are an overlay which accesses Web services from the top of the protocol stack and thus does not allow the degree of fine-grained control that is required for seamless convergence.

An alternative approach is to start from the source, and to replicate the functionality of the Web server on multiple network nodes. Manual procedures for FTP mirroring led to automated mechanisms like Netlib [13], and high traffic Web sites gave rise to sophisticated cluster and geographically distributed server replication schemes [16, 17].

Commercial Content Delivery Networks (CDNs) have approached the problem in a somewhat different way, using HTTP and streaming protocols for client access almost unchanged. This is analogous to the way that online services (e.g. Compuserve and AOL) and ISPs used telephone services. CDNs have instead focused their innovation on the underpinnings of the Internet in order to improve the effectiveness. They use a combination of server side caching, distributed file and database systems and complex streaming and synchronization protocols implemented on private, proprietary international networks of application-specific servers.

CDN Web and DNS servers may be implemented as applications processes, but by using lower layer Internet mechanisms through now-commonplace layering violations (such as topology-sensitive DNS resolution), they use knowledge of network topology and other low level information that is intended to be encapsulated within the Network Layer of the Internet architecture. Modern extensions to the Network layer may allow CDN's to be implemented without such violation of layering, but at the expense of creating a “fatter” and less generic Network layer (see Sect. 4).

Commercial CDNs are thus a kind of Chimera, patched together from proprietary components and standard, low level components of the Internet. They create a proprietary, specialized network with their own services as the spanning layer, using the Internet as tools in their implementation and as a means

of reaching end users. This view is supported by the trend toward using private or non-scalable mechanisms to implement internal communication among centralized and distributed CDN nodes. Since the currently emerging paradigm for edge computing “...extends the CDN concept by leveraging cloud computing infrastructure ... [29],” the non-interoperable nature of these overlay approaches undermines the coherence of Internet-based information service ecosystems. As we argue below, *Exposed Buffer Processing* offers a more interoperable and unified foundation for such next-generation ecosystems.

6 Exposed Buffer Processing: An Approach to Interoperable Convergence

While operating system interfaces such as POSIX provide access to storage, networking and computing services, they do so in ways that conform to the traditional silos.

- File system calls do not have explicit access to general networking or computation resources.
- The socket interface does not provide access to general storage or computation resources.
- The POSIX process management functions have only the minimal necessary overlap with storage and network functions (notably specifying an executable file image in the `exec()` system call).

However, the core resource that is used to implement all these silos is the persistent memory or storage buffer.

- In *storage*, disk/SSD blocks or objects are used in the implementation of higher level file and database systems, along with RAM memory buffers that are used to improve performance, enable application/OS parallelism and allow for flexible exchange of data with other operating system data structures.
- In *networking*, buffers are used at the endpoints for much the same reasons as storage, and are used at intermediate nodes to allow for asynchrony in the operation of store-and-forward packet networking.
- In *computing*, memory pages make up process address spaces, are also used to enable asynchrony in interprocess communication, and hold all other operating system data structures used in the implementation of functions on behalf of processes.

So although convergence of storage, networking and computation is possible through conventional operating system interfaces using the generality of the user process as a gateway between silos, a more interoperable approach is to expose a common abstraction of the underlying resource that all of these high level silos operate on, namely persistent storage blocks or memory buffers. We call this approach to convergence *Exposed Buffer Processing*.

6.1 Core EBP Functionality

Exposed Buffer Processing is a general architectural idea that can have different implementations. The original implementation, which takes the form of the Logistical Networking stack described in Sect. 6.2, encapsulates the service as remote procedure call over TCP.

- **allocate**—This call allocates storage capacity into which data can be stored. It is possible for this allocation to be performed implicitly (as part of the `store` operation described below) or explicitly (as a reservation of resources which subsequent `store` operations can use.) An allocation call specifies several parameters that limit resource utilization, such as duration. One important attribute of an allocation is a local name by which stored data is referenced in subsequent calls.
- **store** and **load**—These calls allow EBP clients to store data in an allocated buffer or to load data that has previously been stored.
- **transfer**—This call transfers data between buffers specified by name on the same or on different depots.
- **transform**—This call applies a named operation to a set of buffers on a single depot, potentially transforming the data of one or more of them. The operation name is local to the depot and must have been defined previous to the transform call by a code upload operation which is specific to the implementation of the depot. Operations will be assigned names through a client-community process that ensures that names are used in a sufficiently consistent manner.

6.2 Logistical Networking as EBP in Overlay

Over the past 15 years the Logistical Networking project [20, 21, 30] has worked to define an approach to Exposed Buffer Processing that is implemented as an overlay on the Internet. An examination of the key components of that implementation provides an EBP proof of concept:

- **Internet Backplane Protocol (IBP)**: IBP is a generalized Internet-based storage service that is encapsulated as remote procedure call over TCP. IBP was designed to be simple, generic and limited following the example of the Internet Protocol (IP) [11]. It is a best effort service, its byte array allocations are named only by long random server-chosen keys (capabilities) and represent leases whose duration and size are limited by the individual intermediate node (in analogy to the IP MTU). The intermediate node that implements IBP is called a *depot*, and it is intended as a storage analog to IP routers. In many ways IBP is closer to a network implementation of `malloc()` than a conventional Internet storage service like FTP. In addition every IBP allocation is a lease of storage resources which can be limited in duration. The IBP depot has been implemented in both C and Java.
- **exNode**: Because IBP is such a limited service, the abstraction of an allocation that it supports does not have the expressiveness of the file abstraction

that users typically expect of a high level data management system. The exNode is a data structure that holds the structural metadata required to compose IBP allocations into a file of very large extent, with replication across IBP depots, identified by their DNS name or IP address [31]. The exNode can be thought of as an analog to the inode used in early Unix file system implementations. The exNode has both standard XML and JSON sequentializations.

- **Logistical Runtime System (LoRS):** The exNode can be used as a file descriptor to implement standard file operations such as **read** and **write**. The Logisticsal Runtime System (LoRS) uses the exNode to implement efficient, robust and high performing data transfer operations. Some of the techniques used in the implementation of LoRS are comparable to those used in parallel and peer-to-peer file transfer protocols [32].
- **Logistical Distribution Network (LoDN):** While the exNode implements topological composition of IBP allocations to implement large distributed and replicated files, it does not deal with the temporal restrictions introduced by IBP’s use of storage leases. LoDN is an active service which holds exNodes and applies storage allocation, lease renewal and data movement operations as required to maintain policy objectives set by end users through a declarative language and manageable by an intuitive human interface.
- **Network Functional Unit (NFU):** The NFU was introduced as a means to allow simple, generic and limited in-situ operations by a depot to data stored in its IBP allocations. The NFU has been used in numerous experimental deployments, and has been shown to enable robust fault tolerance and high performance in a wide variety of applications [33–35]. However, the middleware stack that supports such experimentation has never been fully integrated with the packaged versions of LoRS and LoDN or the Data Logistics Toolkit (discussed below), and so the NFU has never been used in a persistent large scale deployment.

6.3 “Packetization” of Storage and Processing

One way to characterize EBP’s simple, generic, and limited design philosophy for the abstractions of common spanning layer services is to say that it extends the idea of “packetization” from the domain of networking, where it has proved so remarkably successful, to the domains storage/memory and processing as well. Unfortunately, this contradicts the impulses many of service architects who have historically relied on the more complex, specialized and virtually unbounded services. The relevant contrasts between packet-based and circuit-based approaches are familiar and clear in realm of **Networking**:

- **Size:** Circuit-based networks allow an unbounded amount of data to pass over a persistent circuit, in analogy to an electrical connection, masking the underlying digital implementation in terms of MTU-limited packets. The Internet exposed the MTU and required endpoints to concatenate packets into streams.

- **Failure:** Circuit-based networks provide Quality-of-Service (QoS) guarantees sufficient to enable application developers to be protected from occasional communication faults but which fail catastrophically when protection is impossible. The Internet exposed the possibility of failure by exporting a best effort service, requiring endpoints to detect and respond to failures.
- **Locality Independence:** Circuit-based networks can allocate resources and maintain state along a specific path from sender to receiver, helping to ensure fast forwarding and providing a stable platform for implementation of auxiliary services. The Internet allows every packet in a connected flow to be forwarded along a different path, putting the burden for maintaining stability on the packet routing scheme and ruling out connected services that require the maintenance of state, but enabling great resilience in the face of failures and changes in topology.

The similarities between networking and storage make the the analogous set of contrasts relatively easy to work out for the realm of **Storage**:

- **Size:** File-based models of storage allow a very large amount of data (assumed by many applications to be virtually unbounded) to be stored as a single linear data extent. Logistical Networking (i.e., EBP in overlay) exposes a maximum storage allocation size imposed by the storage resource (analogous to the Internet Protocol's MTU) requiring endpoints to explicitly concatenate allocations into files.
- **Failure:** File and database systems provide reliability guarantees sufficient to enable application developers be protected from occassional storage faults but which fail catastrophically when protection is impossible. Logistical Networking exposes a simple failure model (faulty write operations terminate with unknown output state) and by exporting a best effort service, requires endpoints to explicitly detect and respond to failures.
- **Locality Independence:** File-based models of storage can allocate resources and maintain state on a well-connected “site” to manage fault tolerance and replication in terms of where “copies” reside. Logistical Networking allows every allocation comprising a file to be managed independently, potentially spreading them across topologically seperated nodes, moving and storing data on a fine-grained basis as called for by applications (e.g., data streaming).

Finally, an analogous set of contrasts applies to the realm of **Computation**:

- **Size:** Process-based computation allows an unbounded amount of processing to be performed by a set of one or more closely-coupled threads. The Network Functional Unit (i.e., EBP in overlay) exposes a unit of processing that can be limited in many resource dimensions, including CPU cycles consumed, RAM allocated during execution and I/O activity performed, requiring a runtime system to concatenate limited resources to create an unbounded virtual execution model.
- **Failure:** Process-based computation provides QoS guarantees sufficient to enable application developers to either overcome occasional processing faults

or to fail catastrophically when they are detected. The NFU exposes a simple failure model (faulty operations terminate with unknown state for write-accessible storage) and exports a best effort service, requiring endpoints to explicitly detect and respond to failures.

- **Locality Independence:** Process-based computation can allocate resources and maintain state on a set of well-connected processors, enabling successive time slices to execute sequentially in a manner that leverages continuity of operating system and application data state. The NFU allows every allocation comprising a process to be managed independently, potentially moving them and the memory/storage allocations that comprise the state of supervisory and application data state as required (eg fault tolerance and load balancing).

6.4 EBP over Packet

Currently, the IBP protocol is encapsulated over TCP. Using TCP as a substrate offers several benefits and simplifies the engineering effort for the EBP developer, but this convenience comes at a cost. In particular, many design decisions baked into TCP were made to serve the needs of public wide-area networks that must support competing large data transfers that implement fair contention and flow control to facilitate resource sharing.

Not all of these features are necessary for EBP to work efficiently, and in fact, some are detrimental in certain contexts. For example, strict in-order delivery of network packets can inhibit parallelism, and slow start imposes unnecessary latency when an EBP operation involves delivery of a number of packets that is too small to cause congestion. For this reason it is useful to consider a possible implementation of EBP as it would be encapsulated more directly over a packet transport, be that IP v4 or v6, Ethernet or another Link Layer service. Considering such a primitive encapsulation lays bare the relationship between the scale of communication, storage and processing, and how these can be reconstructed in a converged way if they are expressed as the aggregation of limited operations.

- **Operation sequencing:** TCP delivers packets in order, which means that if a packet arrives earlier than it is expected, it will not be delivered to the upper layer until the “missing” packets arrive. When these packets contain parts of a data buffer, then it is essential that they are assembled in the correct order. However, requests for multiple EBP operations—which in a packet encapsulation would be transmitted as multiple packets—might depend on one another or they might be completely independent. When two independent operations are sequenced in a TCP stream, the second may arrive at its destination before the first but not be delivered for processing, causing delays and limiting parallelism. To satisfy the correctness requirements of operations that *do* depend on one another specific operation packets can be tagged and *necessary* order imposed. There are trade-offs between expressiveness and cost (in terms of bits used by different tagging mechanisms for storing the tag) which can be addressed using techniques developed for dynamically scheduled systems for expressing dependencies between tasks as DAGs.

- **Retransmission:** TCP provides mechanisms for retransmitting packets that are not acknowledged within a timeout using exponential backoff to avoid congestion. The approach taken by TCP is geared to maintaining in-order delivery and to keeping the avoidance of congestion paramount in the use of shared wide area networks. In some EBP scenarios, time-critical operations could retransmit packets much more aggressively to minimize the possibility of all packets being lost. However, in order to maintain correctness in the presence of non-idempotent EBP operations (such as those with side-effects) it will be necessary to implement bookkeeping at the destination to avoid *stuttering*.
- **Flow control:** TCP provides flow control that is tied to detection of lost packets at the expense of throttling high bandwidth transfers. In cases where flow control is needed, it will have to be implemented at a higher layer.

Since EBP offers several services simultaneously—storage, networking, and computation—there are several aspects of QoS that can be explored.

- **Allocation of Bandwidth/Storage/Computing.** The most fundamental functionality that EBP provides is allocation of resources. Especially in the scenario where EBP is implemented over IP, and thus the fair contention safeguards of TCP are bypassed, a client application could request to allocate the whole bandwidth capacity of a network route. In this scenario competing allocations would have to be declined. Similarly, one can envision multiple allocations that only require part of the network's capacity to be satisfied simultaneously. In the same spirit, and using similar bookkeeping mechanisms, we could implement allocation of storage and computing resources on the nodes.
- **Hard** versus *soft* allocations. As an additional QoS feature we can provide multiple levels of allocation “hardness” which would trade the level of guarantee for the amount of resources available. At least three levels of hardness can easily be envisioned: (A) best-effort clients, which make no allocation (and thus receive no QoS guarantee) can make unlimited attempts to allocate resources; (B) clients which make *soft* allocations can allocate a resource before it is used but are first to be preempted when that resource is exhausted due to competing system activities or overbooking of hard allocations (see below). Soft allocations can result in the denial of best-effort clients even when resources are not exhausted; and (C) clients which make *hard* allocations can preempt all other types of clients and are last to be preempted in case of total resource exhaustion.
- **Statistical overbooking,** can be used to allow resource reservation to exceed available resources to take advantage of underutilization of reserved resources. Overbooking can be managed through a multi-tiered schema of allocation “hardness” as described above.

6.5 EBP Below the Network Layer

The argument for creating a converged layer to support global distributed services is compelling. The need for distributed systems to have access to and con-

trol over low layer network characteristics including topology and performance is clear in the steps that have been taken to work around the stricture that forbids such direct access in the Internet architecture.

We propose the creation of a platform based on a common service similar to IBP but which models the networking capabilities of the Link Layer. We use the term Exposed Buffer Processing for this as-yet-unrealized service. The central idea of this paper is that the appropriate platform for the creation of distributed systems is some form of EBP. We emphasize that EBP need not follow the design of IBP, as long as it takes appropriate account of the Deployment Scalability Tradeoff. We offer experience with IBP as an overlay form of EBP for the consideration of the community.

7 Applications of EBP

7.1 Scientific Content Delivery

Dissemination of data is one of the fundamental challenges of modern experimental and observational science. There is a general move toward the open sharing of raw data sets, enabling replication of analyses, cross-cutting studies, innovative reexamination of previously collected data and historical examination of collection and analysis techniques [36,37]. In many cases the data collected is large and observation is continuous, as in remote data from satellites and other sensors [38], experiments such as the Large Hadron Collider [39], or broad harvesting of multimedia content [40]. The resources required to make such data streams instantaneously and persistently available can exceed the centralized capabilities of institutions or government agencies.

Commercial CDN or Cloud solutions may be too expensive, and may not adequately serve the entire global user community (see discussion of the Digital Divide below) and may not adequately support the publication by users of secondary data products resulting from their processing of raw data. However, the ICT resources required to address such problems may be affordable, and the community of user institutions may be capable of hosting them in a distributed manner. Using shared EBP infrastructure, we can build a distributed, federated content management system using the resources of the content provider and user communities

7.2 Digital Divide and Disasters

Modern network services take full advantage of the strong assumptions that can be made about the implementation of the Internet in the industrial world. It is common for services to rely on continual low-latency datagram delivery, always-connected servers, stable and uninterrupted datagram routing paths and high bandwidth connectivity to take just a few examples. Services implemented at Cloud Computing centers are among those that place great demands on the Internet backbone and “last mile” connectivity to edge networks.

Many services can be decomposed into synchronous and asynchronous components, and different “Data Logistics” strategies applied to each part [41]. Techniques used in Content Delivery Networks, including caching and prestaging, can be applied on a fine-grained and even per-client basis. It is sometimes the case that the entire service can be implemented using edge resources. In other cases there is a component that can only be implemented using synchronous end-to-end datagram delivery across the backbone, but which requires only low bandwidth. In some cases analysis of the application combined with reconsideration of the truly necessary characteristics of the service delivered to the end-user can reduce the need for high quality synchronous connectivity to the vanishing point. In a sense, reliance on strong network assumptions is often used to trade off unnecessary reliance on excellent network infrastructure for ease of development. This is a useful strategy for those who can afford and support the necessary infrastructure.

Today, some environments cannot support strong network assumptions, even when local IT resources are available. Examples are communities isolated through geography, economic (poverty, discrimination), political circumstances (famine, war), or social factors. Disasters create environments where infrastructure is disrupted even in the most advanced societies. The recent response of modern network technologists has been to bring fixed or mobile wireless technology (satellite, 4G) into remote locations and to the scene of disasters or to create complex wireless infrastructures based on continuous aviation drones such as Google’s balloon-based project Loon and Facebook’s drone-based project Aquila. In contrast, using a mix of interoperable heterogeneous synchronous and asynchronous data transport integrated into a flexible platform to support a variety of distributed applications can be cheap, robust and easily deployed.

7.3 Big Data and Edge Processing

One of the inexorable trends in the collection of data is the emergence of large scale online sensors and instrument that produce data that must be subjected to volume-reducing processing before it can be passed over the network. Growing trends in sensor networks, the Internet of Things, and Smart Cities will severely exacerbate this problem [42]. The historical approach has been to send all such data to computation centers that are either self-contained or connected to their peers through heroic networking that may be private or even proprietary. This is no longer sufficient to address the total size of data, its globally distributed generation and consumption of data that we see today [43]. An alternative possible using EBP is to apply limited edge processing on the in the edge network using a converged infrastructure that can also store and transport data.

7.4 In-Locus Data Analysis

Data Analytics (DA) has emerged as a new paradigm for understanding unreliable and varying environments. Going beyond logging, reporting, and thresholding, DA can perform meaningful analyses of large scale data sources that are

networked through dynamic and distributed infrastructure. (The stage before batch or streaming analytics take place is often called “data assimilation”.) DA is capable of extracting latent knowledge and providing insight from field sensors, computational units, and large mobile networks. Of course the number of these data sources and the corresponding ingest rates are growing dramatically because of increased edge hardware capability (resolution and sampling rate) and hybridization (multi-messenger and multi-sensor data acquisition). These factors require new algorithmic approaches that closely integrate the network, I/O, and computational software stacks to lower the overheads and provide non-trivial data metrics at the edge. Fortunately, the Applied Mathematics and Machine Learning communities have recently produced innovations in the field of approximate and/or randomized algorithms, which combines new methods for matrix approximation via random sampling, that are perfectly positioned to fill this role.

Recent work in randomized and approximate algorithms [44, 45], which attach a probabilistic measure to their results, improves the fit of such methods for inherently unstable and constantly changing distributed environments. In fact, there are many statistical techniques in the Randomized Linear Algebra class of algorithms that lend themselves perfectly to utilization in the converged approach of in-locus computing, e.g., by using IBP’s best effort Network Functional Unit operations as discussed in Sect. 6.2. Such NFU operations can respond algorithmically to assimilate the inherent failures that naturally occur in a widely distributed system at the scale that we target. The iterative nature of most approximate methods allows us to incorporate erroneous responses from a sensor or a network transmission and gradually remove the malformed data from the multidimensional subspace that is being worked on. Similarly, an intermittent lack of response from a sensor or a network element may naturally be incorporated as a sampling and selection operator that is triggered by a system-reported event as opposed to the classical method that uses a pseudo random number generator (PRNG) as an unbiased projector or selector. Also, the probabilistic nature of the approximate algorithms allows us to weigh the data sources based on their history of reliable responses and the quality of the data they delivered (if a measure of quality can be obtained, from, for example, a duplicate sensor). High quality sensors and network connections will, over time, gain larger weights, in turn rendering them highly probable to be approximately correct as envisioned by the Probably Approximately Correct (PAC) learning framework [46].

The fundamental operators of randomized methods are *selection* and *projection*. They may be used in combination or individually, depending on the need. In our approach, we use the transmission errors as a form of selection, while projections would mostly be constructed to incorporate knowledge about the state of the system. In statistical parlance, we strive to obtain unbiased sampling of the incoming data. As long as the sample is representative and limited in size, we are able to process the data at the edge or afford the bandwidth to send closer to the core where more computing power resides. The prior distribution is constructed from known pieces of information, such as hardware specification

of the sensors and their reporting frequency. Over time, we may be able to form a more useful posterior distribution that is informed by changes in the infrastructure deployed in the field and in the surroundings that are being monitored by the sensors. In order to achieve this goal, the integration of the algorithmic methods and the system stack needs to occur in novel ways. Contrast this with the rather idealistic notions of the prevailing fault tolerant paradigms that tend to assume that errors occur discretely in the midst of reliable computing periods. It is assumed that error correction can restore the corrupted data through some form of redundant state management (e.g. checkpoints or error correcting codes) to maintain the logical invariant of the data being unaltered. We move away from these idealizations and incorporate the probability of errors directly into the probability of the answer being correct, using biased sampling to isolate valuable data that has high probability of being representative of the state of the system. The bias in our approach is guided by the information obtained either statically, before execution, or dynamically as the computation progresses.

8 Conclusions

In this paper, we have argued that interoperable convergence of storage, networking and processing is necessary in building a platform to support distributed systems which exhibits deployment scalability, and that the most effective implementation is a form of Exposed Buffer Processing at a layer below that which implements the Internet. Our argument rests on practical historical examples of the problems caused by the Internet's lack of generalized state management and an argument based on a partially formalized design methodology that the spanning layer of any converged infrastructure must be simple, generic and limited.

Acknowledgements. The ideas in this paper were influenced by many spirited discussions with Martin Swany on the integration of storage and processing with scalable networking, and by recent conversations with Glenn Ricart on the definition and justification of interoperable convergence. The concept of “exposed buffer protocol/processing” was coined during discussions between Swany and Beck, although its best definition and implementation are still subject to debate. The authors are also indebted to David Rogers for his professional rendering of the artwork in this any many other papers and presentations, and to Chris Brumgard for his helpful comments.

References

1. Mynatt, E., Clark, J., Hager, G., Lopresti, D., Morrisett, G., Nahrstedt, K., Pappas, G., Patel, S., Rexford, J., Wright, H., et al.: A national research agenda for intelligent infrastructure. *arXiv preprint arXiv:1705.01920* (2017)
2. Networking, I. T. Research, and D. N. Program.: Smart and Connected Cities Framework. <https://www.nitrd.gov/sccc/materials/scccframework.pdf> (2015)
3. Nahrstedt, K., Cassandras, C.G., Catlett, C.: City-scale intelligent systems and platforms. *arXiv preprint arXiv:1705.01990* (2017)

4. Anderson, T., Peterson, L., Shenker, S., Turner, J.: Overcoming the internet impasse through virtualization. *Computer* **38**(4), 34–41 (2005)
5. Tennenhouse, D.L., Wetherall, D.J.: Towards an active network architecture. *Comput. Commun. Rev.* **26**, 5–18 (1996)
6. Foster, I., Kesselman, C.: The Grid: Blueprint for a New Computing Infrastructure, pp. 47–48. Advanced computing. Computer systems design. Morgan Kaufmann Publishers (1999)
7. Chun, B., Culler, D., Roscoe, T., Bavier, A., Peterson, L., Wawrzoniak, M., Bowman, M.: Planetlab: an overlay testbed for broad-coverage services. *SIGCOMM Comput. Commun. Rev.* **33**(3), 3–12 (2003)
8. McGeer, R., Berman, M., Elliott, C., Ricci, R. (eds.): The GENI Book. Springer, Berlin (2016)
9. Clark, D.D.: Interoperation, open interfaces, and protocol architecture. Unpredictable Certainty White Pap. **2**, 133–144 (1995)
10. Uux(1p) posix programmer's manual (2013)
11. Saltzer, J.H., Reed, D.P., Clark, D.D.: End-to-end arguments in system design. *ACM Trans. Comput. Syst.* **2**(4), 277–288 (1984)
12. Carpenter, B., Brim, S.: Middleboxes: Taxonomy and issues. RFC 3234, Feb 2002. Network Working Group
13. Dongarra, J., Golub, G.H., Grosse, E., Moler, C., Moore, K.: Netlib and NA-Net: building a scientific computing community. *IEEE Ann. Hist. Comput.* **30**, 30–41 (2008)
14. Chankhunthod, A., Danzig, P.B., Neerdaels, C., Schwartz, M.F., Worrell, K.J.: A hierarchical Internet object cache. In: Proceedings of the 1996 USENIX Technical Conference, pp. 153–163 (1995)
15. Wessels, D., Claffy, K.: ICP and the Squid web cache. *IEEE J. Sel. Areas Commun.* **16**, 345–357 (1998)
16. Beck, M., Moore, T.: The Internet2 distributed storage infrastructure project: an architecture for internet content channels. In: Computer Networking and ISDN Systems, pp. 2141–2148 (1998)
17. Kirkpatrick, D.: IBM's Olympic Fiasco Department of Groundless Optimism. *Fortune Magazine*, 9 Sept 1996
18. Buyya, R., Pathan, M., Vakali, A. (eds.): Content Delivery Networks. Springer, Berlin (2008)
19. Nygren, E., Sitaraman, R.K., Sun, J.: The Akamai network: a platform for high-performance internet applications. In: SIGOPS Operating Systems Review (2010)
20. Beck, M., Moore, T., Plank, J.S.: An end-to-end approach to globally scalable network storage. In: ACM SIGCOMM 2002 (2002)
21. Beck, M., Moore, T., Plank, J.S.: An end-to-end approach to globally scalable programmable networking. In: Future Directions in Network Architecture, pp. 328–339. ACM Press (2003)
22. David, P.A.: Path dependence: a foundational concept for historical social science. *Cliometrica* **1**(2), 91–114 (2007)
23. Messerschmitt, D.G.: The convergence of telecommunications and computing: what are the implications today? *Proc. IEEE* **84**(8), 1167–1186 (1996)
24. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. *Commun. ACM* **51**(1), 107–113 (2008)
25. Peterson, L.L., Davie, B.S.: Computer Networks: A Systems Approach, 5th edn. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2011)
26. Beck, M.: On the hourglass model, end-to-end arguments, and deployment scalability. *Commun. ACM* (to appear 2018)

27. Will the real end-to-end argument please stand up? <http://mercury.lcs.mit.edu/jnc/tech/end.end.html>
28. Ritchie, D.M., Thompson, K.: The Unix time-sharing system. *Commun. ACM* **17**, 365–375 (1974)
29. Satyanarayanan, M.: The emergence of edge computing. *Computer* **50**(1), 30–39 (2017)
30. Plank, J.S., Bassi, A., Beck, M., Moore, T., Swany, D.M., Wolski, R.: Managing data storage in the network. *IEEE Internet Comput.* **5**(5), 50–58 (2001)
31. Beck, M., Arnold, D., Bassi, R., Berman, F., Casanova, H., Moore, T., Obertelli, G., Plank, J., Swany, M., Vadhiyar, S., Wolski, R.: Logistical computing and inter-networking: middleware for the use of storage in communication. In: In 3rd Annual International Workshop on Active Middleware Services (AMS) (2001)
32. Plank, J.S., Atchley, S., Ding, Y., Beck, M.: Algorithms for high performance, wide-area, distributed file downloads. Technical report, Letters (2002)
33. Beck, M., Liu, H., Huang, J., Moore, T.: Scalable distributed execution environment for large data visualization. In: IEEE Explorer, Nov 2007
34. Liu, H.: Scalable, data-intensive network computation. Ph.D. thesis, University of Tennessee, Knoxville (2008)
35. Liu, H., Beck, M., Huang, J.: Dynamic co-scheduling of distributed computation and replication. In: IEEE International Symposium on Cluster Computing and the Grid, May 2006
36. Kitchin, R.: The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences. Sage, London (2014)
37. Reichman, O.J., Jones, M.B., Schildhauer, M.P.: Challenges and opportunities of open data in ecology. *Science* **331**(6018), 703–705 (2011)
38. Board, S.S., Council, N.R., et al.: Landsat and Beyond: Sustaining and Enhancing the Nation’s Land Imaging Program. National Academies Press (2014)
39. Bird, I.: Computing for the Large Hadron Collider. *Annu. Rev. Nucl. Part. Sci.* **61**, 99–118 (2011)
40. Breeding, M.: Building a digital library of television news. *Comput. Libr.* **23**(6), 47–49 (2003)
41. Wikipedia Contributors. Information logistics (2017) (Online). Accessed 15 Sept 2018
42. Banerjee, S., Wu, D.O.: Final report from the NSF Workshop on Future Directions in Wireless Networking. National Science Foundation (2013)
43. Chen, M., Mao, S., Liu, Y.: Big data: a survey. *Mobile Netw. Appl.* **19**(2), 171–209 (2014)
44. Avron, H., Maymounkov, P., Toledo, S.: Blendenpik: supercharging LAPACK’s least-squares solver. *SIAM J. Sci. Comput.* **32**(3), 1217–1236 (2010)
45. Drineas, P., Mahoney, M.W.: RandNLA: randomized numerical linear algebra. *Commun. ACM* **59**(6), 80–90 (2016)
46. Valiant, L.G.: A theory of the learnable. *Commun. ACM* **27**, 1134–1142 (1984)



QoS for SDN-Based Fat-Tree Networks

Haitham Ghalwash^(✉) and Chun-Hsi Huang

Department of Computer Science and Engineering, University of Connecticut,
Storrs, CT 06269, USA
[{haitham.ghalwash, chunhsihuang}@uconn.edu](mailto:{haitham.ghalwash,chunhsihuang}@uconn.edu)

Abstract. Software-defined Networks (SDNs) are the new network paradigm providing, programmability, agility, and centralized management. In this paper, we show how to leverage the SDN centralized controller to improve the network utilization and the traffic performance. On top of the SDN controller, new modules are added to help finding single and multi-path routes between communicating devices. Flow rules are automatically installed into the designated switches to provide the required paths. The behavior and performance of different types of traffic, namely, UDP, TCP, VOIP, and a Big-data application traffic are investigated. The traffic forwarding is based on either the controller built in layer 2 switching “*odl-l2switch*” feature or single/multi-path selection based on the supplemented modules. Experimental results based on metrics such as delay, jitter and packet drops are presented for each forwarding option. The results disclosed the advantage of having the developed modules on top of the controller for all traffic types. The *OpenDaylight* controller for OpenFlow switches, in a fat-tree network, is used for experiments. For a fair comparison of different traffic types, a monitoring module is built on top of the controller for collecting ports statistics, analyzing and monitoring.

Keywords: QoS · SDN · Fat-Tree · Docker · Hadoop

1 Introduction

The rapid growth in information processing is pushing todays network traffic size and scale to a crazy limit. As the cloud grows and more virtualization techniques are deployed, network engineers are seeking a new architecture to fulfill such demands. Traditional networks are suffering from the rapid traffic growth and will not survive facing such challenge. A forecast from cisco white paper [1] stated that, by the year of 2021, the number of devices connected to IP networks will be three times as high as the global population. Annual global IP traffic will reach 3.3 ZetaByte per year. In 2016, global IP traffic was 1.2 ZB per year. By 2021, global fixed broadband speeds will reach 53.0 Mbps, compared to 27.5 Mbps in 2016. Moreover, the new trends in cloud computing, visualization, wireless networks, IoTs, and bigdata applications, are hardening the need for a software-oriented network architecture.

Software-Defined Networking (SDN) is a promising approach, where the control plane is decoupled from the data plane, thus providing a directly programmable, extremely dynamic, easily used and adaptable network. SDN proved its efficiency over the traditional network in many cases. A study in [2] considering torus and hypercube

topologies demonstrated that SDN outperforms traditional networks by 45% in throughput when 256 servers were used. Moreover, a previous study in [3] showed that using SDN for a Hadoop application superseded normal forwarding mode when tested over different network scales. SDNs also proved its efficiency in solving the lack of information, hard management, and hard QoS guarantee.

In this paper, a new approach is presented, using external modules, to provide a level of QoS in SDN-based Networks. The proposed modules provide monitoring, route determination, rule preparation, and configuration functionalities. The experimental part investigates the network operation, considering different applications, supervised by an SDN controller after and before adopting the developed modules. The recorded results show the superiority of the network operation with the developed module.

The paper is organized as follows. In Section 2, the SDN architecture for the current and the next generation data centers is presented. The Literature review is found in Sect. 3. Section 4 explores the SDN simulation with the proposed modules, and Sect. 5 discusses the experimental setup and configuration. The results are presented in Sect. 6. Finally, Sect. 7 has the concluding remarks and future work.

2 Software Defined Networks Architecture

Recently, the integration between the applications and network configurations has gained a great interest in both industry and academia. SDN has emerged in response to limitations of traditional networking architectures. SDN is now decoupling the control plane from the forwarding plane to help improving security, performance and management issues. The networks' administrators can easily and efficiently manage the services and applications and separate them from the infrastructure forwarding plane. Moreover, SDN provides a global view of the network that turns out to be promising in improving the performance of various network applications including Big-data applications. According to the Open Networking Foundation (ONF) [4], SDN architecture is presented in three distinct planes, namely, Application Plane, Control Plane, and Data Plane, see Fig. 1.

The top layer (Application Plane), consists of a number of applications implementing the desired network behavior and purpose. An application can be added to any SDN controller, either internally in the controller, or externally using open interface for communication (e.g. REST). An application communicates directly with the controller through northbound interfaces. The middle layer (Control Plane), also known as the controller, is logically centralized and may consist of one or more controller. It is responsible for mapping all desired network actions from the top layer (Application Plane) to the lower layer (Data Plane), via open interfaces. Control Plane is also responsible for providing the required abstracted information about the network to the requesting applications. The Infrastructure Layer (Data Plane), comprises all types of physical/virtual network elements. Its main function is packet forwarding. The controller interacts with all network devices through the southbound interface using dedicated control messages and APIs, e.g. OpenFlow [5] protocol.

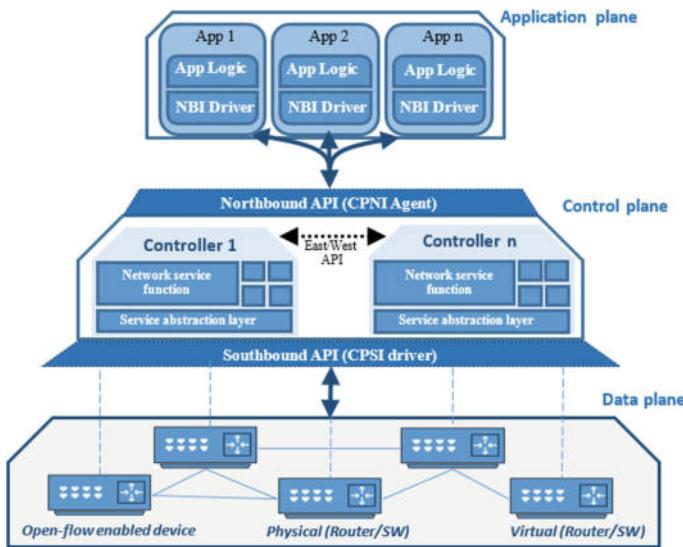


Fig. 1. SDN architecture

Over the past few years, controllers with variety of features and capabilities were developed. *OpenDaylight* (ODL) and the Open Network Operating System (ONOS), appear at the far end of recent controllers. The ODL is a Linux Foundation collaborative project, written in java and is highly supported by various companies. ONOS, on the other hand, is a “private” development effort (managed by ON.Lab). It is mainly designed for wide area networks (WAN) and service provider networks. In this paper, ODL is adopted for being more generic and serving a wide range of applications.

To provide QoS in any Network, a monitoring and management facilities are required. SDNs help network administrators easily and efficiently create automated QoS management frameworks by considering resource reservation, queue management, and packet scheduling. The SDN controller is able to easily acquire the global view/state of the network, that facilitates QoS configuration and monitoring. QoS mainly differentiate traffic within the available resources based on some evaluated metrics such as bandwidth, delay, jitter, and packet dropped. Monitoring is another feature that is easily supported through SDN controllers. Detecting threats and performance issues with SDN controllers becomes handful in nearly real time with the possibility of predicting future network behavior.

3 Literature Review

SDN controller is logically centralized and has a global vision of the whole network. A controller can detect the change of status and dynamically optimizes resources to enhance the network performance. Recently, QoS through SDN is gaining the interest

of researchers. QoS is encountered in SDN networks via flow routing, resource reservation, QoE-Aware mechanisms, monitoring mechanisms, queue management and scheduling, and some other QoS-oriented mechanisms [6].

A work presented in [7], developed an application identification technique based on SDN controller to determine the QoS levels for different types of applications. The application flows were queued with different priorities based on the application type. Experimental results showed a 28% reduction in the average delay. Another work in [8], developed a framework to guarantee QoS for a specific flow. Traffic was classified as it gets into the edge switch and automatically rerouted, when the network was congested, based on a queuing technique that satisfy the required service level. In [9], HiQoS was proposed using multiple paths between source and destination. It also presented queuing mechanisms to guarantee QoS for various types of traffic. In HiQoS, a modified Dijkstra algorithm enabled in assigning multiple paths that satisfy certain QoS constraints. Experimental results showed a reduction in delay and an increase in throughput. The work in [10], proposed a framework for flow classification and rate-shaping. The implemented modules classified the packets and assigned a predefined priority based on the header information. Applying rate-shaping, the appropriate flows are installed.

SDN QoS in Big-data applications was discussed in [11] where Advanced Control Distributed Processing Architecture (ACDPA) was proposed. SDN architecture along with Hadoop were used for network control and processing large amount of data plane. After Hadoop processes data, the SDN controller assigned priorities and programed flow to switches. In [12], an SDN-aware version of Hadoop application was set up to use prioritized queues in the underlying OpenFlow switches. Critical Hadoop traffic was routed before other less critical traffic, and the jobs were completed faster. A similar study in [13] proposed a cross point queued (CICQ) switches to schedule packets for various Big-data applications. The switches scheduled packets based on the bandwidth provisioning table that is set by the controller for different Big-data applications. For monitoring, a work in [14] introduced how collected traffic statistics be used for calculating the available bandwidth on each link in the network.

Most of the presented work considered a single application with very small network topologies, that are customized for the experiment. The network byte load was not considered as an indicator of performance. Moreover, to the best of our knowledge, previous work neglected the built-in operation of “*OpenDaylight*” in the comparison.

4 SDN Simulation and Proposed Modules

In literature, SDN simulations are usually conducted using an open-source controller, and a container-based emulation environment Mininet [15]. Mininet enables creating a set of virtual nodes, which can be connected to form any arbitrary network. It creates actual Linux host instances, which enables the generation of traffic. The controller used is *OpenDaylight*, which is a collaborative open-source project hosted by the Linux Foundation. The topology used is Fat-tree [16] topology which is widely used for its simplicity, and has been considered in a number of recent HPC data centers [17]. Fat-tree is typically adopted as two to three levels [18] of switches, it is well suited for

building large-scale high-performance data centers and providing scalable bandwidth at a moderate cost [16]. In a previous work [19], fat-tree proved to be a better choice as the network scales up.

The *Fat-tree* topology consists of n -*pods*, each with two layers of $n/2$ (*n-port*) switches and $n^2/4$ servers. Each edge switch is directly connected to $n/2$ servers and $n/2$ aggregation switches. Each aggregation switch is further connected with $n/2$ *n-port* core switches. There is a total of $n^2/4$ of *n-port* core switches, where each core switch has one port connected to each pod. A 4-pod *Fat-tree* including, from the bottom, hosts, edge switches, aggregation switches, and core switches, are shown in Fig. 2.

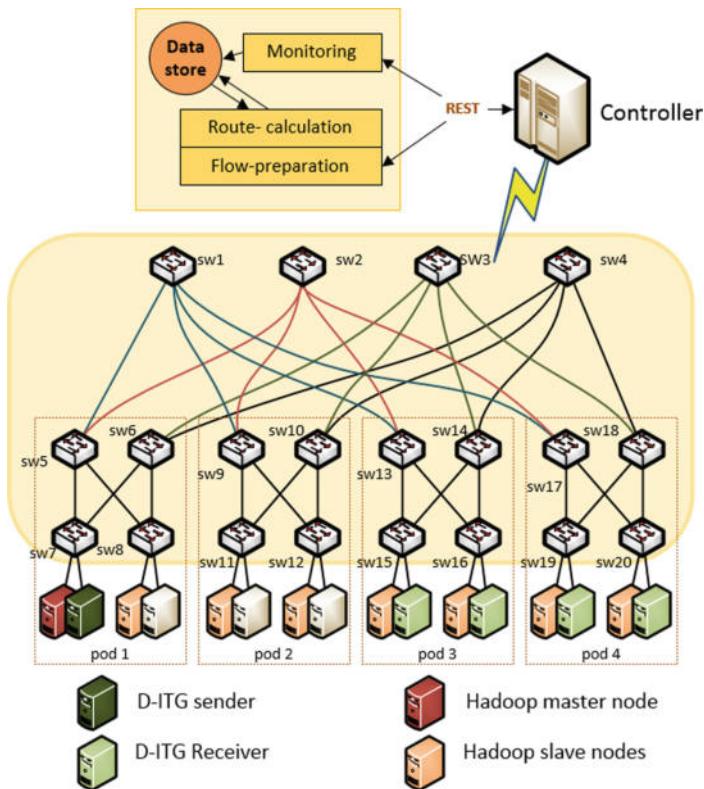


Fig. 2. Fat-tree 4-port 3 levels topology and the controller

Figure 2 shows the reference architecture of our experiment. It consists of an *OpenDaylight*, “carbon” release, controller communicating with OpenFlow switches in a fat-tree topology. A total of 16 hosts are connected to the edge switches. The hosts are either, Hadoop nodes, D-ITG traffic sender or D-ITG traffic receiver. The proposed modules are added on top of SDN as an external SDN application and uses REST to communicate with the controller. The SDN controller performs the basic functions such as discovery, managing and collecting network statistics. It is also responsible for

mapping the application requirements, by installing the needed flows in the designated OpenFlow switch. The developed modules include (1) *Route-calculation* for single-path/multi-path routes, (2) *Flow-preparation* for flow programming on switches, and (3) *Monitoring* module for collecting, storing and analyzing statistics from the controller's operational data store.

In the first module, “*Route-calculation*”, the module computes the appropriate routes based on the source and destination IP's and applications port numbers. A *single* or a *multi-path* route is calculated based on the graph stored in the “*data store*”. In the *single-path* operation, the unweighted shortest path is calculated, based on the source IP, destination IP addresses and the graph information. In the *multi-path* route calculation, the module calculates all possible short paths between the source and destination, compare them using the “*SequenceMatcher*” class from the “*difflib*” python module, and finally considers the two paths with the least similarity ratio value. The two paths are selected with the minimum overlapping hops, maximum disjoint links. This minimizes the shared links and reduces the congestion between communicating nodes. The traffic is classified based on the application type, from the header information of the packet, to be placed on different paths.

Finally, for both route calculations, the selected paths are passed to the “*Flow-preparation*”. The “*Flow preparation*” module defines the forwarding nodes and ports, prepare the required forwarding rules, and inform the controller to install these rules into the designated switches. The last developed module, “*Monitoring*” module, extracts the topology information, collecting node status, and preparing ports utilization statistics. The topology information and ports' statistics are extracted from the controller through REST APIs. All ports' statistics are extracted over an interval of 10 s. The module also identifies the edge switches to be solely monitored. Edge switches Monitoring will help identifying the ingress and egress traffic to and from the network. All statistics and information are maintained locally in Matrix forms and saved time stamped into files. The analyzed data used in this paper presents the switch ports utilization each 10 s, see Figs. 6, 7 and 10.

5 Experiment Configuration

As shown in Fig. 2, the experiment consists of the *OpenDaylight* controller with the added external application modules. The application modules are communicating with the controller through REST APIs. The connecting topology is a fat-tree of 4-pods. The topology is accommodating 4-port capacity switches organized in 3 levels, with a total of 16 hosts and 20 switches. Different types of traffic are examined, namely, UDP, TCP, VOIP, and a Big-data Hadoop read/write traffic. Figure 2 shows the nodes engaged in each type of the traffic generated. In this paper, a two phases experiment is conducted. In the first phase, only the UDP, TCP and VOIP traffic are considered. In the second phase, all types of traffic including the running Hadoop Big-data application are involved.

To generate UDP, TCP and VOIP traffic, *Distributed Internet Traffic Generator* (D-ITG) [20] is used. D-ITG is capable of generating traffic at network, transport, and application layers. D-ITG is used to generate a total of 3 flows, a flow of each type,

from the sender to each receiver. Experiments involves 4 *receiver agents*, one in each pod, forming a total of 12 flows in the network. The packet size is set to 512K bytes, with constant inter-departure time of 2, 4 and 8K packets per second. The execution time is set to 200 s per run with a total of 10 runs. The delay, jitter, and packets dropped are calculated for each communication flow. The ITGDec decoder is the utility used to analyze the results of the experiments conducted. ITGDec parses the log files generated by sender and receiver agents and calculates the average values of bit rate, delay, and jitter either for the whole duration of the experiment or for variable-sized time intervals. For performance analysis, the calculated packets dropped, minimum delay, maximum delay, average delay, and average jitter are adopted as the performance metrics.

In phase two, UDP, TCP, and VOIP traffic are added to a Hadoop running read/write benchmark. Hadoop is widely used in Big-data applications in the industry. It mainly consists of *MapReduce* and *Hadoop Distributed File System (HDFS)*. *MapReduce* is a software framework for processing and generating distributed data from a cluster. *HDFS*, the storage part, mainly consists of a single master node, *NameNode*, and many slave nodes, *DataNodes*. Data is divided into fixed-size blocks and is stored across all *DataNodes*. The *NameNode* handles job requests, divides jobs into tasks, and assigns each task to a *DataNode*. The developed scenario involves a multi-node Hadoop cluster, as shown in Fig. 2. Although the Hadoop cluster may be set using virtual machines, a huge number of CPUs and RAMs are required for this scenario. Therefore, the presented design relies on lightweight software Linux containers, namely, *Dockers*. Hadoop nodes are implemented on separated Docker containers and connected to the Mininet edge switches [21]. For the network traffic, *TestDFSIO*, a common tool for benchmarking single/multi-node Hadoop cluster is used. The *TestDFSIO* benchmark is a read/write test for Hadoop file systems. It is helpful for the discovery of performance bottlenecks in the network, and mainly for testing IO operation. The simulation is conducted using Mininet 2.2 running on Ubuntu machine of 32 GB RAM with a Xeon processor. The Controller is ODL “carbon” release, running on a separate Ubuntu personal machine of 8 GB RAM.

6 Simulation Results

Results of the five performance metrics, namely, Average delay, Minimum delay, Max delay, Average Jitter, and Packets dropped, are shown in Figs. 3, 4, and 5. Each figure displays three different applications traffic, UDP, TCP, and VOIP traffic. Each application is tested for different packet rates of 2K, 4K, and 8K packets per second and the average is considered for plotting. For each traffic type, three modes of packet forwarding are tested, namely, *odl-l2switch*, *single-path* and *multi-path* forwarding.

Figures 3a–c, shows the minimum, average and maximum delays on the same plot. The minimum, average and maximum delays are improved as the forwarding mode changes form *odl-l2switch* to *single-path*, and *multi-path* forwarding with the best performing *multi-path* forwarding. The improvements are noticed over all traffic rates and for all types of traffic. In Fig. 4a–c, using *single-path* and *multi-path* routing outperforms the *OpenDaylight* normal switching, *odl-l2switch*, in terms packet dropped

Minimum, Average, and Maximum Delay

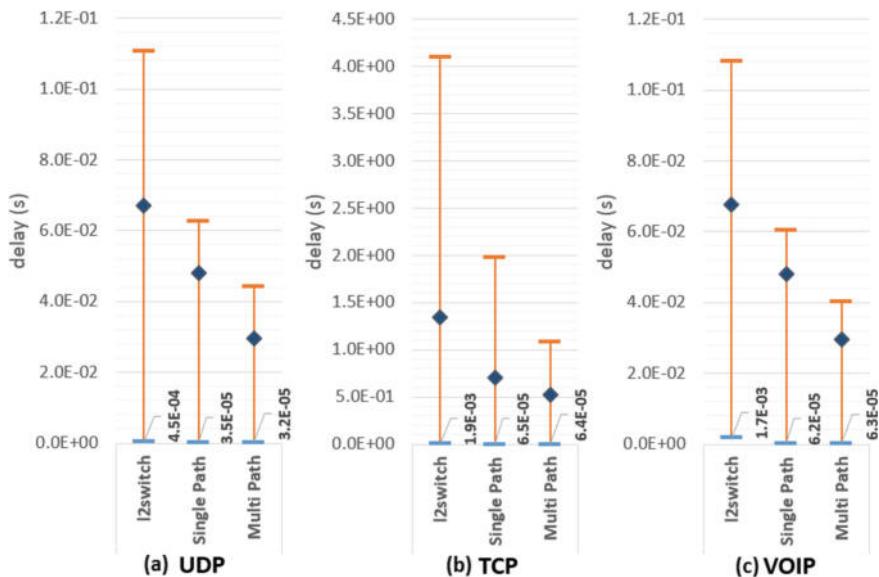


Fig. 3. Delay (UDP, TCP, and VOIP)

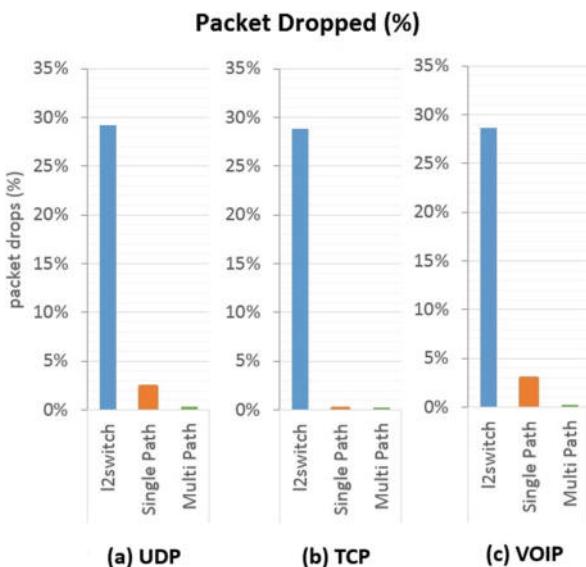


Fig. 4. Packet dropped (UDP, TCP, and VOIP)

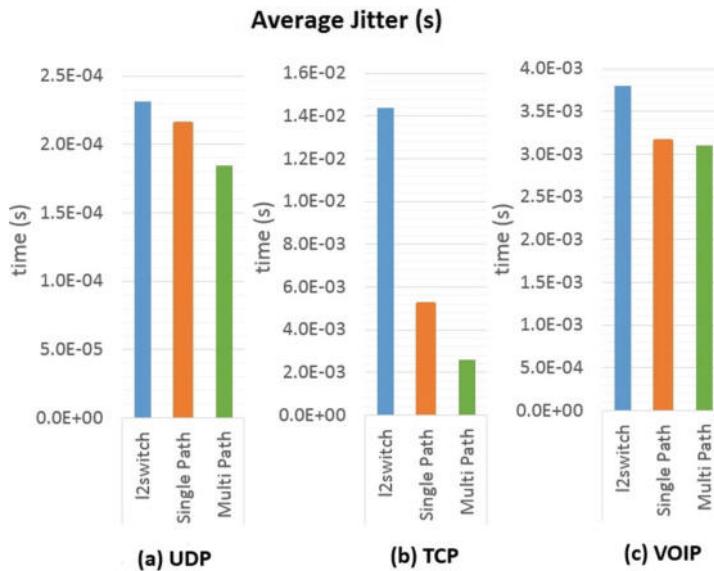


Fig. 5. Average jitter (UDP, TCP, and VOIP)

for all traffic types. *Multi-path* routing again records the best performance over all traffic types. Figures 5a–c *multi-path* forwarding recorded nearly the best average jitter over all traffic types and rates. Meanwhile, the average jitter for *single-path* forwarding outperforms *odl-l2switch* for all traffic rates and types.

The Monitoring module is used to analyze the port status during the experiments as shown in Figs. 6 and 7. Figure 7a–c show the switches traffic load over all ports in bytes for the three forwarding modes *odl-l2switch*, *single-path*, and *multi-path*. Figure 7a shows the highest traffic in the network switches when using the *odl-l2switch* mode, recording 338 M bytes on average, and a max of 695M bytes over all switches traffic. In Fig. 7b and c, using the *single-path* and *multi-path* forwarding reduced the overall average bytes to around 161M and 177M bytes, respectively. The maximum bytes are reduced from 695M in the *odl-l2switch* to 458 M in the *single-path* and down to 440M in the *multi-path*. The monitoring module can also be used to analyze per port traffic as illustrated in Fig. 6. Figure 6, shows how the traffic load is reduced on port #1 of switch #6 when using *multi-path* forwarding over *single-path* forwarding.

For the second phase of the experiment, a Big-data Hadoop traffic is added into the network. The *testDFSIO* benchmark is used for writing a 4 GB file over the 8 different Hadoop cluster nodes. The throughput and execution time are recorded for both read and write operations, as shown in Figs. 8 and 9. *TestDFSIO* is used to test the Hadoop read and write traffic either solely or with the existence of other traffic types in the network. As shown in Figs. 8a, b and 9a, b, the 3 left bars show the Hadoop traffic running solely (without UDP, TCP or VOIP traffic) in the network for each of three forwarding modes *odl-l2switch*, *single-path*, and *multi-path*. In this case, for the *multi-path* forwarding, subset of the Hadoop nodes traffic is routed on the second path. For

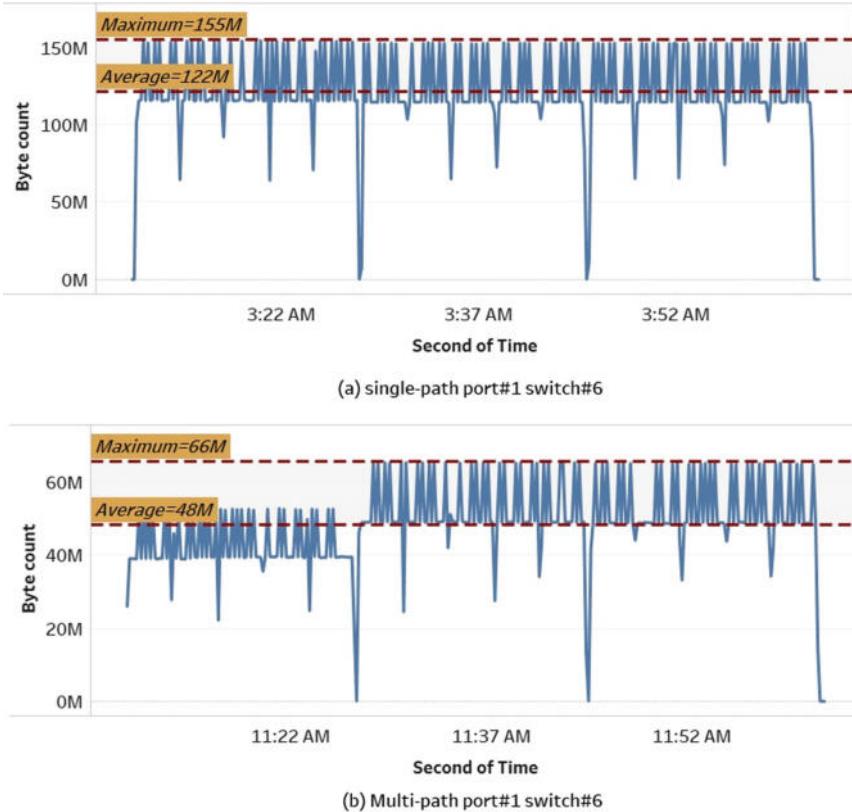


Fig. 6. Monitoring port #1 on switch #6

the *testDFSIO* Hadoop traffic only running, Figs. 8a, b and 9a, b show nearly the same throughput and execution-time for all forwarding modes.

Next, the Hadoop traffic is tested in addition to TCP, UDP, and VOIP. The read/write throughput and execution time is again calculated as the Hadoop is running with other traffic in the network. Figures 8a, b and 9a, b show the *odl-l2switch*, *single-path*, and *multi-path* forwarding modes for read and write operations in right 3 bars. The read and write throughput increases as forwarding mode changes from *odl-l2switch* to *single-path* and *multi path*, see Fig. 8a, b. The execution time, Fig. 9a, b, is also reduced when using the *multi-path* over the *single-path* and *odl-l2switch*.

Figure 10a–c show the monitored forwarded traffic over each switch ports for the Hadoop only traffic in the network (without UDP, TCP or VOIP). Using the developed modules reduces the average traffic from an average of 100K in *odl-l2switch* forwarding to around an average of 18K bytes in the *single-path* and the *multi-path* forwarding. Moreover, the maximum switch ports traffic is reduced from an average of 219K in the *odl-l2switch* to nearly 108K bytes for the *single-path* and *multi-path* modes. Figure 10b, also pointed out the traffic in core switch #4, aggregation switch #6 and the edge switch #7 with a higher average over other switches. Using the *multi-path*

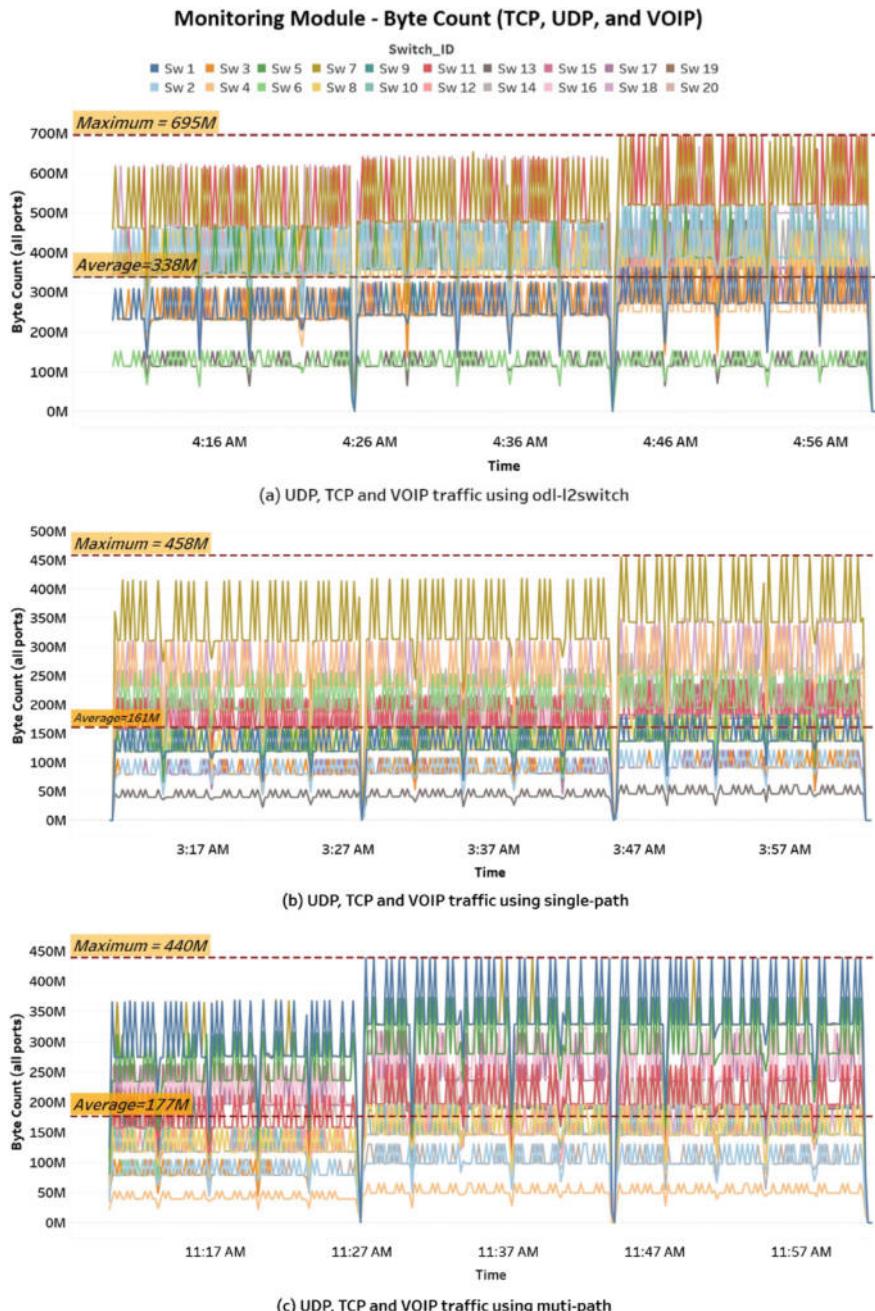


Fig. 7. All switches traffic over time UDP, TCP, and VOIP traffic

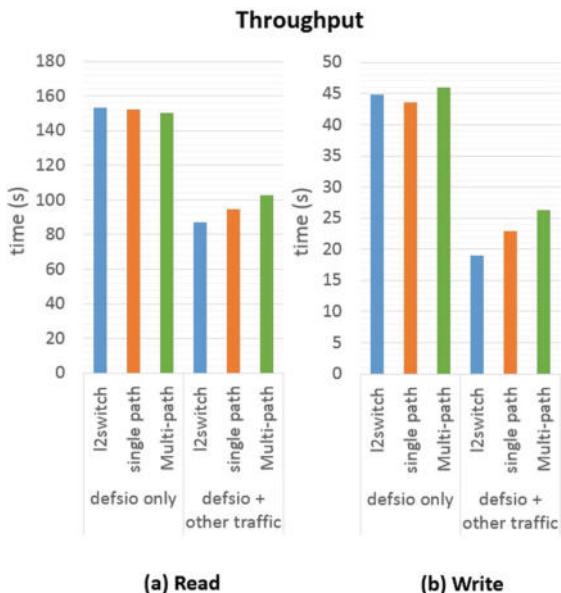


Fig. 8. *TestDFSIO* throughput

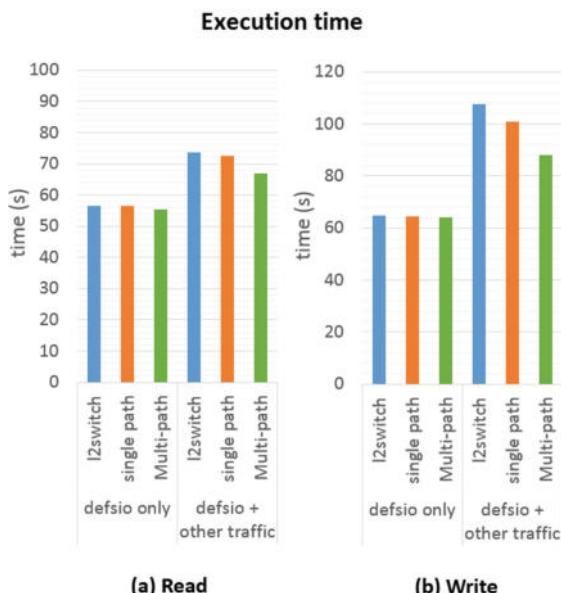


Fig. 9. *TestDFSIO* execution time

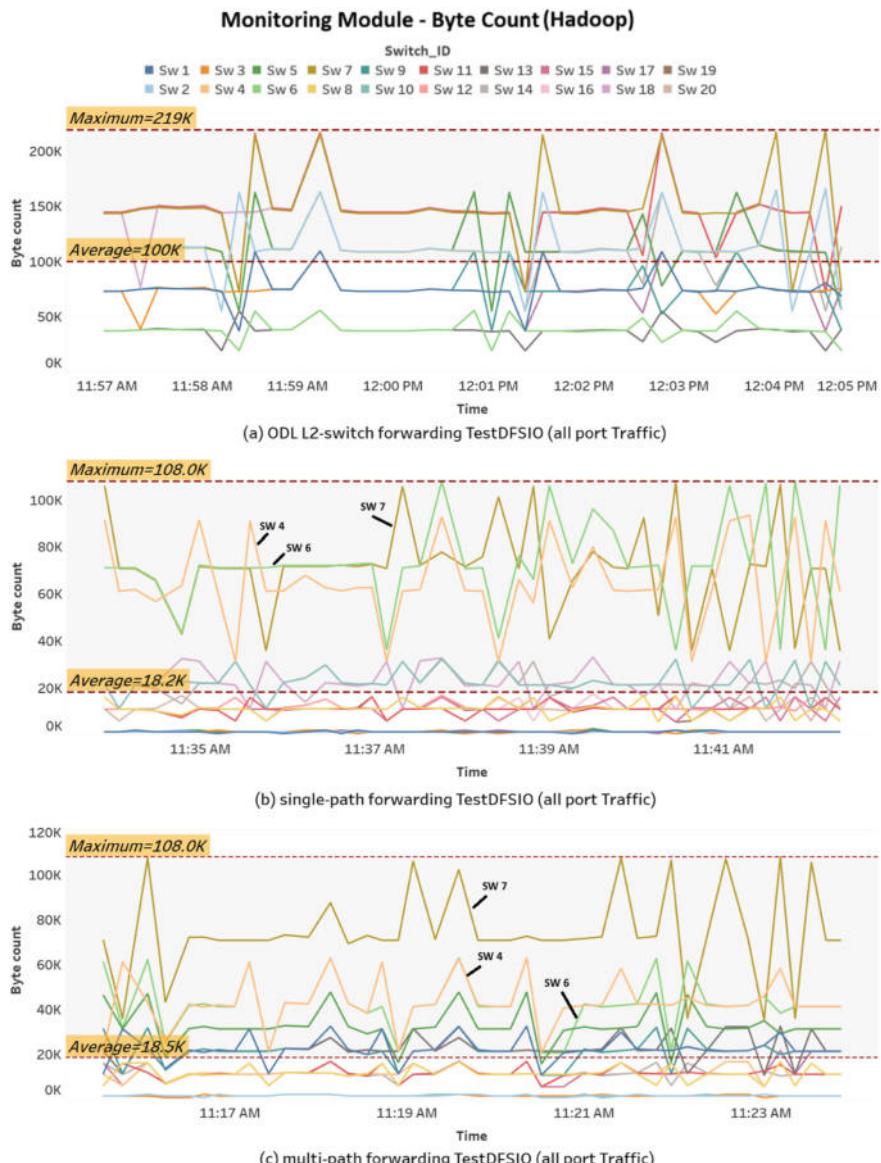


Fig. 10. *TestDFSIO* traffic form monitoring module

forwarding reduced the average traffic for both switch #4 and switch #6, as shown in Fig. 10c. It is worth mentioning that the Hadoop is showing lower overall traffic in this experimented scale, compared to the first phase of UDP, TCP and VOIP traffic.

7 Conclusion and Future Work

This paper presented QoS modules providing classification of data based on the application type. A Route-calculation and network monitoring modules were added to an *OpenDaylight* controller to enhance the traffic flow. The route calculation module helped routing subset of the traffic on a separate link. This simply reduced the load over the original link and provided better performance for all running network applications. Traffic flows are differentiated over the available paths based on the application type. The developed modules are implemented in python and use REST APIs to communicate with the controller. The network topology was a fat-tree of OpenFlow switches that was generated using MININET. Docker containers were attached as hosts to the topology for the Hadoop running application. All traffics in the network were tested under different modes of operation, namely, *odl-l2switch*, *single-path*, and *multi-path*. *Od़l-l2switch* is the build in switching module in *OpenDaylight*.

Adding the presented modules to the SDN controller outperformed the standalone SDN controller with the *odl-l2switch* build in feature. The overall traffic in the network was reduced for all switches. Moreover, *Multi-path* mode further reduced the traffic on congested links. The results indicated higher performance in term of delay, jitter and packet dropped for the *multi-path* over the *single-path* forwarding mode. Both operating modes proved to be more efficient than using the *OpenDaylight* layer 2 switching. For a fair comparison, a module for collecting ports statistics, analyzing and monitoring the switch ports was developed. The monitored traffic over switches showed that *single-path* and *multi-path* forwarding reduced the average and maximum switches traffic over the normal *odl-l2switch* forwarding.

In the near future, further verification research to be conducted on more complicated and large-scale network environment. Queuing mechanisms will be considered for the single and multipath routes. Moreover, the rout selection will consider other metrics.

Acknowledgements. This work was supported by the U.S. Department of Education's GAANN Fellowship through the Department of Computer Science and Engineering at the University of Connecticut.

References

1. Cisco Systems: Cisco Visual Networking Index: Forecast and Methodology, 2015–2020. White Paper (2016)
2. Andrus, B., Vegas Olmos, J.J., Mehmeri, V., Monroy, I.T., Spolitis, S., Bobrovs, V.: SDN data center performance evaluation of torus and hypercube interconnecting schemes. In: Proceedings—2015 Advances in Wireless and Optical Communications, Riga, Latvia, pp. 110–112 (2015)
3. Ghalwash, H., Huang, C.: Software-defined extreme scale networks for bigdata applications. In: High Performance Extreme Computing Conference, Waltham, MA, USA (2017)
4. Fundation ONF: Software-Defined Networking : The New Norm for Networks. ONF White Paper (2012)

5. McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L., Rexford, J., Shenker, S., Turner, J.: OpenFlow: enabling innovation in campus networks. *ACM SIGCOMM Comput. Commun. Rev.* **38**, 69–74 (2008)
6. Karakus, M., Durresi, A.: Quality of service (QoS) in software defined networking (SDN): a survey. *J. Netw. Comput. Appl.* **80**, 200–218 (2017)
7. Li, F., Cao, J., Wang, X., Sun, Y.: A SDN-based QoS guaranteed technique for cloud applications. *IEEE Access* **5**, 229–241 (2017)
8. Xu, C., Chen, B., Qian, H.: Quality of service guaranteed resource management dynamically in software defined network. *J. Commun.* **10**, 843–850 (2015)
9. Yan, J., Zhang, H., Shuai, Q., Liu, B., Guo, X.: HiQoS: an SDN-based multipath QoS solution. *China Commun.* **12**, 123–133 (2015)
10. Trajano, A.F.R., Fernandez, M.P.: uLoBal : Enabling In-Network Load Balancing for Arbitrary Internet Services on SDN, pp 62–67 (2016)
11. Desai, A.: Advanced Control Distributed Processing Architecture (ACDPA) Using SDN and Hadoop for Identifying the Flow Characteristics and Setting the Quality of Service (QoS) in the Network, pp. 784–788 (2015)
12. Narayan, S., Bailey, S., Daga, A.: Hadoop acceleration in an openflow-based cluster. In: *Proceedings—2012 SC Companion: High Performance Computing, Networking Storage and Analysis, SCC* (2012)
13. Hong, W., Wang, K., Hsu, Y.H.: Application-aware resource allocation for SDN-based cloud datacenters. In: *Proceedings—2013 International Conference on Cloud Computing and Big Data*, pp. 106–110, Santa Clara, CA, USA (2013)
14. Hamad, D.J., Yalda, K.G., Okumus, I.T.: Getting traffic statistics from network devices in an SDN environment using OpenFlow. In: *Information Technology and Systems 2015, Sochi, Russia*, pp. 951–956 (2016)
15. Lantz, B., Heller, B., McKeown, N.: A network in a laptop: rapid prototyping for software-defined networks. In: *Proceedings of the Ninth ACM SIGCOMM Workshop on Hot Topics in Networks—Hotnets ’10*, pp. 1–6, Monterey, CA, USA (2010)
16. Al-Fares, M., Loukissas, A., Vahdat, A.: A scalable, commodity data center network architecture. *ACM SIGCOMM Comput. Commun. Rev.* **38**, 63–74 (2008)
17. Saleh, A.: Evolution of the architecture and technology of data centers towards exascale and beyond. In: *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference, Anaheim, California, USA* (2013)
18. Bradonjić, M., Saniee, I., Widjaja, I.: Scaling of capacity and reliability in data center networks. *Perform Eval. Rev.* **42**, 3–5 (2014)
19. Ghalwash, H., Huang, C.: On SDN-based extreme-scale networks. In: *High Performance Extreme Computing Conference, Waltham, MA, USA* (2016)
20. Botta, A., Dainotti, A., Pescap, A.: A tool for the generation of realistic network workload for emerging networking scenarios. *Comput. Netw.* **56**, 3531–3547 (2012)
21. Peuster, M., Karl, H., Van Rossem, S.: MeDICINE : rapid prototyping of production-ready network services in multi-PoP environments. In: *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks, Palo Alto, California, USA* (2016)



LBMM: A Load Balancing Based Task Scheduling Algorithm for Cloud

Yong Shi^(✉) and Kai Qian

Department of Computer Science, Kennesaw State University,
1100 South Marietta Pkwy, Marietta, GA 30060, USA
yshi5@kennesaw.edu

Abstract. As one of the fields in Computer Science research, Cloud Computing has attracted attentions from industries as well as academia in recent years. Numerous topics have been studied related to Cloud Computing, and one of them is task scheduling. Task scheduling is the strategy to assigning various tasks to certain resources. Existing task scheduling algorithms include Min-Min, Suffrage, Max-Min and many more, in which Max-Min is efficient in minimizing the completion time of tasks and producing a good task schedule, however, it has a drawback of load unbalancing. To address this issue, we design an algorithm called LBMM for task scheduling considering load balancing as the key concept. We conduct our experiments using CloudSim package which is a framework for simulating activities in the Cloud systems. The experimental results demonstrate that our algorithm decreases the completion time and improves load balancing of resources, and it outperforms the traditional Max-Min and Min-Min.

Keywords: Task scheduling · Load balancing · Cloud computing · Max-Min algorithm · Cloud simulation

1 Introduction

Cloud Computing provides resource sharing and allocation in an efficient and affordable way so that a large number of parties could benefit from the ecosystem it creates. Please note that the first paragraph of a section or subsection is not indented. The first paragraphs that follows a table, figure, equation etc. does not have an indent, either.

Using services of various levels provides by the Cloud, companies, organizations and schools can easily work on their projects without having to build their own computing infrastructures, and they would have virtually unlimited available resources such as computing power, storage, network, etc., which greatly cuts down the resource provisioning and avoids up-front commitments. In this way, small companies can start without a lot of commitments and they can conveniently adjust their demand depending on their business status.

A well-recognized model for Cloud services is called SPI which represent three layers of services provided by Cloud. In this model, the lowest layer is IaaS which is the Infrastructure as a Service, the middle layer is PaaS which is the Platform as a Service, and the upper layer is SaaS which is the Software as a Service. PaaS uses

services provided by IaaS, and SaaS uses services provided by PaaS. Numerous companies provide these services including Google, IBM, Microsoft, etc.

With benefits mentioned above, Cloud stimulates continuous industry investment and provides growing job market.

There are many research activities related to Cloud Computing, which include network topology, task scheduling, distributed computing, etc. Modern systems are often required to execute multiple processes simultaneously (multitasking) and arrange more than one flows at the same time (multiplexing). Task scheduling uses certain type of strategy to map tasks to resources, to make sure that the resource allocation meets the expectations of users. The resources can be either machines or the processors. To achieve that, task scheduling algorithms are designed to enhance Throughput, and in the meantime decrease Completion Time. Scheduling is not done based on a single paradigm, but it considers various factors such as load balancing, quality of service, etc. Researchers have designed numerous task scheduling algorithms including Round Robin, Min-Min, Max-Min and Sufferage [1].

Different task scheduling algorithms work in their unique ways. For example, First Come First Serve scheduling algorithm makes sure that the task that arrives first in the queue is assigned first, which is simple and fast. In Round-Robin scheduling algorithm, each process is granted the same amount of time, so it has extensive overhead and poor response time. In Min-Min scheduling algorithm, we always select the task that has the potential to be completed at the earliest time and allocate resources to it. In Max-Min scheduling algorithm, however, we always select the task that can potentially be the last one to be completed instead and allocate resources to it. In Most Fit task scheduling algorithm, the task that best fits in the task scheduling list will be selected first, however, it has highest failure ratio. In Priority scheduling algorithm, each task is assigned with priority, and the algorithm always selects the task with the current highest priority to execute. In case of a tie between two tasks of the same priority, the algorithm will select the task that arrives first in the queue to execute [2].

When it comes to resource allocation, we always need to consider load balancing in order to make sure that resources are neither heavily loaded nor idle. There are different kinds of loads such as CPU load, network load, and memory capacity. Load balancing tries to be fair to all the resources in terms of equal work load at any time.

Researchers have designed various types of load balancing algorithms [3, 4]. One type is the family of static algorithms which requires that the users should have prior knowledge of the system. Another type is the family of dynamic algorithms for which prior knowledge is not required [5].

2 LBMM: Load Balancing Based Max-Min Algorithm

Task scheduling is designed to utilize resource efficiently and effectively so we can minimize the scheduling overhead and balance the load for resources.

As discussed in previous section, there are various task scheduling algorithms. Among those algorithms, Min-Min, Sufferage, and Max-Min are closely related. In Min-Min, we assume there is a list of virtual machines available, and each virtual machine currently has some tasks assigned to it, which means that in order for a task to

be executed on a virtual machine, this task needs to wait till all the current assigned tasks on this virtual machine to be executed first. The time we need to execute a certain task on a certain virtual machine depends on the computing power of this virtual machine and the size of this task. Thus the completion time (at what time the task will be completed) of a task on a virtual machine is the sum of two values: (1) how much time it takes for this virtual machine to perform this task and (2) at what time this virtual machine will finish executing all the existing tasks and become ready to execute new tasks. Each time we need to decide which task to run on which virtual machine, we calculate the completion times for all the possible pairs of tasks and virtual machines. At each round we select the task that can be potentially completed first and pair it up with the virtual machine corresponding to the earliest completion time.

In the algorithm Sufferage, for each task, we first calculate how much we would “suffer” if we do not pair it up with the best possible virtual machine, but give it the second best option instead, and use the term Sufferage to represent it. Even a task currently occupied a virtual machine, and there is a new task coming in trying to occupy the same virtual machine, we check whose suffrage value is higher. If the latter task has the higher suffrage value, it will move the former task out of the virtual machine, and moves itself into it. Thus, for each round, the CloudLet with the highest sufferage value will win the competition to occupy a virtual machine if there is any conflict.

Similar to Min-Min, in Max-Min, we assume there is a list of virtual machines available, and each virtual machine currently has some tasks assigned to it. We calculate the ready time for each machine, how much time it takes for each machine to perform each task, and based on these two values, we further calculate at what time each task will be completed on each machine theoretically. Different from Min-Min, instead of selecting a task which can be completed first theoretically, in Max-Min, we select a task that will be completed last theoretically. This is to favor those tasks with big sizes so they do not need to wait in the queue for a very long time before having a chance to be executed.

Max-Min algorithm is efficient in cutting down the waiting time for tasks with big sizes. However, this algorithm does not consider the work load of the resources. Hence, some resources are always busy and some are not, making the entire resource distribution uneven and inefficient. In order to overcome the disadvantages of unbalanced load arrangement, we need to design an algorithm for higher efficiency in resource sharing.

We design a new algorithm LBMM which is the extension of the traditional Max-Min algorithm. In this proposed algorithm two phases are involved.

Phase 1: For all the tasks waiting for resources in Cloud, we calculate the time they need when they are executed on virtual machines.

Since at a given time, each virtual machine might already has tasks executing on it, we need to take in consideration the time point at which a virtual machine is ready to perform a new task, when it finishes executing all the tasks assigned to it previously.

Based on these two factors, in order to find out for a given task, at what time point it can be theoretically completed on a virtual machine, we need to know (1) how much time this virtual machine needs to execute this task and (2) when it is available to perform this task.

However, the second factor varies all the time, because as time changes, a virtual machine might have new tasks assigned to it, resulting in the delay of its available time. Thus we need to constantly check the change of ready time of each virtual machine when determining the strategy to pair up a virtual machine and a task.

Phase 1 is executed in an iterative way. In each iteration, we calculate at what time point a task can be theoretically completed on a virtual machine (which will most likely be changed later). For a task t , we record at what time it will be completed theoretically, as well as the virtual machine v .

Once we acquired the information for all the tasks, we select the task that will be the last to be completed theoretically, and choose the virtual machine to host this task. This task then becomes the one that is assigned to a virtual time at this iteration. In our algorithm, in each iteration of Phase 1, only one task is assigned.

If a virtual machine receives a new task, its ready time for executing another new task is delayed because of this new assignment in the current iteration. Thus we need to record this change. The theoretically expected completion times of other tasks on this virtual machine are all based on the ready time of this virtual machine, and they all need to be recalculated since this virtual machine's ready time is changed. This process has the possibility to change the characteristics of certain tasks, because the completion of execution of these tasks is delayed.

Upon the completion of this iteration, this current task will be deleted from the task pool since it is already assigned to a virtual machine. In the next iteration, the remaining tasks will compete against each other once again, and a machine will be decided to host a certain task just like the previous iterations.

The iterations will continue till all the tasks in the original task pool are mapped to certain virtual machines.

Phase 2: The steps in Phase 1 favor tasks that need more time to be executed, so the waiting times of those tasks are greatly decreased. However, since in each iteration we always choose the task that can be potentially completed last theoretically, the load balance of virtual machines might not be achieved, which makes the resource distribution highly unbalanced and inefficient.

To achieve the work load balance of the virtual machines, we perform the following process. This is also an iterative process as phase 1.

First we find a virtual machine $m(j)$ that has the heaviest work load.

We then check the list of tasks assigned to $m(j)$, and select the one whose value of completion time is the smallest. From the process in phase 1, we can see that $t(i)$ is the first one that $m(j)$ receives.

In phase 1, when we determine which virtual machine should perform task $t(i)$, we calculated $t(i)$'s completion time on all virtual machines. Here we select $m(k)$ that gives $t(i)$ the highest value of completion time.

For all the tasks assigned to $m(j)$, since they are assigned to $m(j)$ at different iterations, naturally they have different completion times on the virtual machine $m(j)$. We choose the last task that $m(j)$ receives in phase 1, and record its completion time.

Next we compare these two values. If the maximum completion time of $t(i)$ on $m(k)$ is less than the completion time of the last task that is assigned to the heavy loaded machine $m(j)$, we remove $t(i)$ from $m(j)$, and reassign it to $m(k)$ instead. This operation will not delay the completion of the entire tasks on $m(j)$ or $m(k)$.

Because of this change, we need to recalculate the available time for both $m(j)$ and $m(k)$.

In the next iteration we find the task related to the virtual machine $m(j)$ with the earliest completion time again, and check if it can be reassigned to another virtual machine as we do in the previous iteration.

We perform this process till no more rescheduling is needed for those tasks originally assigned to $m(j)$.

Figure 1 shows the details of our algorithm. In our algorithm the load is balanced and the tasks are equally distributed between the resources thus no resources will be idle.

Algorithm LBMM:

Phase 1:

For all the tasks

For all the virtual machines

calculate at what time each task will be potentially completed on each machine

For each task

find at which machine this task will be completed at an earliest time t

Select the task with the highest value of t

Dispatch this task to its corresponding virtual machine that give t value

Move this task out of the list

Update the ready time for this virtual machine

Recalculate the time information of tasks on this virtual machine

Phase 2:

Find the heavy load machine $m(j)$

Do until no more scheduling is required for the heavy load machine $m(j)$

For all the tasks currently on $m(j)$

find task $t(i)$ that has earliest completion time on $m(j)$

Find machine $m(k)$ that gives task $t(i)$ the maximum completion time

If the maximum completion time of $t(i)$ (which is on $m(k)$) < maximum completion time of all the tasks that are assigned to the heavy load machine $m(j)$

Assign the task(i) to machine $m(k)$

Update ready time for both machine $m(j)$ and $m(k)$

End If

End Do

Fig. 1. LBMM algorithm for task scheduling

3 Experiment

We implement the LBMM algorithm in CloudSim Package, and the programming language is Java. CloudSim simulates the activities conducted in Cloud systems, and it has core components including data center, host, broker, virtual machine, and CloudLet [6].

There are many benefits of using CloudSim as the tool for our research, including:

- (1) CloudSim is free of charge and its source code is open to the public, thus researchers can easily work on projects related to CloudSim without extra costs.
 - (2) Since CloudSim is a simulation package, the learning phase is much shorter than those based on the real cloud systems. Thus researchers can quickly learn how to run the package and make changes to the package for their research projects.
 - (3) CloudSim provides numerous facilities and packages so researchers can analyze different facets of cloud systems by simulating the activities in CloudSim.
 - (4) CloudSim also provides a lot of examples in fields such as network topology, MapReduce programming model, task scheduling, etc., so researchers can utilize the examples provided by CloudSim and its extension.

A sample code is shown in Fig. 2. We program in Eclipse, one of the well-known IDE high language software development.

In our experiment, we mainly focus on the throughput and turnaround time. The table in Fig. 3 shows the results of the 3 Algorithms. The performance of Load Balancing using Max-Min algorithm has the best performance in terms of Throughput and Turn Around time.

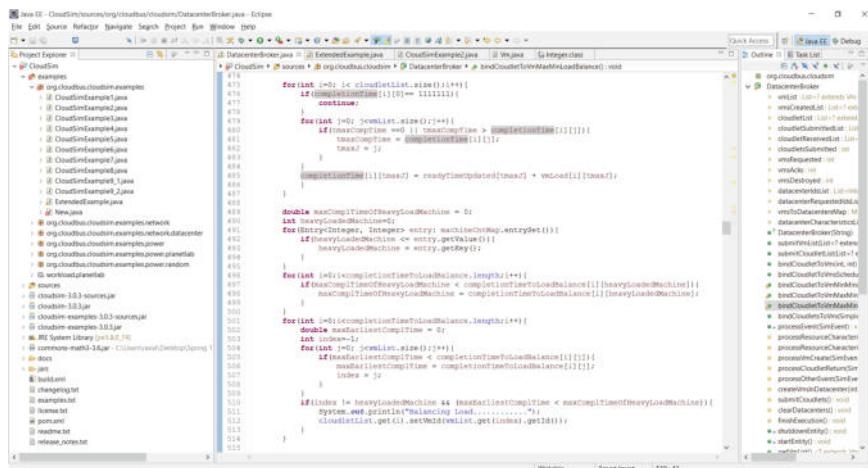


Fig. 2. Sample code of LBMM

	Throughput	Turnaround Time
Min-Min	0.08	290.18
Max-Min	0.0937	259.43
Load Balancing	0.0952	235.09

Fig. 3. Experimental results of the three algorithms

4 Conclusions

To achieve high performance in cloud computing, a new algorithm is proposed with better Throughput and Turnaround time. A package called CloudSim is used to simulate the cloud activities and conduct our experiments. From the experimental results, we can see that our algorithm decreases the completion time and improves load balancing of resources.

References

1. Singh, R.M., Paul, S., Kumar, A.: Task scheduling in cloud computing. *Int. J. Comput. Sci. Inf. Technol.* **5**(6) (2014)
2. Salot, P.: A Survey of Various Scheduling Algorithm in Cloud Computing Environment. M.E, Computer Engineering, India
3. Ghomi, E.J., Rahmani, A.M., Qader, N.N.: Load-balancing algorithms in cloud computing: a survey. *J. Netw. Comput. Appl.* **88**, 50–71 (2017)
4. Xu, Q., Arumugam, R.V., Yong, K.L., Wen, Y., Ong, Y.S., Xi, W.: Adaptive and scalable load balancing for metadata server cluster in cloud-scale file systems. *Front. Comput. Sci.* **9** (6), 904–918 (2015)
5. Padhy, R.P., Goutam, P., Rao, P.: Load Balancing in Cloud Computing Systems. National Institute of Technology, Rourkela (2011)
6. <http://www.cloudbus.org/intro.html>



Exploiting Telnet Security Flaws in the Internet of Things

James Klein and Kristen R. Walcott^(✉)

University of Colorado, Colorado Springs, USA
jklein@uccs.edu, kwalcott@uccs.edu

Abstract. The Internet of Things (IoT) is a developing technology which allows any type of network device inside a home to be linked together. IoT is based on older Wireless Sensor Network (WSN) technology and has been reduced to smaller size and scale for home use. However, both the original WSN and developing IoT technology has inherent security flaws. This paper identifies and evaluates security issues and their underlying causes in IoT technology. We focus on IoT reliance on known exploitable network ports and the difficulty of recovering from such attacks. Most IoT implementations utilize Telnet to communicate between devices. We explore the vulnerability of Telnet connections through a simulated IoT environment. Our results show that Telnet vulnerabilities can be exploited by attackers and grant access over IoT devices allowing the modification of devices and subtle spying on any data being transmitted.

Keywords: Internet of Things (IoT) · Telnet Vulnerabilities · Security

1 Introduction

We live in a world with a constantly evolving battle between new technologies and threats to those technologies. To stay ahead in the evolving technological world, companies often base new systems and technologies on older ones. This practice makes it easier for new technologies to advance more quickly, and to drive the market and increase profit margins. Unfortunately, this haste also leaves in known security vulnerabilities at the fundamental levels in new technologies that never are addressed and which are supplied to the public. The problem is a systemic one as IoT continues to connect various devices together.

As IoT technology has advanced, viruses and attacks against this sort of technology have become more common to the point where they are a constant threat. There are daily attacks against IoT environments because they are easy to gain control of due to lack of security and are high value targets [1]. Part of this is due to modern IoT technology evolving from WSNs and inheriting some of its flaws and vulnerabilities. An example of this is that some IoT environments utilize network port 23 (Telnet) to communicate [2]. This practice originated in the WSN technology and has continued into the modern IoT.

Successful Telnet attacks against IoT networks have jumped since 2014 [3]. Due to exploits like Telnet, every modern device that can connect and utilize this technology has become a risk. In the past, we would not have considered our toasters capable launching Distributed Denial of Service (DDoS) attacks, but today it is possible. Our drive to upgrade everything in our environment with technology means that we do so without considering the security risks and leave ourselves vulnerable. Refrigerators, televisions, and garage doors are just some of the more common devices which now include wireless technology that can be exploited [4].

Every environment with an IoT device is at risk and can be exploited with the right conditions. This has both local and global consequences. Locally, a compromised IoT environment can provide all sorts of personal information such as credit card numbers, usernames and passwords, bank account information, and even give physical access to homes through IoT locks. On a larger scale, a compromised system can be used to compromise more systems, and multiple systems can be used to launch DDoS attacks. These types of security issues are a major risk and if not corrected IoT technology could destroy itself in its infancy because security problems were not exposed and addressed [5].

To better understand how vulnerable IoT is given exploits in IoT, we need to examine IoT and relate exploits at a conceptual and practical level. At the conceptual level, we look at the WSN roots for vulnerabilities that can be addressed. For a practical analysis, a simulated test environment will be constructed to see the extent of IoT's vulnerabilities and to consider ways to make it more secure. One practical vulnerability we focus on is that IoT is reliant on wireless networking, routers, and protocols, such as Telnet. While these devices come with some basic security features, they are not always enabled by default or configured correctly. Moreover, if the security is overcome, then everything connected becomes vulnerable as there are no redundant security measures.

In this research, we look at the history of the IoT technology so we can understand where the inherited security issues originated, examine how IoT can be used both in industry and in the home so we can consider what kind of security is needed, and finally analyze some potential security vulnerabilities to investigate possible mitigation [6]. Utilizing the above, we created a simulated IoT environment with the inherited Telnet security risk in IoT and allowed outside attackers to attempt to exploit the environment. Our results show that having vulnerable Telnet systems in an IoT environment allows potential attackers access to the IoT environment, and any other wireless devices connected to it to a worrying degree.

In this paper, we make the following contributions:

- Analyze potential security vulnerabilities in IoT systems (Sect. 2).
- Experimentally demonstrate Telnet security vulnerabilities in IoT (Sect. 3).
- Discuss the observed vulnerabilities and causes in a simulated IoT system (Sect. 3.5).

2 IoT Technology

IoT attempts to connect all separate wireless devices as well as allow for remote connection and control of them. Most IoT devices use WSN technology as a backbone. In this section, we discuss the role of WSNs in IoT and some vulnerabilities that are then inherited.

2.1 The Role of WSNs in IoT

WSNs were initially limited in use, so there were not many opportunities for hacking or security exploitation. The first Wireless Sensor Networks (DSN) were developed as Distributed Sensor Networks (DSN) at the Defense Advanced Research Projects Agency (DARPA) in 1978 [7]. Later, WSNs were created to utilize wireless and modern networking technology to be easily able to communicate, and the sensors themselves are small and powerful, yet also cheaper and more disposable to make the technology viable.

IoT utilizes the same basic technology as a WSN, but, instead of being used for wide scale corporate technologies, it is scaled more specifically to the home environment. IoT uses home routers and wireless receivers to correlate data useful specifically in a home environment and helps run a home more efficiently and effectively by automating some of the functions normally done manually. IoT took the basic WSN technology to the next step, but instead of reengineering it from the ground up and including new modern security in it, quite a bit has been taken and reused leaving potential security issues waiting to be exploited.

Securing of private information becomes more important when coupled with IoT [8]. When an IoT network is running in a home, it naturally stores more personal information as it is automating personal living preferences and attempting to make normal mundane tasks simpler. Thus, much personal data and preferences are stored and sent over the IoT. If it is not well protected, it could easily be intercepted and stolen. That in turn could lead to an increase in personal security risks, stolen identities, and public data that should be private.

2.2 IoT Security

IoT and the underlying WSN technology have a number of inherent security problems and risks. For the WSN issues, these underlying problems came with the original technology and have carried through to the current IoT implementations. These concerns need to be taken into consideration so they do not continue to undermine the security of the entire technology. The evolution of the technology from WSNs into IoT has also created some new concerns which need to be solved and implemented to ensure that the increased amount of private data contained by the IoT does not cause security vulnerabilities to its users.

WSN Security Concerns Most original WSN wireless sensors do not naturally encrypt the data it will be transmitting. Home wireless setups do not

encrypt the data passed over it wirelessly by default, thus this issue has continued forward into IoT implementations. When you transmit your data, you do not want everyone to be able to grab and use it, so the device must have some form of encryption to protect the data before it is transmitted. Of course data security is one of the largest concerns with use of the technology, especially in the medical field. If the data is tampered with or stolen from a Man-in-the-Middle attack it can cause all sort of problems for both the doctor and patient [9].

Most WSNs can be setup to use Symmetric Key Cryptography to encrypt their messages, but devices using the routing potential of the MANET can also use Public Key Cryptography [10]. Wireless sensors do not need to have the ability to decrypt encrypted messages unless specified. They only need the ability to pass encrypted messages transmitted through them.

Another possible method of keeping the data secure would be to create a network tunnel between the wireless sensor and the collector. Though the tunnel physically goes through many devices, it logically could connect the two together keeping any data transferred between the two secret. The drawback of tunneling is it would require extra power usage from the wireless sensor which would really only be feasible if the wireless sensors were directly hooked up to power and not running on batteries, such as in an IoT implementation.

Security concerns vary depending on the specific usage of the WSN and run the whole range of a minor concern (such as agriculture implementations) to extremely high (such as when used for health concerns). One of the largest security vulnerabilities IoT has is Denial of Service (DoS) attacks [11]. These attacks are effective against IoT technology because of the wireless sensors cannot send their data to the collector, then the whole IoT becomes useless. While for some IoT functions DoS attacks may not matter, it is very important when the data gathered is used in real time and the wireless sensors not reporting could endanger lives, such as in healthcare BAN uses or natural disaster alerts.

Another major threat is physically tampering with a wireless sensor. It may send false data or hacked into for the same effect [12]. A node sending false data could be worse than a node not sending data because it could leave those analyzing the data to believe everything is working when in fact it is not. These two security threats provide the biggest security downfalls to WSN technology.

We also do not currently have reliable solutions to DoS and physical tampering. Some wireless sensors have the ability to sense if they are being tampered with, but that capability utilizes the sensors other resources and makes it less efficient. One can also increase the transceiver size or increase the amount of wireless sensors, neither of which is generally possible due to WSNs lack of resources. Overall these WSN problems also plague IoT systems and will need to be solved in the future for IoT technology to become more widely utilized.

IoT Security Concerns The IoT has taken the underlying WSN technology and advanced it, but with additional functionality comes additional security threats and concerns. An example would be the IoT ability to trigger specific actions based on proximity or GPS location [13]. For example, the IoT can track

a users GPS location through their phone and trigger events, such as lights coming on or going off when the phone and user enters or leaves a room.

Another concern is considering what data the IoT stores [14]. If there are multiple users in a single area with sensors, the system has to be able to tell one from another to properly identify and use their preferences. It is convenient for a user to store a lot of metadata in their profile for the IoT system to use. While things like personal preferences seem obvious, additional data like credit card numbers is also common, to easily allow purchases online through IoT televisions or video game systems. The IoT allows for a number of ways for credit card data to be compromised, such as storing it or transmitting it without encryption. Having stolen credit card data linked to a name or even a social security number are a huge risk, but these types of data can be commonly stored in the IoT with little to no security. Privacy and keeping private data secure is one of the most important security issues with IoT.

Yet another common IoT security threat is giving automated command of appliances to the IoT controller. This use is convenient since the controller could check to make sure the oven is off once people are in bed or to ensure the shower is the perfect temperature. The biggest threat here however is that an outsider will be able to hack into an IoT network or just log into an unsecured one and take over command of these appliances. The examples of this threat are numerous, such as turning on an oven in the middle of the night and causing a fire; causing flooding from a shower, dishwasher, or washing machine; or a refrigerators temperature being turned up and food spoiling. The threats from this type of attack are numerous and range from annoying to life threatening.

A lot of modern houses have home security systems which include door codes, intruder alarms, motion detectors, and possibly cameras. All of these can be considered wireless sensors and linked to the IoT for user control [15]. However, if an outsider is also able to access and control them they gain almost complete control over their targets home. Worse, it is generally in a way the user will not know allowing the attacker to spy through cameras, unlock doors for themselves, and shut down alarms that might go off.

So far the focus has been on using IoT in a home environment, however IoT is also able to connect to the Cloud to stretch its reach even further allowing its user to access and control anything attached to it from anywhere their phone has reception [16]. Utilizing IoT in this manner introduces many additional security issues due to the vulnerability of the Cloud and allows those looking for access a lot more ways to cause problems. Cloud usage also expands the list of IoT vulnerable devices to other computer controlled devices outside the home, such as cars. The additional risks gained by connection the IoT framework to the cloud increase the security risks exponentially.

Medical uses of the IoT bring additional risks as affecting someones medical conditions can be deadly. While being able to remotely access and analyze the data from a pacemaker or insulin pump is extremely useful, lack of security on these devices also make it extremely dangerous [17]. Also, accessing personal medical records through poorly secured or unsecured IoT connections creates

a big problem. Lack of IoT security and correct setup make all of these things possible and shows the dangers of the lack of IoT security even more clearly.

Finally, most modern IoT implementations have a Telnet vulnerability [18]. In some implementations Telnet is used because the sensors have a minimal amount of memory to keep costs down and Telnet does not encrypt so it keeps the overhead low. In other implementations SSH or other encrypted traffic is used, yet still leave the Telnet vulnerability in their systems for unknown reasons. Regardless of why this vulnerability is prevalent in most implementations, it is becoming a massive threat in the IoT world.

These threats represent only some of the possible problems substandard security can cause with the IoT. The threats caused both by the underlying WSN technology and the newer IoT are both pressing and valid. Proper setup and installation of the IoT technology is possible, but the variety and severity of the consequences for setting it up incorrectly seem to indicate an expert should be the one doing so and that is rarely the case. The default settings on IoT technology and the innate security risks need to be addressed by the manufacturers to help solve some of these problems. The alternative is that an IoT security program similar to anti-virus software needs to be developed and offered which can monitor and detect these possible issues in real time. One issue specifically with the IoT is that a lot of the devices have network capabilities added to them haphazardly without proper testing or review and no security is added to them.

3 Testing of IoT Security and Vulnerabilities

The biggest potential problem facing upcoming and emerging technologies is the security issues that threaten to shake the public's confidence in a product, removing any foundation it may have had before it has the chance to shine. The emerging IoT technology could have positive widespread implications for society, but if security is not a priority in its development then there will be risks. When this knowledge becomes public, any opportunity for the technology to take root and grow will be threatened.

3.1 Experiment Design

A primary IoT concern is that IoTs WSN roots have left potential problems and issues in the underlying product that could cause security issues. One example is that some modern IoT implementations have chosen to use the Telnet network port (port 23) to communicate between the sensors and controller over the Wi-Fi or home wireless network like it was used in WSN set ups.

The Telnet vulnerability means that an incorrect Telnet setup on a wireless network running IoT could leave the whole system vulnerable to intruders or other security issues. Also, if another exploit can be used on a wireless network allowing access to the port, everything connected to the IoT in the home would be compromised. To address this challenge, we will be simulating an IoT network

leaving this port open to attack and analyzing the attacks against it to attempt to compile metrics showing this known vulnerability.

The challenges of running and testing IoT technology and trying to exploit possible security faults are numerous. First since the technology is still in its infancy setting up the technology and doing actual testing is a challenge. The technology can be a challenge to find and expensive. All of the current implementations are custom work, there are not any beginners kits you can buy and quickly install to test on. Finally, lack of comprehensive domain knowledge or best practices makes setup and troubleshooting of an IoT system a challenge. While there are plenty of articles on how the technology should theoretically work or conceptual ideas of how to implement it, there currently are not any widely available practical guides or solutions for how to make it work. The lack of practical information means that not only the basic setup is a challenge, but getting it to run in a stable enough way to try to test it for security problems becomes problematic.

To address these challenges in the experiment a laptop will be configured and set up a simulated IoT environment. Realgames has a Home I/O Smart Home Simulation, which will be used. The simulated environment will give a basic, but realistic simulation of an IoT home setup which will show the types of things accessible to both a legitimate user and to those infiltrating the network.

3.2 Security Test Environment Setup

The IoT simulation was set up on a home network and the laptop configured on the network like a honeypot where any intruders would be routed towards the IoT setup. The honeypot setup is important to point possible attackers towards the IoT setup as opposed to anything else on the network and so its security settings can be easily changed as necessary for the experiment.

As shown in Zone A in Fig. 1, the wireless network router was configured to take all the Telnet port 23 traffic and forward it to the experiment laptop in Zone B. This includes forwarding traffic through the routers and laptops firewalls, which were configured with rules to allow that traffic access.

The laptop in Zone B was a Dell Inspiron 15 running an AMD A6 Processor with 4GB of RAM and a 64-bit instance of Windows 10. It had the most recent Windows 10 patches and updates. The test was utilizing a system that is as modern and updated as possible and has as few known vulnerabilities or issues as possible so that the likelihood of those types of issues affecting the results is less likely. Additionally, a user account with normal privileges was be running and accessible the majority of the time to the attackers, while an administrator account was be running in parallel to track the activity on the laptop and the network. A few protected files were created and put into the My Documents folder of the user account to help determine the level of complexity of the attacks.

In addition, the experiment framework is running the AVG Zen Free antivirus and firewall, which is configured with a password to keep it from being turned off. However, there is a setting adjustable by any user. Wireshark is installed and configured on the administrator account to capture all the network traffic passing

over the network and record the data for evaluation. Also HoneyBOT, which is a honeypot software, tracks what ports were being utilized. We additionally created a program that tracked what files were changed and when the system was running on the administrator account so that what files were accessed would be clear. Finally, a telnet server instance was set up on the laptop to allow those trying to access it to have a connection.

Zone C in Fig. 1 shows the simulated IoT setup on the laptop. In the Real-games IoT simulation there are a number of devices. By default lights, doors, alarms, security system, and cameras are all usable and configurable. In addition there is a companion program which allows for programming of additional devices. We configured a television, refrigerator, and home thermostat to add additional devices to the simulation. All the devices in the simulation are Wi-Fi devices that pass data over the wireless network which could be captured, changed, or stolen by an attacker.



Fig. 1. Experiment network

3.3 Experiment Results and Evaluation

In the first phase of the experiment, we configured the environment with the IoT simulation fully running and confirmed that the network traffic was utilizing the honeypot correctly to attract anyone exploiting the network. We accessed the test environment from outside the network. This setup allowed the laptop to act as a vulnerable IoT device where we assume the security of the network has already been breached and that the vulnerable telnet port can be utilized to access other IoT and network devices.

The experiment ran for five days. Due to destructive attacks upon the IoT setup against the equipment, the environment was no longer able to continue the experiment. Once we opened up the security, we expected to need to broadcast out the vulnerable IP address, but within hours, both of the telnet servers sockets were being utilized as shown in Fig. 2.



Fig. 2. Telnet socket utilization

During the five day period when the experiment was running, there were six documented attacks. Three were subtle attacks consisting of viewing the exploited machine for data, attempting to install subtle monitoring programs on it, or attempting to track when a user was or was not using the machine. The other three attacks were more direct and consisted of directly attempting to install viruses, attacking hardware wirelessly attached to the network through Wi-Fi and destroying the wireless router the experiment was being conducted through, as shown in Zone A in Fig. 1.

3.4 Results

We tracked a number of different things in the experiment. One important metric we captured was how frequently the Telnet server sockets were utilized through the HoneyBOT software that was running and tracking port connections. Figure 2 shows that the Telnet connections were nearly at 100% utilization throughout the whole experiment. The dips back to zero in Fig. 2 reflect that each day we would reset the Telnet servers connections to allow new connections to utilize the sockets. Within hours of that reset, they would be reutilized.

Out of the ten possible connections over the five days, five of the connections didn't directly make any changes, or interfere with the system. We equate these with the three subtle attacks as they were either only watching, or making indirect or minor changes to the system. The only sign of these attacks was a hidden program which tracked when a user was on the machine was installed and created a text file on the vulnerable machine showing when the machine was being used. The creation of the text file was noticed by our file tracking software along with the port and traffic data we were capturing for the experiment. This enabled us to readily see the changes, but without those steps which wouldn't be a normal part of IoT security, these attacks would have been invisible.

The other three attacks were much more direct. Six attempts were made to install a keylogger on the vulnerable box during one attack. The installed antivirus stopped these attempts, which lead to them attempting to modify the

antivirus or uninstall it. In the other two direct attacks the attackers were able to access other wireless devices connected to the experiments environment. The printer was accessed wirelessly through the vulnerable machines Wi-Fi connection and it was power cycling constantly until it was rebooted and removed from the wireless network. In the final direct attack the attackers were able to access the wireless routers administration settings (they had been set to the defaults for this experiment) and make it completely inoperable.

We also tracked the changes to the dummy documents that had been placed to see how often they were accessed or changed using the file change tracking program installed. The files were all accessed each day shortly after resetting the Telnet sockets to allow new connections. None of the files were altered. The files with varying protection levels were still only accessed once, not multiple times, as was expected if hackers were attempting to brute force passwords.

3.5 Evaluation and Discussion

We logged the connections and the IP addresses of those connecting and attempted to trace them back through traditional means. Tracing the IP addresses lead back to Akamai Technologies in Massachusetts, which is a cloud provider. The probable scenario is that the attacks were launched from elsewhere and utilized that cloud access point as their last stop before attacking intentionally to mask the origin of the attacks. Without additional tools, it becomes nearly impossible to fully trace the origin of the attacks and while the data would prove interesting, it is not critical to the experiments results.

Next we analyzed what changes were being made to the environment by the malicious attacks. We examined the data captured during the attacks from the honeypot, Wireshark, and the file change capture utility in addition to the Windows logs. Looking at the access record of the dummy files only being accessed once each day shows that the most likely scenario is that the automated bot accessed them when it connected, then when it couldnt break their passwords they were downloaded off the vulnerable machine for later analysis.

In addition through this data we determined that there were connections from six different IP addresses with one that reconnected each time the connections were reset and the others changing each time. The one that continued to connect daily was not responsible for any of the direct attacks. We surmise that those connections were utilizing the vulnerability to wait and watch for vulnerable data they could use or exploit. An example of this would be entering credit card information, which could then be taken and exploited without revealing their presence on the exploited machine. These sort of attacks are difficult to track due to their subtle nature.

The number of connections each day and the speed with which they connected to the vulnerable network leads us to believe that the connections were being made by bots which were scanning IP addresses and attempting to connect to vulnerable Telnet systems. This indicates that more attacks would have been likely had the experiments environment allowed it. We also noted that once an IP had connected, they remained connected until we reset the server. This

implies that had the server not been reset they would have continued monitoring and attempting to exploit the vulnerable system for as long as possible.

When analyzing the antivirus after the experiments conclusion, we found that some changes had been attempted to its settings. In anticipation of this a password had been set on some areas of the antivirus and while the attackers were able to make some modifications to it (such as setting its scan schedule to once a year), they were unable to disable it completely and leave the system vulnerable. This shows that after the initial connection was established by the botnet, control was most likely transferred to a human to launch additional and more complex attacks.

The direct attacks would be particularly effective against IoT devices and networks and that type of environment would have been vulnerable to the keylogger or installation of any other sort of malicious program. In addition the direct attacks did not only exploit the vulnerable box or the text environment, but also other devices without security connected through the Wi-Fi. These attacks emphasize how once an attacker has exploited a vulnerable IoT device it can utilize all its connections to other unprotected devices, gaining more influence over everything connected and attacking them in turn.

Our assessment of the complexity of the attacks derives from the assumption that botnets were being used for the initial discovery of our vulnerable network, and the evidence we found of attacks against the system. The more subtle attacks are of moderate complexity due to the understanding that waiting and listening will gain you more potentially useful information, and the subtle user tracking done. The more direct attacks had a much lower level of complexity, seeming only interested in causing as much direct damage to the environment as possible.

Overall, from our study we learned that there are still attackers actively looking to exploit vulnerable Telnet machines. There are bots scanning IP space for these vulnerable connections and ready to attempt to exploit them. There are enough different attackers doing this that we saw both subtle and direct attacks of varying complexity levels. In an actual IoT environment the security on the devices would not be enough to overcome the types of attacks we logged and detecting or recovering from them could be a challenge in an IoT environment. In summary, IoT systems retaining vulnerable Telnet connections are dangerous and a lot of damage can be done through them to vulnerable systems.

4 Threats to Validity

One potential threat is that to get the Telnet service running for the simulation we had to stand up a telnet server to allow outside hosts to connect. In an actual IoT environment this would not be necessary as the devices are configured to send and receive Telnet traffic. The server was only able to handle two Telnet connections at a time, which limited the potential of the experiment. We believe we still were able to show Telnet's vulnerabilities as both connections were utilized nearly one-hundred percent of the time, however more data could have been captured if this had not been limited.

A second concern is that the environment was set up and configured in a home environment. This was important to the experiment as this technology is primarily being produced for this environment. However, that does mean that uncontrolled or unexpected factors may have been at work in the environment. One example is that the wireless printer was not considered in the environment, yet was utilized by the attackers. While acknowledging that this makes some aspects of the environment potentially uncontrolled, it also allows us a more realistic look at how these attacks may occur in their expected environment.

Finally, in our experiment we decided to focus on the Telnet vulnerabilities in IoT as they can be directly traced back to the original WSN predecessor. While we acknowledge that only some IoT implementations utilize Telnet while others use SSH or other encrypted traffic, the majority of the concepts still apply. While we explore Telnet vulnerabilities in depth, most attacks of these types attack both Telnet and SSH and exploit the fact that in either implementation default usernames and passwords are left in the devices and exploitable [19].

5 Related Work

WSNs are not a new concept and a lot of papers have been written about them [8, 20–27]. These papers often cover the basics of how a WSN works and give a lot of good data about how they can be used. However, they rarely look at security matters and those papers that do not take into account the evolution of the technology and give possible problems, exploits, or vulnerabilities currently applicable.

Most papers discussing WSNs do not cover IoT, and most IoT papers do not mention WSNs [1, 2, 5–7, 10–14, 18, 19]. However, there is an important link when considering the security of modern IoT systems since some of the vulnerabilities that were inherent in the old system and have continued to be a problem when inherited by the IoT when they could have been fixed. Understanding the history of the IoT and how it is evolved is essential to getting a full grasp on the security problems still in the system.

Most of the literature on the IoT discusses the positive points of it and how convenient it is without discussing the security aspects [3, 4, 14]. Even those that do bring up the topic of security usually focus on a single aspect of that problem, instead of addressing the whole scope. There are a number of security issues in the IoT and failing to fix or address a single one still leaves the system vulnerable to data being stolen or exploited.

Finally, privacy must be considered. While information regarding personal information and the dangers of it being stolen are becoming more prevalent, few of these specifically address storing this type of information on your personal network where the IoT could access it [6, 7, 15–17]. Some of the devices on the Internet of Things, particularly those with cameras, provide ability to put those using them and their personal data at risk. Devices that hold and store personal information that can be used to get more information or access other accounts must be better protected in this sort of environment.

In addition, many large technology companies are beginning to slowly put information out about their own IoT approaches, guidelines, and products. The additional research and best practices produced by these companies could give a lot of extra information on how the field of IoT will continue to develop and show additional security information and vulnerabilities not considered here.

6 Conclusions and Future Work

IoT has grown from the WSN technology and has been incorporated into many more modern devices, sensors and network concepts to enable a technological system which provides endless possibilities for the technology in the future. However, due to the reliance on previous technologies, there are innate security issues to be addressed in the new technology. These security flaws must be a main concern in the future of IoT technology.

This research shows that there are known security issues and holes in the current IoT technology. Identification and education is the first step in solving any problem. The data gathered in the experiment shows the damage that can be done through a known Telnet vulnerability. IoT needs to look back at the vulnerabilities it inherited from the WSN technology and fix those problems before continuing to move forward. The potential security risks that comes with that vulnerability are being demonstrated on a daily basis across the world through attacks on various IoT networks.

We learned that having a Telnet vulnerability grants an attacker a variety of ways to exploit data and wireless systems. Once the outer shell of security has been penetrated, an attacker has access to everything else inside the IoT environment including other devices connected through the Wi-Fi network. With the lack of security on IoT networks, attackers who find vulnerable systems can quickly exploit that device(s), gain more access, or install spying programs to capture personal data.

In the future, we will look into other vulnerabilities in IoT environments, especially those inherited from backbone systems such as WSNs. We will also consider automated correction mechanizms for when IoT devices automatically detect intrusion and analyze automated intrusion detection techniques.

References

1. Kolias, C., Kambourakis, G., Stavrou, A., Voas, J.: DDos in the IoT: Mirai and other botnets. *Computer* **50**(7), 80–84 (2017)
2. Pa, Y.M.P., Suzuki, S., Yoshioka, K., Matsumoto, T., Kasama, T., Rossow, C.: Iotpot: a novel honeypot for revealing current iot threats. *J. Inf. Process.* **24**(3), 522–533 (2016)
3. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. *Comput. Netw.* **54**(15), 2787–2805 (2010)
4. Misra, G., Kumar, V., Agarwal, A., Agarwal, K.: Internet of things (IoT)—a technological analysis and survey on vision, concepts, challenges, innovation directions, technologies, and applications (an upcoming or future generation computer communication system technology). *Am. J. Electr. Electron. Eng.* **4**(1), 23–32 (2016)

5. Jing, Q., Vasilakos, A.V., Wan, J., Lu, J., Qiu, D.: Security of the internet of things: perspectives and challenges. *Wirel. Netw.* **20**(8), 2481–2501 (2014)
6. Weber, R.H.: Internet of things-new security and privacy challenges. *Comput. Law Secur. Rev.* **26**(1), 23–30 (2010)
7. Roman, R., Zhou, J., Lopez, J.: On the features and challenges of security and privacy in distributed internet of things. *Comput. Netw.* **57**(10), 2266–2279 (2013)
8. López, T.S., Kim, D., Canepa, G.H., Koumadi, K.: Integrating wireless sensors and RFID tags into energy-efficient and dynamic context networks. *Comput. J.* **52**(2), 240–267 (2008)
9. Roman, R., Najera, P., Lopez, J.: Securing the internet of things. *Computer* **44**(9), 51–58 (2011)
10. Gan, G., Lu, Z., Jiang, J.: Internet of things security analysis. In: 2011 International Conference on Internet Technology and Applications (iTAP), pp. 1–4. IEEE (2011)
11. Suo, H., Wan, J., Zou, C., Liu, J.: Security in the internet of things: a review. In: 2012 International Conference on Computer Science and Electronics Engineering (ICCSEE). vol. 3, pp. 648–651. IEEE (2012)
12. Koliاس, C., Stavrou, A., Voas, J., Bojanova, I., Kuhn, R.: Learning internet-of-things security “hands-on”. *IEEE Secur. Priv.* **14**(1), 37–46 (2016)
13. Zhou, L., Chao, H.C.: Multimedia traffic security architecture for the internet of things. *IEEE Netw.* **25**(3), 35–40 (2011)
14. Tan, L., Wang, N.: Future internet: the internet of things. In: 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), vol. 5, pp. V5–376, IEEE (2010)
15. Airehrour, D., Gutierrez, J., Ray, S.K.: Secure routing for internet of things: a survey. *J. Netw. Comput. Appl.* **66**, 198–213 (2016)
16. Botta, A., De Donato, W., Persico, V., Pescapé, A.: Integration of cloud computing and internet of things: a survey. *Future Gener. Comput. Syst.* **56**, 684–700 (2016)
17. Moosavi, S.R., Gia, T.N., Nigussie, E., Rahmani, A.M., Virtanen, S., Tenhunen, H., Isoaho, J.: End-to-end security scheme for mobility enabled healthcare internet of things. *Future Gener. Comput. Syst.* **64**, 108–124 (2016)
18. Bertino, E., Islam, N.: Botnets and internet of things security. *Computer* **50**(2), 76–79 (2017)
19. Angrishi, K.: Turning internet of things (IoT) into internet of vulnerabilities (IoV): IoT botnets. arXiv preprint [arXiv:1702.03681](https://arxiv.org/abs/1702.03681) (2017)
20. Pathan, A.S.K., Lee, H.W., Hong, C.S.: Security in wireless sensor networks: issues and challenges. In: 2006 8th International Conference Advanced Communication Technology, vol. 2, 6pp.–1048, Feb 2006
21. Cook, D.J., Das, S.K.: How smart are our environments? An updated look at the state of the art. *Pervasive Mob. Comput.* **3**(2), 53–73 (2007)
22. Werner-Allen, G., Lorincz, K., Ruiz, M., Marcillo, O., Johnson, J., Lees, J., Welsh, M.: Deploying a wireless sensor network on an active volcano. *IEEE Internet Comput.* **10**(2), 18–25 (2006)
23. Nittel, S.: A survey of geosensor networks: advances in dynamic environmental monitoring. *Sensors* **9**(7), 5664–5678 (2009)
24. Khamukhin, A.A., Bertoldo, S.: Spectral analysis of forest fire noise for early detection using wireless sensor networks. In: 2016 International Siberian Conference on Control and Communications (SIBCON), pp. 1–4. IEEE (2016)
25. Lai, X., Liu, Q., Wei, X., Wang, W., Zhou, G., Han, G.: A survey of body sensor networks. *Sensors* **13**(5), 5406–5447 (2013)

26. Zulkifli, C.Z., Noor, N.M., Zamzuri, A., Ali, M., Semunab, S.N.: Utilizing active RFID on wireless sensor network platforms for production monitoring. *Jurnal Teknologi* **78**(2), 63–72 (2016)
27. Kumar, P., Lee, H.J.: Security issues in healthcare applications using wireless medical sensor networks: a survey. *Sensors* **12**(1), 55–91 (2011)



Probabilistic Full Disclosure Attack on IoT Network Authentication Protocol

Madiha Khalid^{1(✉)}, Umar Mujahid², Muhammad Najam-ul-Islam¹,
and Binh Tran²

¹ Bahria University, Islamabad, Pakistan

madihazoheb.buic@bahria.edu.pk

² Georgia Gwinnett College, Lawrenceville, USA

Abstract. The Internet of Things (IoTs) is one of the most promising technologies of 5G. The IoTs is basically a system of interconnected computing devices which are provided with unique identification number and capability of transmitting information without human intervention. Since the computing devices (sensors) in IoTs communicate with each other using wireless channel which is accessible for all types of adversaries. Therefore, mutual authentication protocols play an important role for secure communication between the computing nodes. Recently Tewari and Gupta proposed an extremely lightweight authentication protocol to ensure the security and privacy of IoT networks in a cost-effective manner. The proposed protocol uses only two bitwise logical operators; Rotation and XOR and claimed to be one of the most secure Ultra-lightweight Mutual Authentication Protocol (UMAP). In this paper we have highlighted probabilistic full disclosure attack on the said protocol and challenged their security claims. The proposed attack model is passive and success probability is close to unity.

Keywords: Internet of things · Mutual authentication · Full disclosure attack

1 Introduction

The 5th generation communication network is envisioned to provide services like remote access, cloud computing, immersive experience and ubiquitous computing. The ubiquitous computing is a concept of the interconnected embedded systems and the smart sensors which work in collaboration by exchanging data over multiple communication channels. The IoTs is one of the prominent examples of pervasive computing in which data collected by multiple smart sensors is processed to form the valuable information which is further used to improve the human experience. The applications of IoTs are health monitoring, asset management, home automation and predictive management, etc.

The IoTs is a heterogeneous network of the smart electronic devices. In the architecture of IoTs, the perception layer is responsible for collecting data from the sensors and communicating control signals to the actuators [1–3]. To facilitate the real time communication of state variables, the devices associated with perception layer are equipped with the ability to wirelessly connect with each other. The first step of inter-

node communication protocol is basic identification of the devices. This process is referred as Identity Management (IM) system in the literature [4]. The enabling technologies used for the IM system are bar codes, Radio Frequency Identification (RFID) System, and Wireless Sensor Networks (WSN) [2]. The RFID system is preferred among other identification technologies due to the features like low cost, long range, non-line of sight scanning and high speed [5, 6].

The RFID technology comprises of three main entities; the tag, the reader and the database [7]. The tag is a passive low-cost electronic chip attached to the objects that needs to be tracked or identified. The reader identifies and logs the tag present in its vicinity. The database stores data of all the tags associated with the system in order to assist the reader in an identification process.

Since the reader and the tag communicates over a wireless channel which is open to the adversaries as well. Therefore, the communicating parties authenticate themselves before tag identification to avoid cloning, traceability and denial of service attacks. Classical authentication protocols such as Needham-Schroeder protocol, Quantum authentication protocols and Kerberos are efficient but at the same time computationally expensive. The low-cost passive devices cannot implement these protocols due to the resource limitations. According to the EGP C1G2 standards, a low-cost passive tag consists of 10 K gates and maximum 4 K gates can be dedicated to the crypto based operations [8]. Considering the resource constraints of passive devices, the Ultralightweight Mutual Authentication Protocols (UMAPs) were exclusively designed to verify the identity of the communicating parties in a low-cost passive system.

In 2006, Pedro Peris perceived UMAPs as the protocols which can only use bitwise logical operators such as *AND*, *OR*, *XOR* and *modularaddition* for the calculation of *challenge/response* messages [9]. Based on this definition, numerous protocols were presented. Some of the prominent triangular UMAPs were Lightweight Mutual Authentication Protocol (LMAP) [9], Efficient Mutual Authentication Protocol (EMAP) [10] and Minimalistic Mutual Authentication Protocol (M^2AP) [11]. These protocols despite being resource efficient (gate count less than 400) were vulnerable to multiple attacks mainly due to imbalance nature of bitwise logical operators. The cryptanalysis of above mentioned UMAPs includes attacks like Denial of Service (DoS), desynchronization, traceability and full disclosure [12–14].

Later the EGP C1G2 standardized the definition of UMAPs on the basis of gate count [8]. According to the class 1 generation 2 standard, any protocol providing authentication services under 4 K gate count is classified as an ultralightweight solution. This development paved the way for the non-triangular UMAPs. In this class, primitives other than bit wise logical operations are used to enhance the confusion and diffusion properties of public messages. Common examples of non-triangular primitives are rotation, permutation and hash function etc. These operators are basic building block of various UMAPs such as Strong Authentication Strong Integrity protocol (SASI) [15], Robust Confidentiality, Integrity, and Authentication protocol (RCIA) [16], Pseudo-kasami code based Mutual Authentication Protocol (KMAP) [17] and Succinct and Lightweight Authentication Protocol for low-cost RFID system (SLAP) [18]. Multiple vulnerabilities in the design and the equations of these protocols have been highlighted through probabilistic and deterministic attacks [19–21]. The design principles for development of secure authentication protocol are continuously evolving

through the cryptanalysis of existing UMAPs. To the best of our knowledge, all the existing UMAPs have been proven vulnerable to multiple structured and un-structured cryptanalysis techniques.

Recently Tewari and Gupta proposed a non-triangular ultralightweight protocol and claimed to provide a secure solution to IoT node authentication problem in a cost-effective manner [22]. On the contrary, the cryptanalysis of Tewari and Gupta protocol shows several weaknesses in its structure and equations [23, 24]. So far, all the attack models implemented on the protocol are deterministic in nature. These previously proposed attacks are based on brute force principle which generate results at the cost of response time. In this paper we present a probabilistic cryptanalysis model of the IoT authentication protocol. The technique used for structured cryptanalysis is known as tango attack which estimates the tags identification number by exploiting weak diffusion properties of the public messages [25]. The tango attack has success probability close to one for UMAPs with an added advantage of faster response compared to deterministic attacks.

The organization of the paper is as follows: Sect. 2 consists of the working principle of the Tewari and Gupta protocol. The probabilistic full disclosure attack is presented in Sect. 3 followed by performance analysis in Sect. 4. The paper is concluded in Sect. 5.

2 Tewari and Gupta Protocol

The Tewari and Gupta protocol presents an extremely lightweight solution to the node authentication problem. The algorithm uses only two operators; bitwise *XOR* and *Rotation*($\text{Rot}(x, y)$) function for the calculation of *challenge/response* messages. The *Rotation* ($\text{Rot}(x, y)$) function is a non-triangular primitive whose output is the left rotated version of x by the weight of y .

The memory architecture of the tag and the reader implementing the Tewari and Gupta IoT authentication protocol is elaborated in Table 1. Each tag stores a static identification number (ID) and a set of dynamic variables from two previous successful authentication sessions $((IDS^{new}, IDS^{old}), (K^{new}, K^{old}))$. The reader also stores the tag's ID and a pair of latest index pseudonyms (IDS) and keys (K). The size of dynamic variables associated with the tag is denoted by L where L depicts the bit length of the tag's static ID . According to EPG C1G2 standards, the tag's ID can assume values like 32, 64 or 96 bits. The step by step working of the protocol is as follows:

Table 1. Memory Architecture of Tewari and Gupta Protocol

	Storage location: tag & reader (L bit system)				
Variable	IDS^{new}	IDS^{old}	K^{new}	K^{old}	ID
Nature	Dynamic	Dynamic	Dynamic	Dynamic	Static

- (1) The reader sends a “ping” messages to the tag that enters its communication range.
- (2) The tag then responds with a pair of index pseudonyms (IDS^{new}, IDS^{old}). The tag is identified if at least one pseudo identification number is found in the database.
- (3) The reader generates two random numbers n and m and sends a challenge message $P||Q||R$ to the tag. The public messages P, Q, R can be calculated as follows:

$$P = IDS \oplus n \oplus m \quad (1)$$

$$Q = K \oplus n \quad (2)$$

$$R = Rot(Rot(K \oplus n, IDS), K \oplus m) \quad (3)$$

The dynamic variable (IDS) in Eq. (1)–(3) denote the latest value of pseudo identification number found in both the tag and the reader.

- (4) The tag authenticates the reader by calculating a local value of R after extracting the random numbers n and m from Q and P respectively. Once the reader is verified, the challenge messages S is communicated to the reader for the tag authentication. The equation for the calculation of message S is as follows:

$$S = Rot(Rot(IDS \oplus m, K), R \oplus n) \quad (4)$$

- (5) After successful mutual authentication, the dynamic variables of both sides are updated using following equation:

$$IDS^{old} = IDS^{new}; K^{old} = K^{new} \quad (5)$$

$$IDS^{new} = Rot(Rot(IDS^{old} \oplus n, K^{old} \oplus n), IDS^{old} \oplus m) \quad (6)$$

$$K^{new} = Rot(R \oplus n, IDS^{old} \oplus m) \quad (7)$$

A block diagram presenting a pictorial representation of the protocol is given in Fig. 1.

3 Probabilistic Full Disclosure Attack

The cryptanalysis of UMAPs includes several variants such as Denial of Service (DoS), desynchronization, traceability and full disclosure attacks. In full disclosure attack, the adversary aims to retrieve the tag’s static identification number (ID). Once the tag identity is successfully retrieved, it can be used for tag cloning, traceability attacks and identity thefts.

In this section a probabilistic full disclosure attack is performed by implementing the tango cryptanalysis technique [25]. The main idea of this model is to retrieve secret values associated with the tag by exploiting weak diffusion properties of public messages. The general working principle of the tango attack can be divided in two parts i.e.

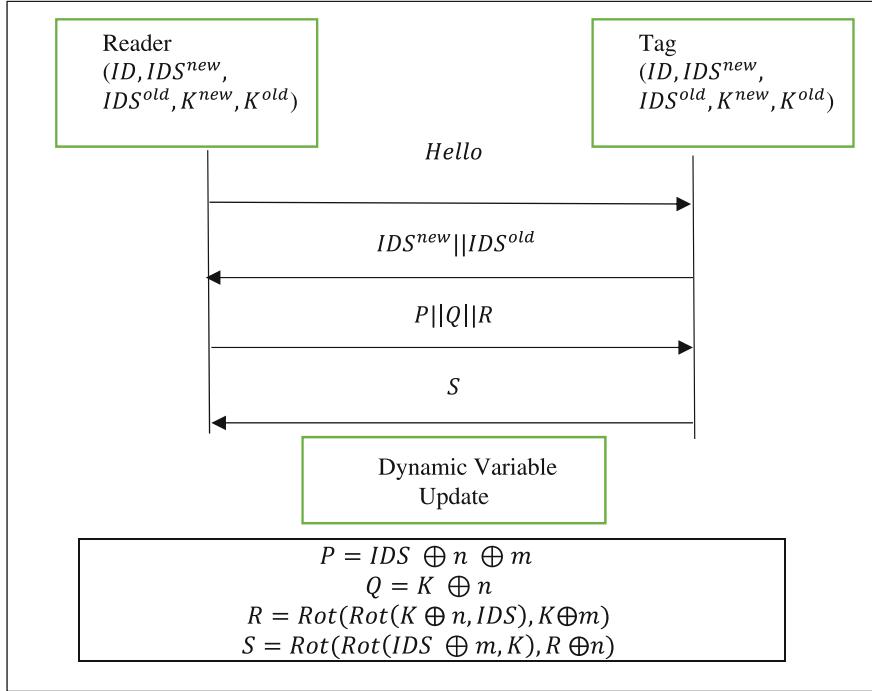


Fig. 1. Block Diagram of Tewari and Gupta Protocol

Good Approximation (GA) equations calculation and tag *ID* estimation. The details of these steps are as follows:

- (1) Good Approximation equations: In this step the linear (*XOR*) combinations of public messages exhibiting weak diffusion properties are selected. An equation can be classified as GA equation if the hamming distance between the equation results and tag's *ID* is less than half of identification number's bit length ($L/2$).
- (2) Tag *ID* estimation: After obtaining a set of x GA equations, identification number of any tag implementing the protocol under consideration can be calculated. The procedure of *ID* estimation is as follows:
 - I. Eavesdrop y number of authentication session between the tag and the reader.
 - II. Calculate good approximation equation results for each sniffed session.
 - III. As a last step, tag's *ID* is estimated in a bitwise fashion. If the sum of values present at certain bit position in GA equations result, is greater than γ then value 1 is assumed at the position under consideration otherwise 0 is placed as an estimate of tags *ID*. The expression of γ is as follows:

$$\gamma = 0.5 * x * y \quad (8)$$

By traversing in a bitwise manner, estimates for all bit locations of ID to form conjecture tag identification number can be calculated.

The Tewari and Gupta protocol comprises of only two functions i.e. XOR and $Rotation$. The concept of double rotation introduced by the authors affects the ability of the protocol to conceal secret values associated with the tag in public messages. The use of only two operations for public message calculation makes the algorithm resource efficient at the expense of making it vulnerable to multiple security threats. The tango cryptanalysis of Tewari and Gupta protocol exploits the above stated weaknesses to obtain results. Following are the step for the execution of attack:

- (1) By assuming random initial values of the tag's static and dynamic variables, the protocol is executed to obtain public messages P , Q , R and S . This process is repeated for 100 sessions. The linear combinations of these variables are considered as candidates of good approximation equations. The equation whose average hamming distance with the tag's ID is less than $L/2$ qualifies as the GA equation. Appendix A shows the average hamming distances of multiple equations for 64 bit system.
Table 2 presents 15 linear combinations of public messages whose average hamming distance with the tag's ID is less than $L/2$. These equations will be used as GA equation in step number 2.
- (2) For demonstration purposes, full disclosure attack is executed on an 8 bit system.

Table 2. Good approximation equations for tag's identification number

Target	Good approximation
ID	$GA - ID$ $= \{(P), (Q), (R), (S), (P \oplus Q), (P \oplus R), (P \oplus S), (Q \oplus R), (S \oplus R),$ $(P \oplus Q \oplus R), (P \oplus Q \oplus S), (P \oplus S \oplus R), (P \oplus Q \oplus R \oplus S),$ $(S' \oplus R'), (Q' \oplus S' \oplus R')\}$

Figure 2 represents the working example of the tango cryptanalysis. In this example two consecutive sessions are eavesdropped to obtain a pair of public messages. The results of good approximation given in Table 2 are calculated by using the snuffed data. An intermediate estimation vector $A = [A_{L-1}, A_{L-2}, A_{L-3}, \dots, A_0]$ is formed by counting number of one's in GA equations at a particular bit position and replacing the result at respective location in vector A . The conjecture ID can be obtained by using algorithm given in Fig. 3.

The results of Fig. 2 shows that by sniffing the public messages of at least two authentication sessions, the 8 bit secret identification number associated with the tag can be revealed. The above stated procedure can be used to recover the tag's ID of any length. The probability to retrieve the correct value of tag's ID depends on the number of good approximation equation (x) and the number of sessions eavesdropped (y). Since the number of good approximation equations are fixed through working in step

Concealed Variables	Values	
<i>ID</i>	0 x 5E	01011110
<i>IDS</i>	0 x 31	00110001
<i>K</i>	0 x CD	11001101
<i>n</i>	0 x 9E	10011110
<i>m</i>	0 x DE	11011110

Variables	Values
<i>x</i>	15
<i>y</i>	2
γ	$0.5 \times 15 \times 2 = 15$

<i>Public messages of Session i</i>	
<i>P</i>	0 x 71
<i>Q</i>	0 x 53
<i>R</i>	0 x 4D
<i>S</i>	0 x FB
<i>Public messages of Session i+1</i>	
<i>P</i>	0 x D7
<i>Q</i>	0 x 26
<i>R</i>	0 x 4C
<i>S</i>	0 x 95

<i>Good approximation equation (session i)</i>	
<i>P</i>	0 1 1 1 0 0 0 1
<i>Q</i>	0 1 0 1 0 0 1 1
<i>R</i>	0 1 0 0 1 1 0 1
<i>S</i>	1 1 1 1 1 0 1 1
$P \oplus Q$	0 0 1 0 0 0 1 0
$P \oplus R$	0 0 1 1 1 1 0 0
$P \oplus S$	1 0 0 0 1 0 1 0
$Q \oplus R$	0 0 0 1 1 1 1 0
$S \oplus R$	1 0 1 1 0 1 1 0
$P \oplus Q \oplus R$	0 1 1 0 1 1 1 1
$P \oplus Q \oplus S$	1 1 0 1 1 0 0 1
$P \oplus S \oplus R$	1 1 0 0 0 1 1 1
$P \oplus Q \oplus R \oplus S$	1 0 0 1 0 1 0 0
$S' \oplus R'$	1 0 1 1 0 1 1 0
$Q' \oplus R' \oplus S'$	0 0 0 1 1 0 1 0
<i>Good approximation equation (session i+1)</i>	
<i>P</i>	1 1 0 1 0 1 1 1
<i>Q</i>	0 0 1 0 0 1 1 0
<i>R</i>	0 1 0 0 1 1 0 0
<i>S</i>	1 0 0 1 0 1 0 1
$P \oplus Q$	1 1 1 1 0 0 0 1
$P \oplus R$	1 0 0 1 1 0 1 1
$P \oplus S$	0 1 0 0 0 0 1 0
$Q \oplus R$	0 1 1 0 1 0 1 0
$S \oplus R$	1 1 0 1 1 0 0 1
$P \oplus Q \oplus R$	1 0 1 1 1 1 0 1
$P \oplus Q \oplus S$	0 1 1 0 0 1 0 0
$P \oplus S \oplus R$	0 0 0 0 1 1 1 0
$P \oplus Q \oplus R \oplus S$	0 0 1 0 1 0 0 0
$S' \oplus R'$	1 1 0 1 1 0 0 1
$Q' \oplus R' \oplus S'$	0 0 0 0 0 0 0 0
<i>Column Vector A</i>	14 15 13 17 16 15 16 14
<i>Conjecture ID</i>	0 1 0 1 1 1 1 0

$$\gamma = 0.5 * 15 * 2$$

$$\text{Conjecture } ID_t = \begin{cases} 1 & A_t \geq \gamma \\ 0 & A_t < \gamma \end{cases}$$

Fig. 2. Tango Attack on 8-bit Tewari and Gupta Protocol

```

 $\gamma = 0.5 * 15 * 2$ 
for ( $i = (L - 1), i < 0, i --$ )
{
  if ( $A_i \geq \gamma$ )
     $A_i = 1$ 
  else
     $A_i = 0$ 
}

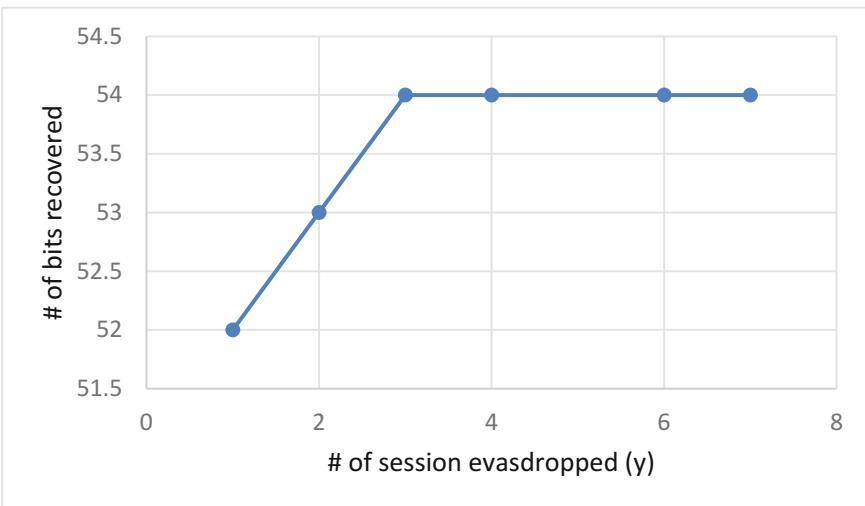
```

Fig. 3. Tag's *ID* estimation criteria

one, the results of probabilistic full disclosure attack can be improved by increasing the number of sniffed sessions between legitimate tag reader pair.

4 Performance Analysis

For tango cryptanalysis, the success probability depends on two factors i.e. the number of GA equations and the number of sessions eavesdropped for estimation. As the number of GA equations are fixed by step one of tango cryptanalysis, the probability of success to achieve tag's *ID* can be improved by sniffing larger number of public messages. Figure 4 presents the performance analysis of 64 bit system against full disclosure attack. According to the results, 84.37% of tag's *ID* can be successfully

**Fig. 4.** Performance analysis of 64-bit system against Tango Attack

retrieved by eavesdropping maximum 6 authentication session. Since average scan speed of a passive RFID system is 300 scan/sec therefore the adversary needs 0.02 s to collect information for the passive attack execution. Similar performance trend is observed for 32 and 96 bit systems.

The trend of graph in Fig. 4 proves the direct relationship between the accuracy of tag's *ID* estimate and the number of sessions eavesdropped. The static behavior of output after three sessions prevents complete disclosure of tag's *ID*. The use of genetic algorithm for GA equation calculation can facilitate the retrieval of complete value of *ID*. The genetic tango cryptanalysis will be implemented on the protocol in future to improve the success probability of full disclosure attack.

5 Conclusion

The IoTs is one of the key features of 5th generation communication systems. The IoT network comprises of low cost electronic nodes that work in collaboration to achieve user specified objectives. In this paper cryptanalysis of ultralightweight node authentication protocol is presented. The Tewari and Gupta authentication protocol is subjected to the tango cryptanalysis to obtain 15 linear combinations of public messages which exhibits weak diffusion properties. These equations are then used to disclose node's identification number. The probability of successful attack is found to be directly proportional to the number of sessions eavesdropped.

Appendix A

Good approximation equation calculation to estimate tag's *ID* (100 tests).

X	$dist(ID, X)$	GA equations
P	30.48	•
Q	31.75	•
R	31.69	•
S	30.59	•
$P \oplus Q$	31.69	•
$P \oplus R$	31.93	•
$P \oplus S$	28.89	•
$Q \oplus R$	29.40	•
$Q \oplus S$	32.28	▲
$S \oplus R$	31.42	•
$P \oplus Q \oplus R$	31.74	•
$P \oplus Q \oplus S$	31.60	•
$Q \oplus R \oplus S$	32.03	▲
$P \oplus S \oplus R$	31.8	•

(continued)

(continued)

X	$dist(ID, X)$	GA equations
$P \oplus Q \oplus R \oplus S$	31.23	•
P'	33.52	▲
Q'	32.25	▲
R'	32.31	▲
S'	33.41	▲
$Q' \oplus S'$	32.28	▲
$S' \oplus R'$	31.42	•
$Q' \oplus R' \oplus S'$	31.97	•
$P' \oplus S' \oplus R'$	32.2	▲

▲: Equation not selected as good approximation

•: Equation selected as good approximation

References

1. Atzori, L., et al.: The social internet of things (siot)–when social networks meet the internet of things: concept, architecture and network characterization. *Comput. Netw.* **56**(16), 3594–3608 (2012)
2. Lin, J., et al.: A survey on internet of things: architecture, enabling technologies, security and privacy, and applications. *IEEE Internet Things J.* **4**(5), 1125–1142 (2017)
3. Mahmoud, R., et al.: Internet of things (IoT) security: current status, challenges and prospective measures. In: 2015 10th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE (2015)
4. Babar, S., et al.: Proposed security model and threat taxonomy for the Internet of Things (IoT). In: International Conference on Network Security and Applications, Springer (2010)
5. Čika, D., Draganić, M., Šipuš, Z.: Active wireless sensor with radio frequency identification chip. In MIPRO, 2012 Proceedings of the 35th International Convention. IEEE (2012)
6. Tan, J., Koo, S.G.: A survey of technologies in internet of things. In: 2014 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), IEEE (2014)
7. Finkenzeller, K.: RFID handbook: fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication. Wiley (2010)
8. Class, E.: Generation-2 Class-1 Generation 2 UHF Air Interface Protocol Standard Version 1.2. 0. Gen. **2**: p. 2008
9. Peris-Lopez, P., et al.: LMAP: A real lightweight mutual authentication protocol for low-cost RFID tags. In: Proceedings of 2nd Workshop on RFID Security (2006)
10. Peris-Lopez, P., et al.: EMAP: an efficient mutual-authentication protocol for low-cost RFID tags. In OTM Confederated International Conferences On the Move to Meaningful Internet Systems. Springer (2006)
11. Peris-Lopez, P., et al.: M2AP: a minimalist mutual-authentication protocol for low-cost RFID tags. In: International Conference on Ubiquitous Intelligence and Computing. Springer (2006)
12. Islam, S.: Security analysis of LMAP using AVISPA. *Int. J. Secure. Netw.* **9**(1), 30–39 (2014)

13. Li, T., Deng, R.: Vulnerability analysis of EMAP-an efficient RFID mutual authentication protocol. In: The Second International Conference on Availability, Reliability and Security, 2007. ARES 2007. IEEE (2007)
14. Bárász, M., et al.: Passive attack against the M2AP mutual authentication protocol for RFID tags. In Proceedings of First International EURASIP Workshop on RFID Technology (2007)
15. Chien, H.-Y.: Sasi: a new ultralightweight rfid authentication protocol providing strong authentication and strong integrity. *IEEE Trans. Dependable Secure Comput.* **4**(4), 337–340 (2007)
16. Mujahid, U., Najam-ul-Islam, M., Shami, M.A.: Rcia: a new ultralightweight rfid authentication protocol using recursive hash. *Int. J. Distrib. Sens. Netw.* **11**(1), 642180 (2015)
17. Mujahid, U., Najam-ul-Islam, M., Sarwar, S.: A new ultralightweight RFID authentication protocol for passive low cost tags: KMAP. *Wireless Pers. Commun.* **94**(3), 725–744 (2017)
18. Luo, H., et al.: SLAP: succinct and lightweight authentication protocol for low-cost RFID system. *Wireless Netw.* **24**(1), 69–78 (2018)
19. Sun, H.-M., Ting, W.-C., Wang, K.-H.: On the security of Chien’s ultralightweight RFID authentication protocol. *IEEE Trans. Dependable Secure Comput.* **8**(2), 315–317 (2011)
20. Avoine, G., Carpent, X., Martin, B.: Strong authentication and strong integrity (SASI) is not that strong. In International Workshop on Radio Frequency Identification: Security and Privacy Issues. Springer (2010)
21. Safkhani, M., Bagheri, N.: Generalized desynchronization attack on UMAP: application to RCIA, KMAP, SLAP and SASI + protocols. *IACR Cryptology ePrint Arch.* **2016**, 905 (2016)
22. Tewari, A., Gupta, B.: Cryptanalysis of a novel ultra-lightweight mutual authentication protocol for IoT devices using RFID tags. *J. Supercomput.* **73**(3), 1085–1102 (2017)
23. Adat, V., Gupta, B.: Security in internet of things: issues, challenges, taxonomy, and architecture. *Telecommun. Syst.* **67**(3), 423–441 (2018)
24. Safkhani, M., Bagheri, N.: Passive secret disclosure attack on an ultralightweight authentication protocol for internet of things. *J. Supercomput.* **73**(8), 3579–3585 (2017)
25. Hernandez-Castro, J.C., et al.: Cryptanalysis of the David-Prasad RFID ultralightweight authentication protocol. In International Workshop on Radio Frequency Identification: Security and Privacy Issues. Springer (2010)



Multilevel Data Concealing Technique Using Steganography and Visual Cryptography

Chaitra Rangaswamaiah, Yu Bai, and Yoonsuk Choi^(✉)

California State University Fullerton, Fullerton, CA 92831, USA
chaitrarangaswamaiah@csu.fullerton.edu,
{ybai, yochoi}@fullerton.edu

Abstract. Steganography is a data hiding technique which uses images, audio or video as a cover medium. Cryptography has become an essential part of security. Image Steganography is one such way to hide secret messages in an image to reduce vulnerability to cryptanalysis. We overcome the drawbacks of using only textual steganography as it is easier to intercept and decipher. We encrypt the plaintext with a randomly generated key using XOR and One Time Pad (OTP) Algorithm and in turn embedding it into the Least Significant Bit (LSB) of the cover image. We embed the cipher text in LSB of the pixels of the cover image to form Stego image. To enhance and ensure security, we use visual cryptography along with image scrambling. Image scrambling is a technique in which the location of pixels is scrambled to provide extra protection to the Stego image. Visual cryptography is a method used to encrypt the visual information by breaking it into shares. Using both image scrambling and visual cryptography makes the system not only more secure but also difficult to decrypt. A decryption algorithm for the same is also constructed in this paper.

Keywords: Image scrambling · Mean square error (MSE) · One time pad (OTP) algorithm · Peak signal to noise ratio (PSNR) · Stego image and data security

1 Introduction

Steganography is derived from Greek words ‘steganos’ meaning protected and ‘graphein’ meaning writing. This method is used to hide data from unauthorized party which has made the technique popular as it cannot be detected easily. In recent time, steganography has improved. Vital information is being transmitted to the receiver in the presence of third party or unauthorized user without being intercepted. The most popular file formats that are being used are the digital images due to their high availability on the internet [1]. Vital Data in the form of text, image, audio or video can be encrypted and hidden into another form of text, image, audio or video.

The method of hiding data in a text file is known as textual steganography. It was very popular before the emergence of the internet. Now textual steganography has become very easy to decipher and is also not preferred as the text file cannot contain more data. Another popular method uses image as its cover medium to hide data. This method is called image steganography. Using an implanting algorithm, the data is

implanted over the image which is referred to as a Stego image and sent to the receiver. It is then processed at the receiver end using the extraction algorithm process. This method allows the intruder to know that the information is being transmitted but does not allow them to see the hidden data.

Audio steganography is another method that deals with encrypting the vital data in a cover speech which does not allow the unauthorized user to access the data. The audio steganography methods/software that are currently available can embed data in MP3 and WAV sound files.

Secret data can be hidden in any image using many steganographic techniques, there are many ways in which this can be done [10]. They must have the following requirements:

- (1) Data as plain text or cipher text or digital image or any data.
- (2) Cover medium to contain secret message
- (3) Steganographic techniques.

Additional techniques can be incorporated to maximize the level of security to increase diffusion and confusion.

The embedded secret message can be in plain text or cipher text format, any encryption algorithm can be used to generate cipher text based on type of message and medium for transmission used. In this paper we make use of XOR encryption and One Time Pad algorithm.

Image scrambling is another technique where locations of pixels are modified by scrambling to provide extra protection to the Stego image [12]. In visual cryptography a two toned secret image is hidden into a set of binary transparencies. It is an encryption technique that encrypts the modified pixel image.

2 Proposed System

Least Significant Bit (LSB) alteration procedure is used in steganography for two main reasons:

- (1) It is simple and easy to perform various experiments related to it.
- (2) Security can be increased by adding a technique that has more randomness.

About substituting LSBs, it is more preferred to add randomness to the images [4]. We should also make sure that the Stego image and the cover image seem similar. The randomness that we propose in his paper makes use of image scrambling technique and visual cryptography.

2.1 Project Description

In this paper we make use of MATLAB for simulation of the proposed algorithm. We used cover image to hide the data. We use XOR or One Time Pad (OTP) algorithm for encrypting the data. A random key is generated in the case of OTP algorithm where as in XOR encryption the key used is given by the user as an input. For XOR the key length may or may not be the same but in OTP the key length must be same.

Preprocessing procedure must be carried out for the cover image before hiding the cipher text. Now we obtain the LWT of the cover image, resulting transformed matrix consists of four sub bands namely LL, LH, HL and HH. We use the LH sub band to hide the secret data.

After this process scrambling algorithm is carried out where pixel locations are scrambled. This technique of Stego image provides extra layer of protection. Along with scrambling technique another technique called visual cryptography is applied on the Stego image. In visual cryptography method, Stego image is split into two different shares based on a threshold value and transmitted to the receiving end. At the receiver end the same threshold level must be added and reverse process of scrambling is done. The secrete message is then decrypted from this image. If the threshold value used for the generation of the shares is unknown at the receiver end it will be impossible to reveal the secrete message from the image making this highly secure and strong.

3 Methodology

In this chapter we will see how each block of the proposed system is simulated using MATLAB. We make use of modularization from the menu option in MATLAB since it makes it easier to modify the code as and when needed. In the following sections we can understand the encryption and decryption processes.

3.1 Encryption and Decryption

The encryption and the decryption processes are as shown in Figs. 1 and 2, respectively. Encryption process is carried out at the sender site where the data is converted in such a way that the unauthorized party cannot read it [9]. Plain text is referred to the original data or the secrete message that as to be encrypted. Encryption using any encryption algorithm leads to the formation of cipher text. This process of obtaining the cipher text from plain text is also called as Enciphering. Two inputs are needed for encryption of any data namely a plain text and a key which result in a cipher text. The process which gets back the plain text from cipher text is called deciphering or decryption.

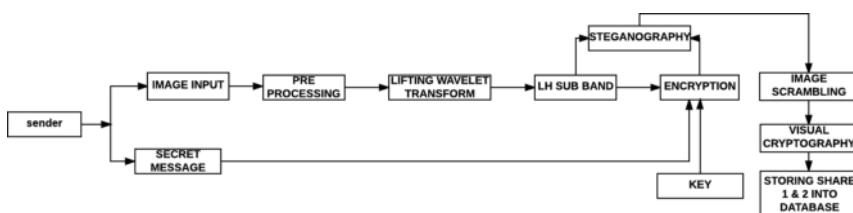
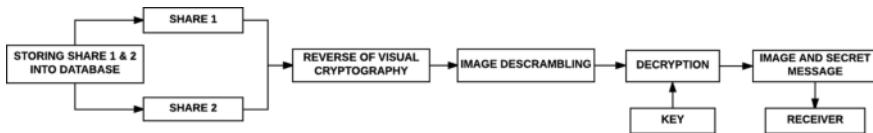


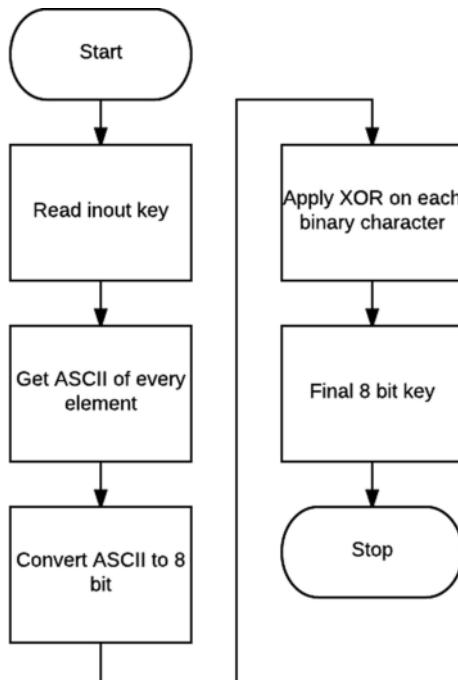
Fig. 1. Encryption process

**Fig. 2.** Decryption process

3.2 XOR Cryptography

It is one of the simplest cryptographic methods which uses the XOR operation. The data or message is encrypted using a key and XOR operation. The encryption and decryption are carried out by XORing the message and the cipher text respectively. The plain text is converted to its ASCII value and in turn converted to its binary form in 8, 16 or 32-bits form.

The key that is being generated is also converted into 8, 16 or 32 bits form respectively. The key could be a single character or a string. In the case of a string each and every character's bits are XORed to form a key. Now each character's bits of the data are logically XORed with bits of the key. The resulting value is the cipher text for the given message. In this system we make use of an 8 bit key for the encryption process. The user provides the input key which is converted into 8 bit binary key. Figure 3 shows the flow chart for key generation.

**Fig. 3.** XOR key generation

After the key is obtained as shown in Table 1, XOR encryption is performed. Figure 4 shows the flow chart for the XOR encryption process. The plain text characters are given as an input from the user, which is converted into its ASCII format and in turn into its corresponding binary format.

Table 1. XOR key generation

Key	ASCII	Binary format	Applying XOR
Hello	h—104	h—0110 1000	h—0110 1000
	e—101	e—0110 0101	he—0000 1101
	l—108	l—0110 1100	hel—0110 0001
	l—108	l—0110 1100	hell—0000 1101
	o—111	o—0110 1111	hello—0110 0010
			key—0110 0010

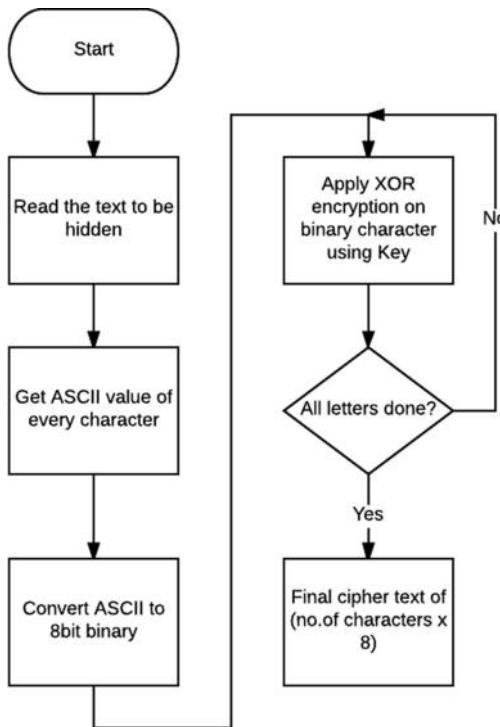


Fig. 4. XOR encryption process

Combining all the obtained values from Table 2, cvalue_1, cvalue_2, cvalue_3, cvalue_4 will give us the final cipher text. The final length of the cipher text depends on the number of characters in the plain text. Number of cipher text bits = number of characters in plain text * number of bits in key.

Table 2. XOR encryption method

Plain text	ASCII value	Binary format		
India	105—i	i—0110 1001	key—0110 0010	cvalue_1—0000 1011
	110—n	n—0110 1101	key—0110 0010	cvalue_2—0000 1111
	100—d	d—0110 0100	key—0110 0010	cvalue_3—0000 0110
	105—i	i—0110 1001	key—0110 0010	cvalue_4—0000 1011
	97—a	a—0110 0001	key—0110 0011	cvalue_5—0000 0011

3.3 One Time Pad Encryption

This encryption algorithm was developed as an improved version of the vernam cipher. OTP provides a lot of security. In this algorithm we generate a random key that has the length same as the length of the plain text. The randomness of the output and the randomness of the key makes the algorithm unbreakable. In this paper, we are making use of a 27×27 table with 26 English alphabets and 27th is a space or an underscore.

We create the 27×27 table in an excel sheet. The user gives the plain text as an input and then the length of the plain text is determined. Random key is generated in MATLAB in the decimal range 96 through 122. The generated key length is same as the plain text length. The excel sheet containing the look up table is parsed in MATLAB and the random key is generated after which the encryption process takes place. Algorithm for OTP encryption:

Input: key, Plain text Output: Cipher text

Step 1: Read the excel sheet containing 27×27 OTP table

Step 2: Search for key character in the key-column

Step 3: Search for message character in the plain text-row

Step 4: Obtain the cell value of the intersecting row and column

Step 5: Repeat from step 2 for each of the message characters.

For example: If the plain text input is India, length of this message is 5. The random key that is generated will be of the same length of 5. Let us assume that the key generated randomly is dfarj, the cipher text will be lsdzj. Next, this cipher text is further converted into its binary format with each character length equal to 8 bits.

3.4 Steganography Process

It is the process of hiding the plain text or cipher text in an image. The LSBs of the pixels of the image are used to hide the data. All images have smooth color variations (low frequency variations) and sharp color variations (high frequency variations). Both these variations together form a complete image. Separation of these two variations are

carried out in many ways, one such way that we are incorporating in our project is the discrete Wavelet transform (DWT) to obtain the frequency bands [8]. The wavelet transforms of any image gives 4 bands of each 1/4th the size of the image. The lower frequency components or the smooth variations forms the base of the image is present in the LL band. This will have the approximate image of the input image [12]. LH and the HL are the middle frequency bands which extract the horizontal features and vertical features respectively. We can either use LH or HL band for steganography, in this paper we make use of LH band. The algorithm that we use for the steganography process is:

Input: Cipher Text Output: Stego Image
Step 1: Read the cover image and resize to 128×128
Step 2: Apply LWT to obtain sub bands
Step 3: Extract Red plane of LH band
Step 4: Represent the extracted plane in binary format
Step 5: Embed the binary cipher text into LSBs of step 4
Step 6: Convert binary to decimal and convert into matrix
Step 7: Concatenate Green and Blue plane matrix with step 6's matrix
Step 8: Apply ILWT to form Stego image as in Fig. 10a, b.

3.5 Scrambling Process

Now the Stego image is formed as in Fig. 11a, which is the image where in the message or data is hidden or encrypted. To make it more secure we make use of image scrambling. Image scrambling is a technique in which the pixel location of the image is modified or rather scrambled to form a chaos image or the scrambled image [12]. This can be seen in Fig. 11b. This scrambled image can be reconstructed only if the scrambling method and variables are known [2]. We make use of the scrambling method for scrambling the image column wise and row wise [3]. In this paper, we first shuffle the column pixels and then the row pixels. The algorithm for shuffling is as follows:

Input: Stego Image Output: Shuffled Image
Step 1: Read the Stego image which is of size 128×128
Step 2: Shuffle group of first 8-pixel columns with last 8-pixel columns
Step 3: Move to next groups of pixels respectively and repeat Step 2 until all the columns are shuffled
Step 4: Shuffle group of first 8-pixel rows with last 8-pixel rows
Step 5: Move to next groups of pixels respectively and repeat Step 4 until all the rows are shuffled.

3.6 Visual Cryptography Encryption

Visual cryptography is an encryption technique where the visual information is encrypted by breaking it into shares [5]. It can be decrypted only if the person has all 'n' shares of the image. In visual cryptography, we break the image into two or more shares based on a threshold value. Depending on the threshold, much different number

of shares can be formed [5]. The threshold that we make use of can be separating white pixels from the black pixels or separating the even numbered rows from the odd numbered rows or even numbered columns from odd numbered columns [6]. In this paper we have chosen the threshold to produce shares by separating even numbered columns and odd number of columns. All columns are numbered alternatively as 1 and 2. The first share has all columns numbered 1 and the second share has all columns numbered 2. The algorithm that we use for this is as follows:

Input: Stego Image Output: 2 shares of image

Step 1: Read the scrambled image generated in Fig. 12.

Step 2: Split the image into 2 shares based on threshold of row numbers. Share 1 containing even numbered columns and Share 2 containing odd numbered rows as in Fig. 13.

Step 3: Save the 2 shares separately.

These two shares are transmitted over any wireless medium or a wired medium which are prone to less error. This transmission can happen in real time or not. If it's a real time transmission, the two shares must be sent simultaneously with the least time delay. If the message is revived late, it loses its importance. Both the sender and the receiver will have all the details of the encryption and decryption algorithms used.

3.7 Visual Cryptography Decryption

In the decryption process, the two shares which are received in Fig. 14 are combined by using the same threshold that we used for encryption [11]. The two shares are added to form the scrambled image as in Fig. 15. The algorithm for the decryption is as follows:

Input: 2 shares of image Output: Recovered Scrambled Image

Step 1: Read the 2 image shares

Step 2: Apply addition operation on both the shares column wise and save the overlapped image.

3.8 Descrambling Process

Descrambling is the process in which the pixel locations are shuffled [7]. This is the reverse process of the scrambling algorithm. The image that we have recovered after visual cryptography decryption is used and the descrambling process is applied to obtain the Stego image in which the message is hidden as depicted in Fig. 16a, b. The descrambling algorithm is as flows:

Input: Shuffled image recovered from Visual Cryptography

Output: Reformed Stego Image

Step 1: Read the recovered image from visual cryptography

Step 2: Shuffle group of first 8-pixel rows with last 8-pixel rows

Step 3: Move to next groups of pixels respectively and repeat Step 2 until all the rows are shuffled

Step 4: Shuffle group of first 8-pixel columns with last 8-pixel columns

Step 5: Move to next groups of pixels respectively and repeat Step 4 until all the columns are shuffled.

3.9 Message Extraction

This process is used to extract the hidden message from the Stego image. In this paper we have hidden the message in the LSB's of the image. Once the message is obtained, we must check if the message is in plain text format or cipher text format. If it is in the cipher text format, we must perform the required decryption process and obtain the plain text. The algorithm used is as follows:

Input: Reformed Stego Image Output: Hidden message
 Step 1: Read the reformed Stego Image
 Step 2: Apply LWT to obtain sub bands
 Step 3: Extract Red plane of LH band
 Step 4: Represent the extracted plane in binary format
 Step 5: Read out the binary cipher text embedded in LSBs
 Step 6: Save all the binary values in a variable with 8 bits representing 1 character.

3.10 XOR Decryption

The key that we used for encryption process must be known to do the decryption process. XOR operation is performed on the cipher text bits with the key bits to obtain the plain text as in Fig. 17. The detailed XOR decryption method for the key 'hello' is shown in Table 3.

Table 3. XOR decryption method

Key	ASCII	Binary format		
hello	h—104	h—0110 1000	cvalue_1—0000 1011	key—0110 0010
	e—101	e—0110 0101	cvalue_2—0000 1111	key—0110 0010
	l—108	l—0110 1100	cvalue_3—0000 0110	key—0110 0010
	l—108	l—0110 1100	cvalue_4—0000 1011	key—0110 0010
	o—111	o—0110 1111	cvalue_5—0000 0011	key—0110 0010
		Decimal format		Plain text
pvalue_1—0110 1001		105—i		India
pvalue_2—0110 1101		110—n		
pvalue_3—0110 0100		100—d		
pvalue_4—0110 1001		105—i		
pvalue_5—0110 0001		97—a		

3.11 One Time Pad Decryption

OTP is one of the most secure and unbreakable algorithms. It can only be decrypted with the help of a key that was used for encryption. The key is a randomly generated one which makes the process unbreakable. If the intruder manages to find two different keys, then two different plain texts are obtained which will make it difficult for the intruder to guess the right plain text.

The drawback of this algorithm is that random key that is generated is of the same length as the plain text. If a plain text of 100 length must be encrypted, then a key of length 100 must be generated which requires a super computer. Because of this OTP is used only when high level of security is expected. For this paper we make use of a 27×27 look up table for encryption as well as decryption. The obtained message is grouped into 8's, where 8 bits represent a character. Now they are converted back into their respective ASCII character hence forming cipher text. Now the decryption of this cipher text is carried out to get the plain text as in Fig. 18. The decryption algorithm is as follows:

Input: Key, Cipher Text Output: Plain Text

Step 1: Read the excel sheet containing 27×27 OTP table

Step 2: Search for key character in the key-column

Step 3: Search for message character in the key-row

Step 4: Obtain the cell value of the corresponding plain text-row

Step 5: Repeat from step 2 for each of the message characters.

4 Results

The simulation results are obtained from Graphics User Interface (GUI) as depicted in Figs. 5 and 8. The GUI for this is developed using MATLAB. Measurable parameter such as Mean Square Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) are also obtained experimentally.

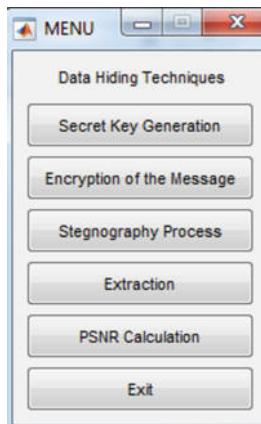


Fig. 5. GUI menu for XOR encryption process

4.1 XOR Cryptography

The result of the key generation and XOR encryption process is depicted in Figs. 6 and 7, respectively and the OTP encryption process is depicted in Fig. 9a, b.

Secret Key:

Input: hello

Output:

```
Secret key generation
Enter a input key
hello

AsciiCode =
hello

secretKey =
1 0 0 0 0 0 0
```

Fig. 6. XOR key generation

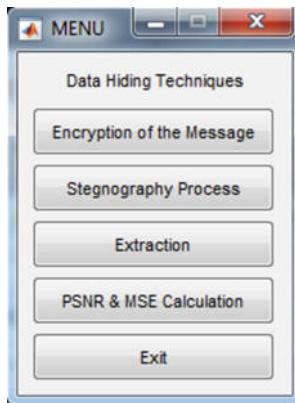
Encryption Process:**Input:** India**Output:**

```
encMsg =  
  
Columns 1 through 15  
  
0 0 0 0 1 0 1 1 0 0 0 0 1 1 0  
  
Columns 16 through 30  
  
0 0 0 0 0 0 1 1 0 0 0 0 0 1 1 0  
  
Columns 31 through 40  
  
1 1 0 0 0 0 0 0 1 1
```

Fig. 7. XOR encryption process

4.2 One Time Pad Encryption

See Figs. 8 and 9.

**Fig. 8.** GUI menu for OTP encryption process

```
Encryption of the message
Enter text to hide
india

rkey =
104    114    114    101    100
```

```
key =
hrred

The Encrypted text is

ct =
PDUMD
```

```
AsciiCode =
1x5 uint8 row vector

80    68    85    77    68
```

(a)

```
binVal =
01010000
01000100
01010101
01001101
01000100
```

```
secMsg =
Columns 1 through 9

0    1    0    1    0    0    0    0    0
```

Columns 10 through 18

```
1    0    0    0    1    0    0    0    1
```

Columns 19 through 27

```
0    1    0    1    0    1    0    1    0
```

Columns 28 through 36

```
0    1    1    0    1    0    1    0    0
```

(b)

Fig. 9. a and b depict the OTP encryption process

4.3 Steganography Process

See Fig. 10.

Input: Original Image



(a)

Output: Stego Image



(b)

Fig. 10. Steganographic encryption **a** Resized cover image, **b** stego image obtained

4.4 Scrambling Process

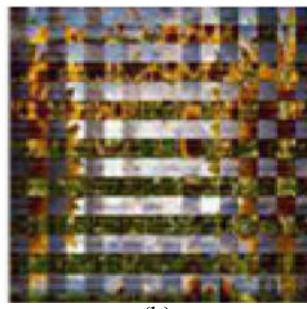
See Fig. 11.

Input: Stego image



(a)

Output: Scrambled image



(b)

Fig. 11. Scrambling process **a** Stego image obtained, **b** scrambled image

4.5 Visual Cryptography Encryption Process

See Figs. 12 and 13.

Input: Scrambled Image

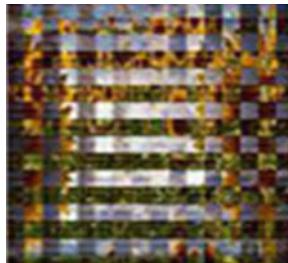
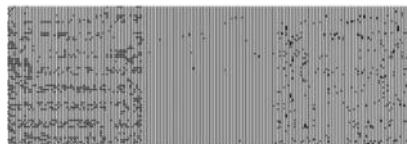


Fig. 12. Scrambled image

Output: Share 1



Share 2

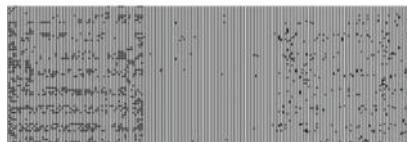
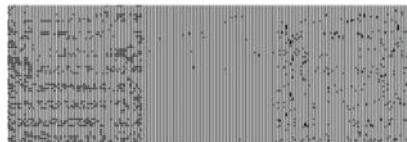


Fig. 13. Shares of an image after visual cryptography

4.6 Visual Cryptography Decryption Process

See Figs. 14 an 15.

Input: Share 1



Share 2



Fig. 14. Shares of an image after visual cryptography that has to be combined

Output: Scrambled image

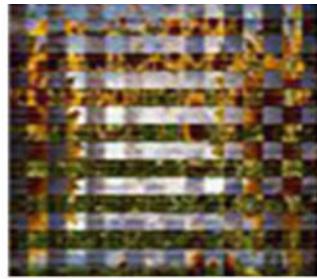


Fig. 15. Scrambled image after combining the shares

4.7 Descrambling Process

See Fig. 16.

Input: Scrambled Image



(a)

Output: Stego Image



(b)

Fig. 16. Descrambling of the image **a** Scrambled image, **b** stego image obtained

4.8 XOR Decryption Process

See Fig. 17.

Sec Msg is

ans =

india

Fig. 17. XOR decryption

4.9 One Time Pad Decryption Process

See Fig. 18.

The Decrypted text is

pt =

india

Fig. 18. OTP decryption

4.10 Parameter Measurement

Any data file that needs to be exchanged must be compressed to use less bandwidth for transmission. We make use of Mean Square Error (MSE) and Peak Signal to Noise Ratio (PSNR) to compare the original image and the compressed image and the formula used for this calculation is as below:

$$MSE = \frac{1}{MN} \sum_{x=1}^M \sum_{y=1}^N [K(x, y) - K'(x, y)]^2 \quad (1)$$

$$PSNR = 20 \log_{10} \left(\frac{MAX_{pi}}{\sqrt{MSE}} \right) \quad (2)$$

where MAX_{pi} is the minimum value any pixel can have, M, N are the dimensions of the image, $K(x, y)$ is the original image, $K'(x, y)$ is the other version of image. The MSE and PSNR are measured for cover image and the Stego image which is generated; the obtained values are indicated in Table 4.

Table 4. MSE and PSNR values

	MSE	PSNR (dB)
XOR	0.2206	54.6925
OTP	0.2330	54.4552

If the value of MSE is low, it indicates that the error is less, whereas if the value of PSNR is high, it indicates that the noise value is less than signal value.

5 Conclusion and Future Scope

The proposed method has multiple encryption and decryption processes such as XOR, OTP, steganography, image scrambling and visual Cryptography. All these processes when used individually would give protection but not as much protection as it would give when all of them are combined. It increases confusion as well as diffusion to the unauthorized party.

It can be concluded that with the use of multiple cryptographic techniques, it is made difficult for the unauthorized party to get the message. The One-time pad algorithm is one of the most secure type of encryption which makes it impossible to decrypt the code because of the random key used. Thus, the proposed system can withstand any kind of attacks. Using visual cryptography and image scrambling we have been able to enhance the regular image security.

This method can be used for implementing video steganography where in the secret message can be encrypted using video as a cover medium.

References

1. Natarajan, S., Prema, G.: Steganography using genetic algorithm along with visual cryptography for wireless network application. In: International Conference on Information Communication and Embedded Systems (ICICES) (2013)
2. Shao, L., Qin, Z., Liu, B., Qin, J., Li, H.: Image scrambling algorithm based on random shuffling strategy. In: 3rd IEEE International Conference on Industrial Electronics and Applications (ICIEA), Singapore (2008)
3. Tseng, L.-Y., Chan, Y.-K., Ho, Y.-A., Chu, Y.-P.: Image hiding with an improved genetic algorithm and an optimal pixel adjustment process. In: Eighth International Conference, vol. 3, on Intelligent Systems Design and Applications (ISDA), Kaohsiung (2008)
4. Li, H., Du, W., Yao, X., Wu, H.: A steganographic scheme based on image scrambling and coding techniques. In: International Conference on Communications, Circuits and Systems (ICCCAS) vol. 1, Chengdu, China (2013)
5. Blesswin, J., Rema, Joselin, J.: Recovering secret image in visual cryptography. In: International Conference on Communications and Signal Processing (ICCSP), Calicut (2011)
6. Jena, D., Jena, S.K.: A novel visual cryptography scheme. In: International Conference on Advance Computer Control (2009)
7. Che, S., Che, Z., Ma, B.: An improved image scrambling algorithm. In: Second International Conference on Genetic and Evolutionary Computing (WGEC), Hubei (2008)
8. Ghasemi, E., Shanbehzadeh, J., Fassih, N.: High capacity image steganography using wavelet transform and genetic algorithm. Manuscript received November, 2010; revised (2011)
9. Al-Bahadili, H.: A secure block permutation image steganography algorithm. Int. J. Crypt. Inf. Secur. (IJCIS) 3(3) (2013)
10. Usha, B.A., Srinath, N.K., Narayan, K., Sangeetha, K.N.: A secure data embedding technique in image steganography for medical images. Int. J. Adv. Res. Comput. Commun. Eng. 3(8) (2014)

11. Luo, H., Yu, F., (Correspondence author), Pan, J.-S.: Data hiding in non-expansion visual cryptography based on edge enhancement multitone. In: The Fourth International Conference on Information Assurance and Security (2008)
12. Shang, Z., Ren, H., Zhang, J.: A block location scrambling algorithm of digital image based on arnold transformation. In: The 9th International Conference for Young Computer Scientists (2008)



Ring Theoretic Key Exchange for Homomorphic Encryption

Jack Aiston^(✉)

Newcastle University, Newcastle upon Tyne, England, UK
j.aiston@newcastle.ac.uk

Abstract. We propose a key exchange protocol that works in a polynomial ideal setting. We do this so that the key can be used for a homomorphic cryptography protocol. The advantage of using key exchange over a public key system is that a large proportion of the process needs to be carried out only once instead of needing a more complicated encryption function to use for each piece of data. Polynomials rings are an appropriate choice of structure for this particular type of scheme as they allow universal computation. This paper will examine how we can perform computation correctly on cipher texts and address some of the potential weaknesses of such a process.

Keywords: Cryptography · Homomorphic · Key exchange · Rings · Ideals

1 Introduction

The number one goal of cryptography has always been to ensure readable versions of sensitive or private data do not get into the wrong hands, even if it is potentially accessible. It was suggested in the 70s however that it may be possible to do more than just safely store data; in fact, useful operations could be performed on encrypted data [1].

The importance of computation on encrypted data has become apparent in more recent years due to the adoption and ever-increasing use of the cloud. Despite the potential for faster data analysis, organizations such as hospitals and banks are hesitant to make use of the cloud due to the lack of privacy. A successful homomorphic encryption protocol would solve this problem as data would remain encrypted until taken back off the cloud for example Fig. 1 gives an overview of the process for computing the mean would work.

The problem with a lot of current methods of homomorphic encryption is that only one operation is possible (usually, but not always, addition or multiplication) [2]. More meaningful applications would become possible if both the typical operations were available.

This work was supported by the Engineering and Physical Sciences Research Council, Centre for Doctoral Training in Cloud Computing for Big Data [grant number EP/L015358/1].

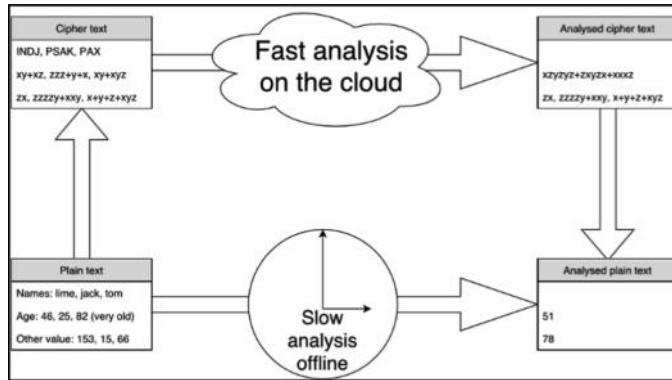


Fig. 1. The advantage of being able to analyze encrypted data

There are examples of security systems allowing processing of hidden messages for a specific purpose. One example allows a group to decide whether to veto a decision or not. This is done whilst providing anonymity to the person who made the decision to veto [3].

For more general-purpose systems, protocols such as ring learning with errors [4] have attracted a lot of attention as good “quantum resistant” solutions that allow for universal computation. This particular protocol has even drawn Google’s attention [5]. This method is a particular case of larger collection of cryptosystems known as SHE’s (somewhat homomorphic encryption) that are based on lattice problems [6]. The issue with these systems is that the encryption functions introduce error. The decryption methods can handle the initial errors however, as operations are performed, the error further increases, potentially to the point at which errors in decryption occur.

The purpose of this paper is to present a homomorphic system that may be useable on devices weaker than computers. The IOT world is rapidly growing in size and many sensors are collecting sensitive data. We of course would like to encrypt that data; however, a lot of schemes have costly encryption functions. We suggest the use of key exchange between sensor and cloud that need only be calculated once. Once the key exchange has taken place, a simple encryption system, that a weak device can handle, is used.

One concern with using homomorphic encryption protocols is that, they provide a weaker privacy guarantee than something like randomized encryption [7]. Other than just the hard problem the protocol is based on, we will consider a few other ways a potential attacker could eavesdrop. We will attempt to adapt the protocol to make those attacks harder/impossible.

In this paper we will begin by looking at how the structure of a ring differs from the more common group theoretic setting. This will lead into what is believed to be a hard problem in this setting, allowing us to build a key exchange protocol. Then using the key from this protocol, we shall define a homomorphic encryption function and

consider the trade off in security and computation needs. Finally, we will look at a potential weakness of the encryption method and suggest ways to lessen the likelihood of a successful attack of that kind.

2 Preliminaries

To begin with, we will briefly revisit the definitions of the groups and rings.

Definition A group is a set G , with a binary operation $* : G \times G \rightarrow G$, where the image of (g_1, g_2) is $g_1 * g_2$ and the following conditions hold:

- **Associativity:** $g_1 * (g_2 * g_3) = (g_1 * g_2) * g_3$ for all $g_1, g_2, g_3 \in G$.
- **Identity:** There exists an *identity element* $e \in G$ such that for all $g \in G$, we have $g * e = e * g = g$.
- **Inverse:** For each element $g \in G$, there exists an *inverse element* $g^{-1} \in G$ such that $g^{-1} * g = e = g * g^{-1}$.

Furthermore, we say that a group is *abelian* if for all $g_1, g_2 \in G$ we have that $g_1 g_2 = g_2 g_1$.

Definition A ring is a set R with two binary operations $+$ and $*$, which we call addition and multiplication, with an identity element 0 for addition, identity element 1 for multiplication, where the following axioms:

- R is an abelian group with respect to addition.
- $(r_1 * r_2) * r_3 = r_1 * (r_2 * r_3)$ for all $r_1, r_2, r_3 \in R$.
- **Distributivity:** $r_1 * (r_2 + r_3) = (r_1 * r_2) + (r_1 * r_3)$ and $(r_1 + r_2) * r_3 = (r_1 * r_3) + (r_2 * r_3)$ for all $r_1, r_2, r_3 \in R$.
- $0 \neq 1$

Next, we describe the particular rings we use in our protocol.

Definition The set of non-commutative monomials in the variables $\{x_1, x_2, \dots, x_n\}$ is defined as the set of sequences

$$\{x_{i_1}, \dots, x_{i_k} | i_1, \dots, i_k \in \{1, \dots, n\}, k \in \mathbb{N} \cup \{0\}\}.$$

The *degree* of the monomial $m = x_{i_1}, \dots, x_{i_k}$ is $\deg(m) = k$ (where the empty sequence has degree 0).

Multiplication of monomials is performed with concatenation of sequences. Standard conventions such as dropping the commas between variables and $x_i^1 = x_i$, $x_i^a x_i^b = x_i^{a+b}$ apply. For example, x_2, x_1 is written as $x_2 x_1$ and the product of $x_2 x_1$ and $x_1 x_3$ is $x_2 x_1^2 x_3$.

We shall need particular types of order on sets of monomials. An order $<$ on a set S is called *total* if either $s_1 \leq s_2$ or $s_1 \geq s_2$, for all $s_1, s_2 \in S$.

Definition An admissible order is a total order on the set of monomials satisfying the following property:

$$m_i \leq m_j \Rightarrow m_i m_k \leq m_j m_k,$$

for all monomials m_i, m_j, m_k .

Often orders are required to be well-orders. An order $<$ on monomials is a well order if

- $<$ is a total order and
- $1 \leq m$, for all monomials m .

Definition The lexicographic ordering on monomials $m_1 = x_{i_1} \cdots x_{i_k}$ and $m_2 = y_{i_1} \cdots y_{i_k}$ defines $m_1 < m_2$ if

- m_1 is a prefix of m_2 or
- If j is the minimum value such that $x_j \neq y_j$ then $x_j < y_j$.

Definition For two non-commutative monomials $m_1 = x_{i_1} \cdots x_{i_k}$ and $m_2 = x_{j_1} \cdots x_{j_{k'}}$, the degree lexicographical (DegLex) ordering sets $m_1 < m_2$ if

$$\deg(m_1) < \deg(m_2) \text{ i.e. } k < k'$$

or

$$\deg(m_1) = \deg(m_2) \text{ and } m_1 < m_2$$

under the lexicographical ordering.

Unless stated otherwise, anytime an ordering is used in the remainder of this paper, DegLex will be chosen.

Definition The polynomial ring, Rx_1, \dots, x_n in x_1, x_2, \dots, x_n over a field R is defined as the set of expressions, called polynomials in x_1, x_2, \dots, x_n , of the form

$$p = \sum_{j=1}^k a_j m_j,$$

where each a_j is a non-zero element of R and the m_j are distinct.

Suppose we have another polynomial $q = \sum_{i=1}^{k'} b_i m'_i$, then addition in the ring is defined as

$$p + q = a_1 m_1 + \cdots + a_k m_k + b_1 m'_1 + \cdots + b_{k'} m'_{k'}$$

where we make the substitution $a_j m_j + b_i m_i = (a_j + b_i) m_i$ if $m_i = m_j$.

Multiplication in the ring is defined as

$$p * q = (a_1 m_1 + \dots + a_k m_k) * (b_1 m'_1 + \dots + b_{k'} m'_{k'})$$

$$\sum_{j=1}^k \sum_{i=1}^{k'} (a_j * b_i) m_j m'_i$$

where we can make the same substitution as addition.

The 3 properties of a polynomial defined next will be the key to establishing “easy” variants of our cryptographic problem later.

Definition Given an order on monomials, suppose a non-commutative polynomial is of the form

$$p = a_1 m_1 + \dots + a_k m_k,$$

where $m_1 > \dots > m_k$. We define

- The leading monomial of p $\text{LM}(p) = m_1$.
- The leading coefficient of p $\text{LC}(p) = a_1$.
- The leading term of p $\text{LT}(p) = a_1 m_1$.

The main purpose of defining an order on the terms in a polynomial is to establish which term leads.

As the title suggests, we are going to be interested in encrypting a plain text using a ring homomorphism.

Definition Given two rings, R and S , then a ring homomorphism is a function $f : R \rightarrow S$ s.t.

- $f(a + b) = f(a) + f(b)$ for all $a, b \in R$.
- $f(ab) = f(a)f(b)$ for all $a, b \in R$.
- $f(1_R) = 1_S$, the multiplicative identity in each ring.

A polynomial ring in n variables will be the setting for our protocol. The entire space however will not be of interest as the problem we are looking to introduce won’t be difficult there. This leads us to look at specific subsets.

Definition If I is a subset of a commutative ring $(R, +, \cdot)$, it is called a two sided ideal of R if:

- $(I, +)$ is a subgroup of $(R, +)$.
- $\forall x \in I, \forall r_1, r_2 \in R : r_1 x r_2 \in I$.

One area we will look at is rewriting systems that produce “nice” (but equivalent) generators of any given ideal.

3 The Ideal Membership Problem

3.1 Suitably Hard Problem

Definition (The ideal membership problem). Given $f_0, f_1, \dots, f_m \in K\langle x_1, x_2, \dots, x_n \rangle$, is it true that $f_0 \in \langle f_1, \dots, f_m \rangle$ (The ideal generated by f_1, \dots, f_m)?

In general, this problem has high computational complexity or is unsolvable. We now have a difficult problem to use as a security basis. The next step is to understand how the problem can be formulated so that the right people can get access to the important data.

The commutative version of the ideal membership problem can be solved using Buchberger's algorithm [8]. This algorithm takes a basis of an ideal and outputs what is known as a Gröbner basis.

Definition (Gröbner basis). A set of non-zero polynomials $G = \{g_1, \dots, g_t\}$ contained in an ideal I , is called a Gröbner basis for I if and only if for all $f \in I, f \neq 0$, there exists $i \in \{1, \dots, t\}$ such that $\text{LM}(g_i)$ divides $\text{LM}(f)$.

Access to a Gröbner basis gives us a straightforward approach to perform membership testing. The goal is to see if a polynomial equals zero after taking the remainder with respect to the ideal. We try dividing a polynomial p by any element g of the basis of I to obtain $p = q_1 g q_2 + r$. If r is 0 then p is in I . Otherwise we repeat using r instead of p . If at any point we have a non-zero polynomial which is not divisible by any of the leading monomials, we can conclude the polynomial is not a member of the ideal.

3.2 Overlapping Polynomials and Reduction

We now need to justify why any arbitrary ideal will not suffice to perform membership testing. The example below, while very simple, illustrates the premise of the problem.

Example Is the polynomial $p = xzy + yzx$ a member of the ideal

$$\langle xy + zx, yx - xz \rangle ?$$

There is no way to multiply a generator on the left or right to give p . However, p can be written as

$$p = -(yx - xz)y + y(xy + zx).$$

so, p does belong to the given ideal. As the number of elements in a basis increases, the number of ways for terms, in particular leading terms, to cancel can increase dramatically. It therefore becomes difficult to determine which polynomials belong to the ideal.

To deal with this issue, we need to consider how terms in polynomials may cancel.

Definition For a monomial m in a non-commutative polynomial ring define:

- prefix (m, i) to be the initial subword of length i of m ,

- suffix (m, i) to be the terminal subword of length i of m ,
- subword (m, i, j) to be the subword starting at position i and ending at position j in m .

Definition Two monomials m_1 and m_2 with degrees $d_1 \geq d_2$ (respectively) overlap if any of the following conditions are true

- prefix $(m_1, i) = \text{suffix}(m_2, i)$ ($1 \leq i < d_2$),
- subword $(m_1, i, i + d_2 - 1) = m_2$ ($1 \leq i \leq d_1 - d_2 + 1$),
- suffix $(m_1, i) = \text{prefix}(m_2, i)$ ($1 \leq i < d_2$).

In order to avoid the kind of problem we saw in the example above, we consider the ways in which leading terms of polynomials overlap.

Definition Assume we have polynomials p_1 and p_2 such that

$l_1 * \text{LM}(p_1) * r_1 = l_2 * \text{LM}(p_2) * r_2$, where we choose l_1, l_2, r_1 and r_2 in a way such that at least one of l_1 and l_2 and at least one of r_1 and r_2 are equal to 1. The S-polynomial associated with this overlap is

$$S - \text{pol}(l_1, p_1, l_2, p_2) = c_1 l_1 p_1 r_1 - c_2 l_2 p_2 r_2,$$

where $c_1 = \text{LC}(p_2)$ and $c_2 = \text{LC}(p_1)$.

S-polynomials will be used in the algorithm to find a Gröbner basis. At the end of the algorithm we may need to remove some of these from the output basis as it possible that they are a sum of the other basis elements. We fix this problem by performing the division algorithm.

Algorithm (Noncommutative division algorithm)

Input: A nonzero polynomial p and a set of polynomials $P = \{p_1, \dots, p_m\}$ over a polynomial ring $R\langle x_1, \dots, x_n \rangle$; an admissible monomial ordering

$r = 0$.

while ($p \neq 0$) **do**

$u = \text{LM}(p)$, $c = \text{LC}(p)$, $j = 1$, found = false;

while ($j \leq m$) **and** (found == false) **do**

if ($\text{LM}(p_j) | u$) **then**

found = true;

choose u_l and u_r such that $u = u_l \text{LM}(p_j) u_r$;

$p = p - (c \text{LM}(p_j)^{-1}) u_l p_j u_r$;

else

$j = j + 1$;

end if

end while

if (found == false) **then**

$r = r + \text{LT}(p)$, $p = p - \text{LT}(p)$;

end if

end while

Output: $\text{Rem}(p, P) = r$, the remainder of p with respect to P .

3.3 Finding a Gröbner Basis

It turns out that finding all the S-polynomials from the original basis of an ideal is all that needs to be done to find a Gröbner basis. The following algorithm iterates over pairs of generators of an ideal and adds all S-polynomials as required.

Algorithm (Mora)

Input: $F = \{f_1, \dots, f_s\}$, a basis for an ideal I over a non-commutative polynomial ring.

G = F ;

A = \emptyset ;

For each pair of polynomials (g_i, g_j) in G , add an S-polynomial to A for each of the possible overlaps of the lead monomials of g_i and g_j .

While $A \neq 0$,

 Remove the first entry s_1 from A ;

$s'_1 = \text{Rem}(s_1, G)$;

If $s'_1 \neq 0$ **then**,

 Add s'_1 to G and then add all the S-polynomials of the form S-pol (l_1, g_i, l_2, s'_1) to $A \forall g_i \in G$.

Output: $G = \{g_1, \dots, g_t\}$ a Gröbner basis for I (Assuming termination).

3.4 Finding the Truly Hard Instances of the Problem

Buchberger's algorithm in a commutative ring is guaranteed to terminate due to Hilbert's basis theorem [9]. There is no equivalent of Hilbert's basis theorem in the non-commutative setting. Thus, we have no such guarantee. The non-commutative Polly Cracker cryptosystem takes advantage of the fact that some ideals do not have a finitely generated Gröbner basis [10]. An example is given in the following theorem,

Theorem Let K be a finite field and $K\langle x, y, z \rangle$ the polynomial algebra over K . Let $g_1 = xzy + yz$ and $g_2 = yzx + zy \in K\langle x, y, z \rangle$. Then, $I = \langle g_1, g_2 \rangle$ does not have a finite Gröbner basis under any admissible order.

We now have all the tools needed to explain how both our key exchange and encryption protocols work.

4 The Protocol and Operations

4.1 Key Exchange Protocol

Alice and Bob publically agree on an ideal $I \subset K\langle x_1, \dots, x_n \rangle$, with the property that Mora's algorithm will not terminate if this ideal is chosen as an input.

Alice in secret chooses a second ideal, I_A , that has a Gröbner basis

$$G_A = \{g_{A,1}, g_{A,2}, \dots, g_{A,n_A}\},$$

Such that $I_A \supset I$.

Alice then generates a polynomial as a linear combination of her private key

$$S_{A,0} = \sum_i^{n_A} l_i g_{A,i} r_i \in I_A,$$

Where $l_i, r_i \in K\langle x_1, \dots, x_n \rangle$. She also generates a collection of elements that aren't in her private ideal but are "close". The m elements generated are denoted

$$S_{A,1} = S_{A,0} + \epsilon_1 \notin I_A,$$

⋮

$$S_{A,m} = S_{A,0} + \epsilon_m \notin I_A.$$

Figure 2 gives an idea of how Alice generates her polynomials. Although the public ideal (and by extension Alice's private ideal) are infinite in size, it gives an impression of how the polynomial ring has been split up and how Alice chooses one polynomial per coset.

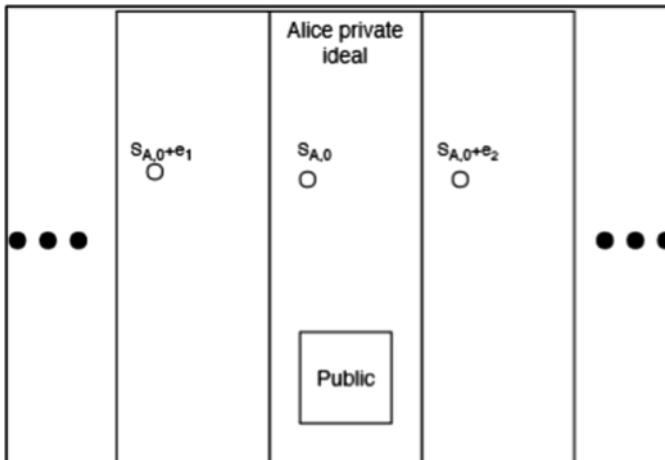


Fig. 2. Alice selects polynomials from her ideal and its cosets

The idea of “close” and our choice of ϵ s will be made clearer in Sect. 4 but for now recall that a ring has the property of being an abelian group under addition. With this in mind, we choose the ϵ s in such a way that each $S_{A,i}$ is in a unique coset of the ideal.

Alice randomises the order of the polynomials and then sends them to Bob over a public channel, making sure to remember at which position she placed $S_{A,0}$. At this point there should be no means for an eavesdropper to learn anything useful about Alice’s choices with only access to the public ideal.

Bob now picks an element of the public ideal, denoted S_B , that he would like to use as the shared secret key. He also picks at random one of the polynomials, $S_{A,j}$ which Alice has sent. The difference between these polynomials

$$B_{\text{diff}} = S_B - S_{A,j}$$

determines the polynomials sent back to Alice. Bob sends

$$\begin{aligned} S_{B,0} &= S_{A,0} + B_{\text{diff}} \notin I, \\ S_{B,1} &= S_{A,1} + B_{\text{diff}} \notin I, \\ &\vdots \\ S_{B,j} &= S_{A,j} + B_{\text{diff}} \in I, \\ &\vdots \\ S_{B,m} &= S_{A,m} + B_{\text{diff}} \notin I, \end{aligned}$$

crucially in the same order he received them.

Bob’s goal is to use an element of the public ideal as the secret key. He does so in Fig. 3 by adding an appropriate polynomial to $S_{A,2}$ but he also does the same to each of the other polynomials that Alice sent as well.

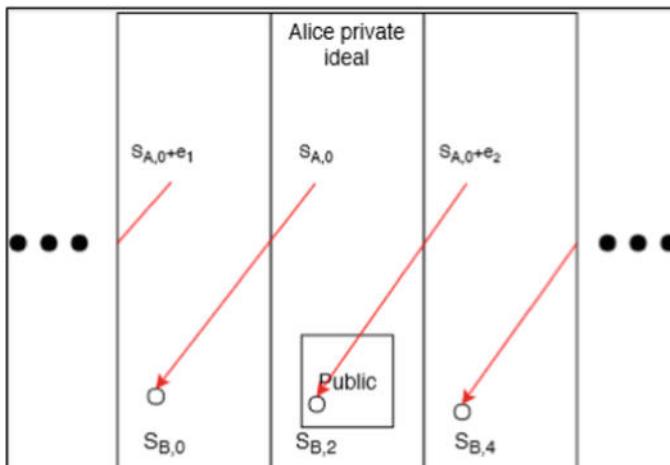


Fig. 3. Bob adds a polynomial to each of Alice’s so one lies in the public ideal

At this point the polynomial chosen to be used as the secret key is being sent over a public channel. To test the polynomials however would require a Gröbner basis for I so it should reach Alice without the eavesdropper gaining access to S_B .

$$\begin{aligned} S_{B,0} &= S_{A,0} + B_{\text{diff}} \\ &= S_{A,0} + (S_B - S_{A,0} - \epsilon_j) \\ &= S_B - \epsilon_j, \end{aligned}$$

where S_B reduces to zero as any element of the public ideal will reduce to zero and therefore will also in Alice's private ideal.

Figure 4 illustrates why it was so important that Bob added the same polynomial to each polynomial that he received from Alice. Even though Alice doesn't know which polynomial that Bob's sent back is his public key choice, she can locate $S_{B,0}$ as the order is maintained. She then computes $S_{B,0} - S_{A,0} = (S_{A,0} - S_B) + \epsilon_j$ and can find ϵ_j since $S_{A,0} - S_B \in I$.

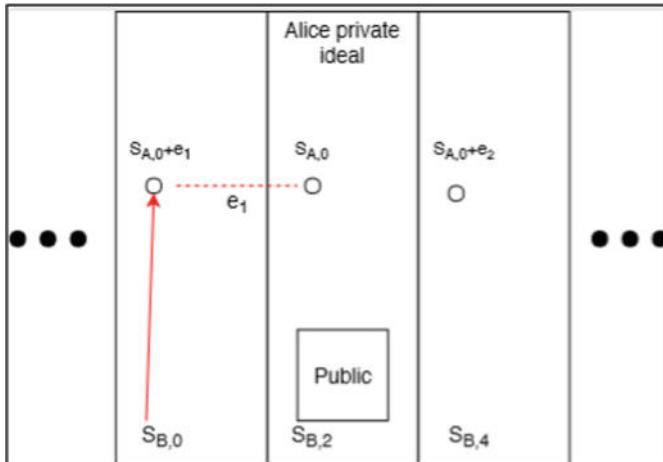


Fig. 4. Alice knows the difference between her polynomials so can get back to her $S_{A,0}$

4.2 Encryption and Evaluation Process

Cryptosystems based on braid groups take advantage of the idea that the conjugacy problem is hard. Computationally conjugacy is a fairly simple operation to perform, so we base our encryption method on this. For the purposes of this particular paper, the operation may seem a little arbitrary however, further work is currently being done to test this system in a Hecke algebra setting, which ties in closely with braid groups [11].

We will denote the private key p and use this to generate a second private key $q = p^{-1}$. The key q is not the inverse of p in the original ring R , but in the factor ring R/J where

$$J = R\langle pq - 1, qp - 1 \rangle R.$$

Encryption method:

$$\text{Enc}(m) = pmq + qmp$$

While the encryption method needs to be simple to make it possible for weak devices to compute, the computational operations performed on the encrypted data and the decryption are performed on more powerful devices. The time-consuming parts of the method therefore are concentrated on the latter.

Decryption method:

$$\begin{aligned} \text{Dec}(c) &= q \cdot \text{Enc}(m) \cdot p \\ &= qpmqp + q^2mp^2 \\ &= m + q^2mp^2 \end{aligned}$$

In our quotient space the second term will be killed off, thus leaving us with just our original message.

Homomorphisms: Encryption is a ring homomorphism since

$$\begin{aligned} \text{Enc}(m_1) + \text{Enc}(m_2) &= pm_1q + qm_1p + pm_2q + qm_2p \\ &= p(m_1 + m_2)q + q(m_1 + m_2)p \\ &= \text{Enc}(m_1 + m_2) \text{ and} \\ \text{Enc}(m_1) \cdot \text{Enc}(m_2) &= (pm_1q + qm_1p) \cdot (pm_2q + qm_2p) \\ &= pm_1qpm_2q + pm_1q^2m_2p + qm_1p^2m_2q + qm_1pqm_2p \\ &= pm_1m_2q + qm_1m_2p + pm_1q^2m_2p + qm_1p^2m_2q \\ &= \text{Enc}(m_1 \cdot m_2) + pm_1q^2m_2p + qm_1p^2m_2q \end{aligned}$$

Once again, an appropriate choice of ideal will be necessary to only leave the terms for an exact homomorphism.

Example Alice and Bob agree to use the ideal from the earlier theorem as their public information. Alice creates her private ideal

$$\langle xzy + yz, yzx + zy, xz + x, z^4y - z^3y \rangle$$

and chooses her first polynomial

$$\begin{aligned}
S_{A,0} &= y(xzy + yz)x + z(yzx + zy)y \\
&\quad + (xz + x)z + y(z^4y - z^3y)y \\
&= yxzyx + y^2zx + zyzxy + z^2y^2 \\
&\quad + xz^2 + xz + yz^4y^2 - yz^3y^2,
\end{aligned}$$

followed by selecting coset elements $\epsilon_1 = y^3 + yzy - xz^2 + xz$ and $\epsilon_2 = z^5 - z^2yz$. She sends to Bob the polynomials $S_{A,1} = S_{A,0} + \epsilon_1$, $S_{A,0}$ and $S_{A,2} = S_{A,0} + \epsilon_2$ over a public channel, keeping in mind the order she sent the polynomials.

Bob now chooses an element of the public ideal

$$\begin{aligned}
S_B &= xy(xzy + yz)x + z(yzx + zx)yz \\
&= xyxzyx + xy^2zx + zyxzy + z^2xyz.
\end{aligned}$$

He then chooses at random one of Alice's polynomials, say $S_{A,1}$. The difference between the two polynomials is

$$\begin{aligned}
B_{diff} &= xyxzyx + xy^2zx + zyxzy + z^2xyz \\
&\quad - yxzyx - y^2zx - zyzxy - z^2y^2 \\
&\quad - y^3 - yzy - yz^4y^2 + yz^3y^2.
\end{aligned}$$

He then sends back S_B alongside

$$\begin{aligned}
S_{B,0} &= yxzyx + y^2zx + zyzxy + z^2y^2 \\
&\quad + xz^2 + xz + yz^4y^2 - yz^3y^2 \\
&\quad + xyxzyx + xy^2zx + zyxzy + z^2xyz \\
&\quad - yxzyx - y^2zx - zyzxy - z^2y^2 \\
&\quad - y^3 - yzy - yz^4y^2 + yz^3y^2
\end{aligned}$$

and

$$\begin{aligned}
S_{B,2} &= yxzyx + y^2zx + zyzxy + z^2y^2 \\
&\quad + xz^2 + xz + yz^4y^2 - yz^3y^2 \\
&\quad + z^5 - z^2yz \\
&\quad + xyxzyx + xy^2zx + zyxzy + z^2xyz \\
&\quad - yxzyx - y^2zx - zyzxy - z^2y^2 \\
&\quad - y^3 - yzy - yz^4y^2 + yz^3y^2.
\end{aligned}$$

Having received the 3 polynomials, Alice now looks at $S_{B,0}$.

$$\begin{aligned}
& [y(xzy + yz)x + z(yzx + zy)y \\
& \quad + (xz + x)z + y(z^4y - z^3y)y \\
& \quad + xy(xzy + yz)x + z(yzx + zx)yz \\
& \quad - y(xzy + yz)x - z(yzx + zy)y \\
& \quad - y(z^4y - z^3y)y] \\
& \quad - y^3 - yzy.
\end{aligned}$$

Alice can reduce everything within the square brackets to zero using her generators leaving just $-y^3 - yzy$. This tells Alice that Bob chose to use the polynomial associated with ϵ_1 and she can therefore determine the shared private key S_B .

Now either Alice or Bob can look at encrypting and operating on data. We already discussed that an appropriate choice of quotient space is needed to make the calculations work. Let's assume Alice and Bob repeat the process once again to get a second secret key. For the sake of brevity, we'll say $yxy + z$. Now consider the ideal

$$\langle S_B(yxy + z) - 1, (yxy + z)S_B - 1 \rangle.$$

Our requirement that we have two secret keys that satisfy the inverse property will hold if we quotient out our polynomials in our ciphertext by this ideal i.e. anytime we see the product of our two keys appear in a ciphertext, we can replace it with a 1.

The example of encryption, basic operations and decryption are given in the appendix. Security and practical concerns.

4.3 Choosing an Appropriate Quotient Space

When looking at both the decryption function and the homomorphic property of our system, there is a strong reliance on the choice of quotient space to make sure the calculations are performed correctly. The issue with the example given is that we have made public, the product of our 2 secret keys in the quotient. Unlike in the integers, factoring 2 polynomials is not a difficult problem. This means any eavesdropper that can observe the operations being performed can find both keys.

The solution to this problem is to only use that quotient space in the decryption where we can be surer that no one will see the function used. This does mean that whilst encrypted our polynomials are growing in size faster as we don't have the ability to cancel terms.

It isn't all bad news however as our choice of our second key was independent of our original key. It does seem reasonable therefore that we can quotient out by the square of that key. This will enable us to remove some of the unwanted terms from multiplication whilst not revealing information about our first key.

4.4 Implementation

Questions that may need to be addressed if there was to be a practical implementation of the key exchange protocol are:

- There is an important requirement that the polynomials sent back and forth maintain their order. What can be done to ensure Alice and Bob are confident that they have received the polynomials in intended order? Pairing the polynomials with a list number is an obvious approach but then implementing some kind of error correction would be necessary.
- Depending on how the ordering is recorded, even if only a subset of polynomials fails to be transferred, all may have to be resent. Due to how the polynomials are related, it may be possible to send equivalent information without needing to send all of the polynomials in whole.

4.5 Partial Gröbner Bases

For a given polynomial and a fixed ideal, it isn't necessary to find all Gröbner basis elements in order to successfully perform the membership test. Only the elements used to generate the given polynomial are needed. As before we stick with the DegLex ordering throughout. Although we stated that many different orderings will suffice when looking at Gröbner basis, the advantage of this choice here is that we can place an upper bound on the degrees of polynomials we need in our basis. If we could say for sure that we'd found all the polynomials up to that degree, then we could end Mora's algorithm.

Figure 5 gives the degree of the polynomial output on particular examples at each iteration of the algorithm. As you can see there is a clear upward trend. However, it would be difficult to say for sure the last time a degree n polynomial would appear.

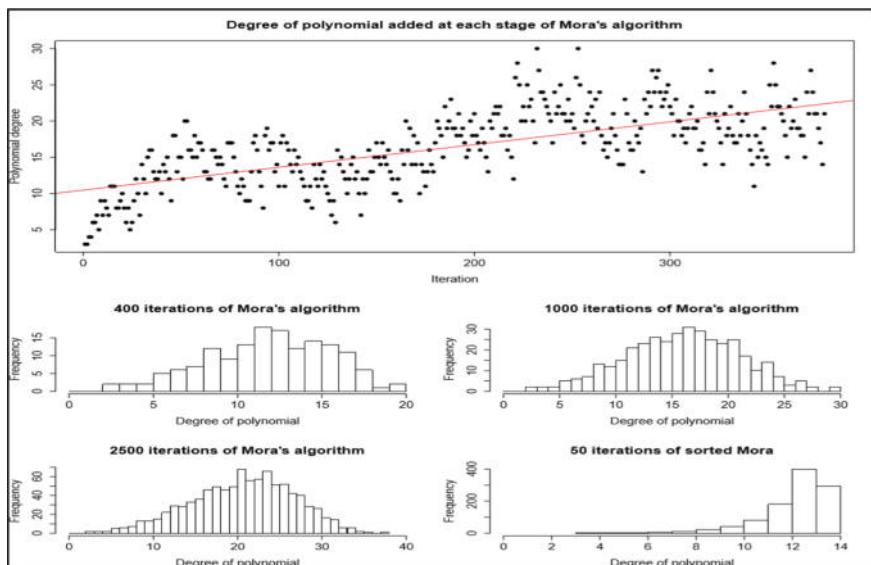


Fig. 5. Summary stats for the distribution of degrees output by Mora's algorithm

This is backed up by looking at the distribution of degrees after different number of iterations. There is no clear sign that we are exhausting the lower degree polynomials.

A potential solution to that problem may be to sort all of the S-polynomials first. This way you would most likely include all the degree 1 polynomials first, then move onto degree 2 and so forth. The bottom right plot gives an example of such a process.

As you can see, there appears to be exponential growth in the number of polynomials as degree increases. The algorithm may not need to run forever, but it appears that it will last for an unreasonable length of time for any meaningful attack.

5 Potential Attack and Protection Against It

5.1 Weakness of Key Exchange

Although Eve does not have access to the private ideal that Alice has generated, she could create her own ideal with a finite Gröbner basis that contains the public ideal. This would allow her to successfully perform the membership test on Bob's choice of secret key. While this is a concern, Eve's choice of ideal may also accept Bob's other polynomials as members of the ideal too, thus making it unclear which is the actual choice.

We will denote the polynomials that Bob sends back to Alice that falls into Eve's ideal as the set $\{S_{Bj_1}, S_{Bj_2}, \dots, S_{Bj_m}\}$.

Our goal to prevent an attack like this is to maximize the value of m , or at the very least try increase as much as possible the likelihood of m being at least 1. This turns our interest towards the intersection of ideals. Figure 6 gives an impression of how we want to set up our polynomials to fool and eavesdropper.

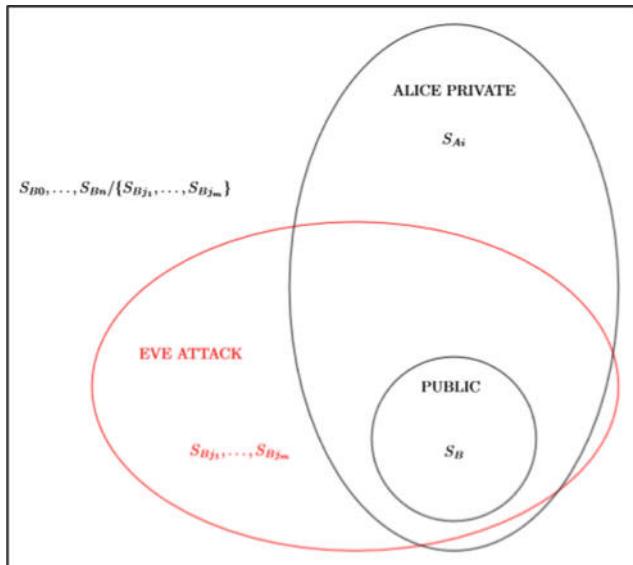


Fig. 6. Alice wants to choose polynomials that likely will be captured by Eve's ideals

Definition Let j and n be natural numbers such that $1 \leq j \leq n$. A monomial order is of the j -th elimination type provided that any monic monomial involving x_1, \dots, x_{j-1} or x_j is greater than any monic monomial of $K\langle x_{j+1}, \dots, x_n \rangle$.

Theorem For 2 ideals $I = \langle f_1, \dots, f_k \rangle, J = \langle g_1, \dots, g_l \rangle \in K\langle x_1, \dots, x_n \rangle$ then $I \cap J = H \cap K\langle x_1, \dots, x_n \rangle$ where

$$H = (tf_i, (1-t)g_j, tx_m - x_mt | 1 \leq i \leq k, 1 \leq j \leq l, 1 \leq m \leq n) \\ \in K\langle x_1, \dots, x_n, t \rangle.$$

Corollary Let G be a Gröbner basis for H according to the elimination order in $K\langle x_1, \dots, x_n, t \rangle$ with $t \geq x_1, \dots, x_n$. Then $G \cap K\langle x_1, \dots, x_n \rangle$ is a Gröbner basis for $I \cap J$.

Note that so far, we have not used the variable t in any of our polynomials. This means that the elimination ordering and DegLex ordering would have acted the same way regardless of which we chose to use.

5.2 Choosing Good Polynomials

Now that we know all the properties required to generate Alice's polynomials, we will return to our example from the previous section. Recall that we chose 3 polynomials to transfer to Bob. We will now show that these choices come from distinct cosets of Alice's private ideal.

Example

1. $\langle I, xz + x, z^4y - z^3y \rangle$,
2. $\langle I, yz + y, xz^4 - yzxz \rangle$,
3. $\langle I, zx + z, xz^4 - z^2yx, y^3x + yzy, z^3y - zyz \rangle$.

To make it easier to find polynomials that lie in one ideal but not the others we will find what is known as a reduced Gröbner basis. While still generating the same ideal, it will make it much easier to work out how to choose polynomials fitting that requirement.

Definition Let $G = \{g_1, \dots, g_p\}$ be a Gröbner basis for an ideal in a polynomial ring. Then G is a reduced Gröbner basis if the following 2 conditions hold:

- $\text{LC}(g_i) = 1$ for all $g_i \in G$.
- None of the terms in the polynomials $g_i \in G$ are divisible by any $\text{LT}(g_j), j \neq i$.

Algorithm (Reduced Gröbner basis)

Input: $G = \{g_1, \dots, g_m\}$ a Gröbner basis for an ideal J over a noncommutative ring

$G' = \emptyset$

For each $g_i \in G$ **do**

Multiply g_i by $\text{LC}(g_i)^{-1}$;

If $(\text{LM}(g_i) = l\text{LM}(g_j)r)$ for some monomials l, r and some $g_j \in$

G ($g_j \neq g_i$) **then**

$G = G \setminus \{g_i\}$;

End if

End for

For each $g_i \in G$ **do**

$g'_i = \text{Rem}(g_i, G \setminus \{g_i\}) \cup G'$;

$G = G \setminus \{g_i\}$, $G' = G' \cup \{g'_i\}$;

End for

Output: $G = \{g'_1, \dots, g'_p\}$ a unique reduced Gröbner basis for J .

Reduced Gröbner basis

1. $\langle xy, yz, xz + x \rangle$,
2. $\langle xy + y^2, zy + y, y^2 + zy, -yx + zy, yz + y, xz^4 - yzxz \rangle$,
3. $\langle -zy^3 + (zy)^2, zy^2x + yzy, zyx + zy, z^2y^2 - z^2yz, z^5 - z^2yz, yz - zy, xz^2y - yz^2x, xzy + yz, zx + z, xz^4 - z^2yx, y^3x + yzy, z^3y - zyz \rangle$.

To begin with we will start with generating an element from the 3rd ideal, this should be the simplest place to start as there are fewer leading monomials in the other two ideals to be concerned about. From the first ideal we can see that the leading monomial cannot contain the substrings xy, yz, xz . From the second ideal we see that the leading monomial cannot contain the substrings zy, y^2, yx, xz^4 . The first thing we notice is that there is no possible combination of letters containing a y allowed. Also, due to the xz monomial the only polynomials left that we can form out of the 3rd ideals generators have a leading monomial of the form $z^m x^n$, for $m, n \geq 1$.

We could spend time working out all the possible lead monomials for the polynomial in ideal 2, however for sake of example we note that this is the only ideal that contains a lead monomial with only ys .

As for ideal 1, once again, any monomial containing y will also be divisible by the lead monomial of at least one polynomial in ideal 2. This eliminates all but one of our generators from ideal 1. Our polynomial used here therefore will be of the form $l(xz + x)r$, ensuring that l and r aren't chosen in such a way to contain a substring of the lead monomials from the other ideals.

5.3 Checking the Overlap of Ideals

Now that we understand how Alice chose her ϵs , we would like to ensure that these choices of ideals were suitable by examining their intersection. Our starting point for creating these ideals was to use the public infinite ideal, so our goal is to see each pairwise intersection be a larger ideal than that.

Even for seemingly small ideals finding a Gröbner basis for the intersections takes a long time. We don't need to find the full basis however, just examples of elements that aren't in the public ideal. In the intersection of the first 2 ideals we have the polynomial

$$xzy + xy = (xz + x)y = x(zy + y).$$

The middle representation shows it as a sum of ideal 1's basis, whereas the right representation shows it as a sum of ideal 2's basis. It cannot be a member of our public ideal as the second term doesn't contain a z . Every term in the public ideal does contain a z so it would be impossible for it to be a member.

Now within the intersection of ideal 1 and ideal 3 we have the polynomial

$$z^4y - z^3y = z(z^3y - zyz) + z^2(yz - zy).$$

The left side of the equation is a generator of ideal 1 and the right-hand side is expressed as a sum of generators from ideal 3 [10]. Discusses how the public ideal could not contain polynomials whose leading term of the form $z^m y$, therefore our polynomial could not be a member.

As you can see, the public ideal is included (not equal) to the intersection of Alice's ideal and each of her decoys. This means that there is a chance that an Eavesdropper may fall for the trap.

6 Discussion

While access to cipher texts that are in ring spaces could lead to meaningful data analysis, the computational resources needed for a system such as the one described in this paper is still too high. Regardless of the protocol, polynomials are going to be a more difficult data structure to use as opposed to large integers. A lot of code that finds Gröbner bases focusses on commutative versions as it has a simpler structure to operate on. As suggested earlier in this paper, further research into something like Hecke algebras may help alleviate this problem. Making use of something like the conjugacy problem may allow smaller keys to be used.

Despite spotting a potential attack on the key exchange, the current "fix" only provides a flavor as to how someone may try to prevent it. Some cryptanalysis should be performed to try and get an understanding of what choice Eve may make to attempt an attack and focus on stopping those in particular. On top of that, to prevent a brute force attack, it would be greatly helpful to increase the number of possible keys without Alice and Bob needing to transmit much more information. One such approach may be to permute or take subsets in some way from the collection of polynomials.

7 Conclusion

In this paper we examined the ideal membership problem as the basis of a homomorphic encryption scheme. We were able to setup an encryption function that was simple and allowed for both addition and multiplication of cipher texts.

There is still concern about how badly the time it takes to perform operations increases as cipher texts grow. This appears to be a common problem in many attempts to develop homomorphic cryptosystems. It seems one of the best approaches to making something that can work in a reasonable time is Gentry's fully homomorphic scheme [12].

The problem of long computation was made worse because we couldn't cancel a lot of our terms until the very end. This was salvaged somewhat by being able to quotient out terms that included a square of our second key. Further work needs to be done to see if there is a subtler choice of quotient space that would still enable the required terms to cancel.

We have shown that the intersection of Alice's ideal with her decoys is larger than the public ideal. How much of Alice's ideal is overlapped by the decoys however isn't clear. Further work needs to be done to ensure a significant probability of an eavesdropper making a bad choice of attack ideal.

Finally, the biggest question brings us back to the first sentence in this paper, can this system guarantee the safety of users' data? Problems such as factoring integers has been around for a long time and no well-known solution (outside of a quantum solution) is out there. We can therefore be fairly trustworthy of protocols such as RSA. The ideal membership problem is a lot newer and has yet to prove its resilience to the same extent from cryptanalysis. Despite our attempts to cover the potential alternative attacks, it may be that the initial problem itself needs a better understanding of which instances can be considered safe.

Appendix

To prevent tedious amounts of expanding brackets, we will keep a lot of the polynomials factored and we will also make the substitutions

$$\begin{aligned} A &= xyxzyx, B = xy^2zx, C = zyxyz, D = z^2xyz, \\ W &= yxy. \end{aligned}$$

This means that we have the public ideal $\langle (W+z)^2 \rangle$ that we can use to reduce terms while in the cloud. We will also have the private ideal $\langle (A+B+C+D)(W+z)-1, (W+z)(A+B+C+D)-1 \rangle$ to be used for further reduction once offline. Now let's do a simple calculation.

$$\begin{aligned}
& (Enc(x) + Enc(y)) * Enc(z) \\
&= ((A + B + C + D)x(W + z) + (W + z)x(A + B + C + D) \\
&\quad + (A + B + C + D)y(W + z) + (W + z)y(A + B + C + D)) \\
&\quad * ((A + B + C + D)z(W + z) + (W + z)z(A + B + C + D)) \\
&= ((A + B + C + D)(x + y)(W + z) + (W + z)(x + y)(A + B + C + D)) \\
&\quad * ((A + B + C + D)z(W + z) + (W + z)z(A + B + C + D)) \\
&= (A + B + C + D)(x + y)(W + z)(A + B + C + D)z(W + z) \\
&\quad + (A + B + C + D)(x + y)(W + z)^2z(A + B + C + D) \\
&\quad + (W + z)(x + y)(A + B + C + D)^2z(W + z) \\
&\quad + (W + z)(x + y)(A + B + C + D)(W + z)z(A + B + C + D).
\end{aligned}$$

Our public ideal space kills off multiples of $(W + z)^2$, so that leaves us with

$$\begin{aligned}
& (A + B + C + D)(x + y)(W + z)(A + B + C + D)z(W + z) \\
&\quad + (W + z)(x + y)(A + B + C + D)^2z(W + z) \\
&\quad + (W + z)(x + y)(A + B + C + D)(W + z)z(A + B + C + D).
\end{aligned}$$

Performing the decryption method, we have

$$\begin{aligned}
& Dec((Enc(x) + Enc(y)) * Enc(z)) \\
&= (W + z)((A + B + C + D)(x + y)(W + z)(A + B + C + D)z(W + z) \\
&\quad + (W + z)(x + y)(A + B + C + D)^2z(W + z) \\
&\quad + (W + z)(x + y)(A + B + C + D)(W + z)z(A + B + C + D))(A + B + C + D) \\
&= (W + z)(A + B + C + D)(x + y)(W + z)(A + B + C + D)z(W + z)(A + B + C + D) \\
&\quad + (W + z)^2(x + y)(A + B + C + D)^2z(W + z)(A + B + C + D) \\
&\quad + (W + z)^2(x + y)(A + B + C + D)(W + z)z(A + B + C + D)^2.
\end{aligned}$$

Once again reducing according to our public ideal we are left with

$$(W + z)(A + B + C + D)(x + y)(W + z)(A + B + C + D)z(W + z)(A + B + C + D).$$

Now that we are offline we can use the private ideal which gives us

$$1 * (x + y) * 1 * z * 1 = (x + y) * z,$$

as required.

References

1. L. A. M. L. D. Rivest, R. L.: On Data Banks and Privacy Homomorphisms. Foundations of secure computation (1978)
2. X. P. R. a. B. E. Yi: Homomorphic Encryption and Applications. Springer (2014)
3. F. a. Z. P. Hao: The power of anonymous veto in public discussion. In: Transactions on Computational Science, Springer, Berlin (2009)

4. L. K. V. V. Naehrig, M.: Can homomorphic encryption be practical? In: Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop (2011)
5. G. A.: Google tests new crypto in chrome to fend off quantum attacks. www.wired.com/2016/07/google-tests-new-crypto-chrome-fend-off-quantum-attacks/ (2016)
6. V. d. P. J., Lattice-Based Cryptography. Eindhoven (2011)
7. C. J. E. P. R. M. N. W. E. M. S. B. H. a. Z. N. Curino: Relational Cloud: A Database-as-a-Service for the Cloud (2011)
8. Buchberger, B.: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. *J. Symbolic Comput.* 475–511 (1965)
9. W. a. L. P. Adams.: An Introduction to Grobner Bases. American Mathematical Society (1994)
10. Rai, T.: Infinite Grobner Bases and Noncommutative Polly Cracker Cryptosystems (2004)
11. Garber, D.: Braid group cryptography. In: Braids: Introductory Lectures on Braids, Configurations and Their Applications, pp. 329–403 (2010)
12. Gentry, C.: A Fully Homomorphic Encryption Scheme. Stanford University (2009)



Response-Based Cryptographic Methods with Ternary Physical Unclonable Functions

Bertrand Cambou¹✉, Christopher Philabaum¹, Duane Booher¹, and Donald A. Telesca²

¹ School of Informatics Computing and Cyber Systems, Northern Arizona University, Flagstaff, AZ, USA

{Bertrand.cambou, cp723, duane.booher}@nau.edu

² Air Force Research Laboratory, Information Directorate, Rome, NY, USA

Donald.telesca@us.af.mil

Abstract. Physical Unclonable Functions (PUFs) are used as hardware fingerprints for access control, and authentication in mobile and wireless networks and Internet of Things. However, it is challenging to use PUFs to extract cryptographic keys, because a single bit mismatch in the keys is not acceptable to most encryption algorithms. PUFs are aging; they are sensitive to temperature drifts, and other environmental effects. Successful implementation of PUFs, as key generators, requires power hungry error correcting schemes that add latency, and vulnerability to attacks such as differential power analysis. This work proposes methods to generate cryptographic keys directly from the un-corrected responses of the PUFs. The secure server, driving the network, manages the differences between the PUF responses and the original PUF challenges, through matching algorithms, mitigating the need to use heavy error correction schemes. In these methods, both the server and the client devices independently generate the exact same un-corrected responses of the PUF. These responses are therefore suitable for cryptographic protocols such as public key infrastructure or highly secure ledger protecting blockchain technology. The method presented in this paper, which is based on ternary PUFs, was successfully implemented and tested in a PC environment.

Keywords: Mobile security · Access control · Cryptography

1 Introduction: PUF-Based Cryptography

Physical unclonable functions (PUF) are the equivalent of human fingerprints. The manufacturing variations created during fabrication of the devices make each device authenticable from each other. Authentication protocols based on various physical unclonable functions (PUF) [1–10], embedded in each internet of thing (IoT) node, can be effective when the PUFs are showing intra-PUF stability, offer inter-PUF randomness, and when the drifts in the PUF characteristics are small enough. Memory structures [11–19], SRAM [12, 13], DRAM [14], Flash [15], ReRAM [1, 16–18], and MRAM [18, 19], are suitable to generate strong PUFs. One of the methods to generate PUFs from memory devices is to characterize a particular parameter \mathcal{P} of the cells of

the array. The values of parameter \mathcal{P} vary cell to cell, and follow a distribution with a median value T. The cells with $\mathcal{P} < T$ generate “0” states, and the others generate “1” states.

The following definitions are stated for unambiguousness as other authors have used different, however equivalent terminology to describe PUFs. This work is agnostic on which terminology is preferred.

- The “*Client devices*” are the components secured by PUFs. It could be IoT or any other peripheral devices;
- The “*Server*” is the component driving a set of “client devices”.
- The PUF “*Challenges*” are the initial data streams generated during enrolment of the client devices by the server. In some protocols, the Challenges can be the generated by processing, or averaging, multiple queries and measurements of the PUFs. They can also be the result of computations, and statistical analysis. The challenges can be represented by binary, ternary, or any other radix;
- The “*Enrolment*” of the constellation of PUFs is the operation in which the “challenges” of the PUFs located in each client devices are downloaded in a data base, or look up table in the server. This operation is generally done in a secure environment. A block diagram of this operation is shown Fig. 1. Additional PUFs can be added to the network, and enrolled over time;

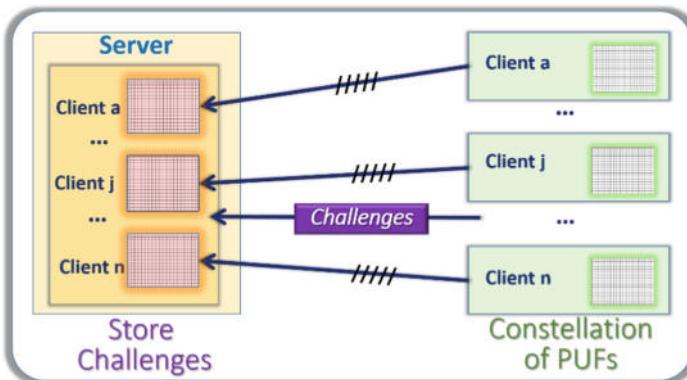


Fig. 1. Enrollment of a constellation of PUFs. The initial PUF responses, the challenges, are stored as references by the server for future authentication.

- The PUF “*responses*” are the data streams generated by the PUFs during the life of the client devices. These PUFs are physical elements that can age, and be subject to temperature changes, electro-magnetic interferences, and other environmental effects;
- “*Challenge-response-pairs*” (CRP) are generated on demand by the server of the PUFs of the client devices;

- When the PUF is a strong PUF having multiple ways to query it, the server needs to send “*instructions*” to the client devices to find the particular “*address*” in the PUF, and to generate challenge-response-pairs;
- The “*helper*” is the data stream generated by the server from the challenges, and transmitted to the client devices for error correction of the responses.

When the CRP error rates are relatively low, the responses can be used as part of authentication protocols, which has significant commercial value to protect cyber physical systems. Error rates between challenges and responses below 10% are low enough for satisfactory authentication, acceptable false rejection rate (FRR), and false acceptance rate (FAR). When the CRP error rates are too high, the use of error correcting methods, and a helper, improve both FAR and FRR. The use of PUFs to generate cryptographic keys from the responses is more challenging than generating responses for authentication. This requires schemes that fully correct the responses of the PUF; a single-bit mismatch in a cryptographic key is not acceptable for most encryption protocols. Ciphers, in particular block ciphers such as DES (data encryption standard), AES (advanced encryption standard), RSA (Rivest Shamir, Adleman encryption), ECC (elliptic curve cryptography), DH (Diffie Hellman), cannot be decrypted with keys having a single-bit mismatch. A typical architecture to drive a network of client devices with PUFs is shown in Fig. 2.

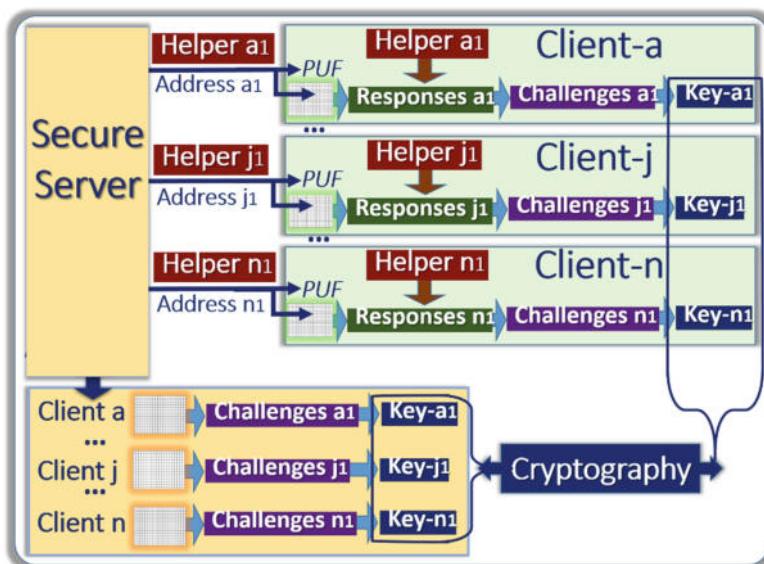


Fig. 2. Diagram of PUF-based cryptography with error correction. The PUF responses generated by the client devices are corrected with the helpers.

- As part as the handshake process, the server initiates the process by sending instructions to the client device j that incorporate the addresses $\text{Address-}j_1$ within the PUF on where to extract responses $\text{Responses-}j_1$.
- The server independently analyses the challenges $\text{Challenges-}j_1$ stored in its look up tables at the corresponding address, and generates $\text{Helper-}j_1$ with fuzzy extractor, and other error correcting methods from $\text{Challenges-}j_1$ [20–28]. The helpers are transmitted to the client devices as part of the handshake.
- The client devices generate $\text{Responses-}j_1$ at the address $\text{Address-}j_1$, and correct them with the helper with fuzzy extractors, and other correcting methods. To be acceptable for encryption, the corrected responses of the client devices, and the challenges of the server should be identical. Thereby the same key $\text{Key-}j_1$ is independently generated for encryption schemes.

Such protocols have two fundamental weaknesses: first, the client devices are burdened, and need to consume additional computing power to run the error correcting codes; second, such protocols increase the vulnerability to side channel attacks, differential power analysis, and potential exposure of the helpers.

This work has identified a new method for addressing these weaknesses. The remaining sections of the paper detail this research and are structured as follows: Sect. 2 describes the generation of cryptographic keys from PUF responses with response-based cryptography (RBC). The burden is transferred from the client devices to the server, which can have access to considerable computing power, and the highest level of security, while the clients need to operate at low power, and in a non-secure environment. In Sect. 3, the performance in terms of latency, and false rejection rate (FRR) is modeled. Section 4 is a generalization of the RBC concept to enhance the security of blockchains and peer-to-peer communication.

2 Response Based Cryptographic Methods

The protocol described in Sect. 1 and Fig. 2 can be called “challenge-based cryptography”. The new method detailed in this manuscript can be called “response-based cryptography” (RBC), which uses un-corrected responses from a network of PUFs securing client devices [29]. A block diagram of a network protected by RBC is shown in Fig. 3. As done in the challenge-based cryptography, the secure server sends instructions to the client device j , the $\text{Addresses-}j_1$, to generate the PUF responses $\text{Responses-}j_1$.

The cryptographic key $\text{Key-}j_1$ is directly generated from the responses. The purpose of the RBC Engine which is driven by the server is to generate from the challenges $\text{Challenges-}j_1$ the responses matching the ones generated by the PUFs, and extract the exact same cryptographic key $\text{Key-}j_1$. The server does not need to generate and transmit helper messages anymore, and the client devices do not need to have error correcting schemes to extract streams matching the challenges stored by the server. The computing power needed at the client level is thereby lowered, which allows the use of less powerful microcontrollers, smaller memory components, and simpler architectures.

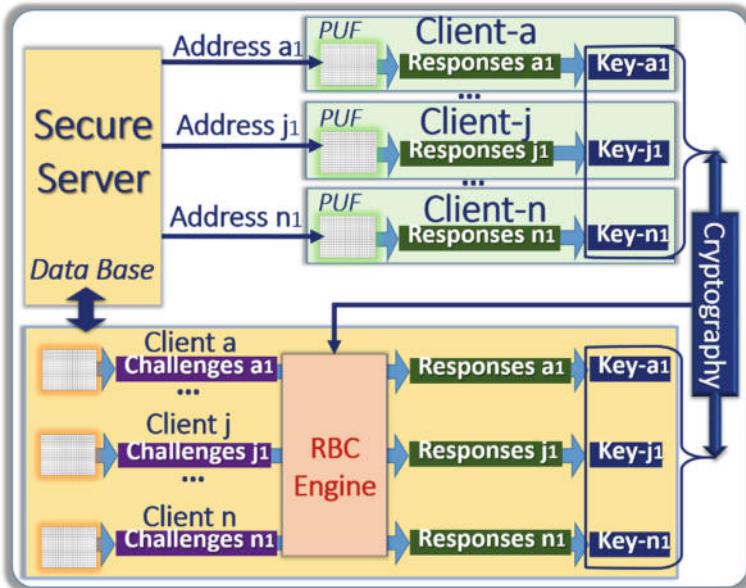


Fig. 3. Diagram of response-based cryptography (RBC). The server extracts identical responses from the challenges with the RBC.

The elimination of the helpers also simplify the communication between server and clients. The latency at the client device level is significantly reduced, giving less time for malicious observers to extract relevant information from the transaction.

2.1 Authentication Protocol with RBC

As part of the authentication protocol, the client device j encrypts a message, $MA-j_1$, with symmetrical scheme such as Advanced Encryption Standard (AES), and the cryptographic key $Key-j_1$. The encrypted message $E(MA-j_1, Key-j_1)$ is transmitted to the server, see Figs. 4 and 5, and analyzed by the RBC engine. Examples of message $MA-j_1$ could be the user ID of client j or other identification scheme. The cipher $E(MA-j_1, Key-j_1)$ can only be decrypted with the cryptographic $Key-j_1$, which is out of reach for malicious parties attacking the network without this key. This protocol does not replace other access control schemes such as multifactor authentication; the main objective of RBC is to validate the key exchange.

We are defining $Responses\ j_{10}$ as the data stream having a hamming distance of zero with $Challenges\ j_1$, which means $Challenges\ j_1 = Responses\ j_{10}$. The key Kj_{10} is extracted from $Response\ j_{10}$, and is used to encrypt $MA-j_1$ to generate the cipher $E(MA-j_1, Kj_{10})$, see block diagram in Fig. 5. If the responses $Responses-j_1$, from client- j , and responses $Responses-j_{10}$ are identical, the authentication of the client device j is positive and:

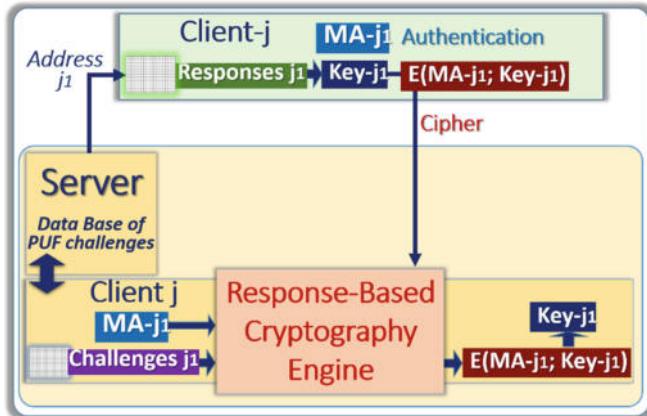


Fig. 4. Protocol of authentication with RBC. The server exploits the encryption of the authentication message sent by the client devices to find the same cryptographic keys.

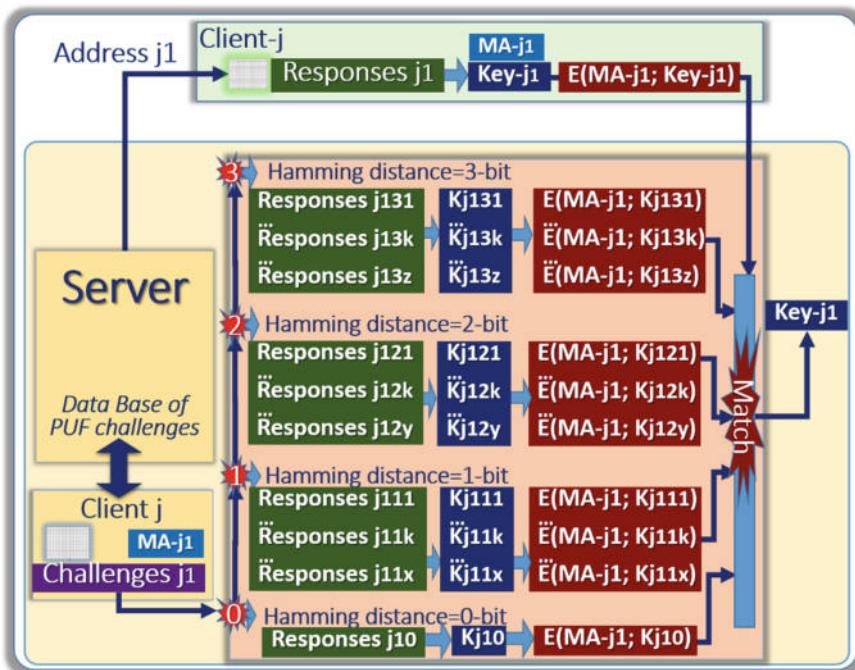


Fig. 5. Method to match the responses from PUF challenges. The server generates data streams with increasingly large Hamming distances to find a match with the cipher of the client device.

$$E(MA - j_1, \text{Key} - j_1) = E(MA - j_1, Kj_{10}) \quad (1)$$

$$\text{Key} - j_1 = Kj_{10} \quad (2)$$

If the error rate is not zero, Eq. (1) is false. We are defining **Responses** j_{11k} , with $k \in \{1, x\}$, as all data streams with Hamming distances with **Challenges** j_1 of 1. For example if the data streams are 256-bit long, then there are 256 streams having Hamming distance of 1 with **Challenges**- j_1 , and $x = 256$. The keys Kj_{11k} , with $k \in \{1, x\}$, are generated from these streams, and are used to encrypt $MA-j_1$, and generate the ciphers $E(MA-j_1, Kj_{11k})$, with $k \in \{1, x\}$.

If the CRP error rate between **Challenges** j_1 and **Responses** j_1 is one, one of these ciphers will be equal to $E(MA-j_1, \text{Key}-j_1)$. The response **Responses** j_{11k} , which generates the cipher equal to $E(MA-j_1, \text{Key}-j_1)$ is therefore equal to **Responses** j_1 , and the authentication of j is positive:

$$\text{Key} - j_1 = Kj_{11k} \quad (3)$$

If the CRP error rate between **Challenges** j_1 and **Responses** j_1 is two, one of the cipher $E(MA-j_1, Kj_{12k})$, with $k \in \{1, y\}$, is equal to $E(MA-j_1, \text{Key}-j_1)$. The response **Responses** j_{12k} , which generates the cipher equal to $E(MA-j_1, \text{Key}-j_1)$ is therefore equal to **Responses** j_1 , and the authentication of j is positive:

$$\text{Key} - j_1 = Kj_{12k} \quad (4)$$

In a similar way, if the CRP error rate between **Challenges** j_1 and **Responses** j_1 is three, one of the cipher $E(MA-j_1, Kj_{13k})$, with $k \in \{1, z\}$, is equal to $E(MA-j_1, \text{Key}-j_1)$. The response **Responses** j_{13k} , which generates the cipher equal to $E(MA-j_1, \text{Key}-j_1)$ is therefore equal to **Responses** j_1 , and the authentication of j is positive:

$$\text{Key} - j_1 = Kj_{13k} \quad (5)$$

The process described above, which is summarized in Fig. 5, can be extended to the generation of data streams having Hamming distances higher than 3. The limitations due to the available computing power and expected latency of such an authentication protocol are studied in Sect. 3; Hamming distances greater than 5 are not practical for mainstream servers.

2.2 Example of Matching Algorithms for RBC

Increased Hamming Distance The authentication protocol described in Sect. 2.1 is the first step of the protocol driven by the RBC engine. Assuming that the computing power of the server device is infinite, the authentication process can iterate, as shown in Fig. 6, to find a stream having the same cipher as the client, $E(MA-j_1, \text{Key}-j_1)$. At one point, if there are no responses matching when the Hamming distance is reaching $a = a_{\max} - 1$, the authentication is considered negative. The trade-off computing power at various level of PUF quality is presented in Sect. 3. When their ciphers are matching,

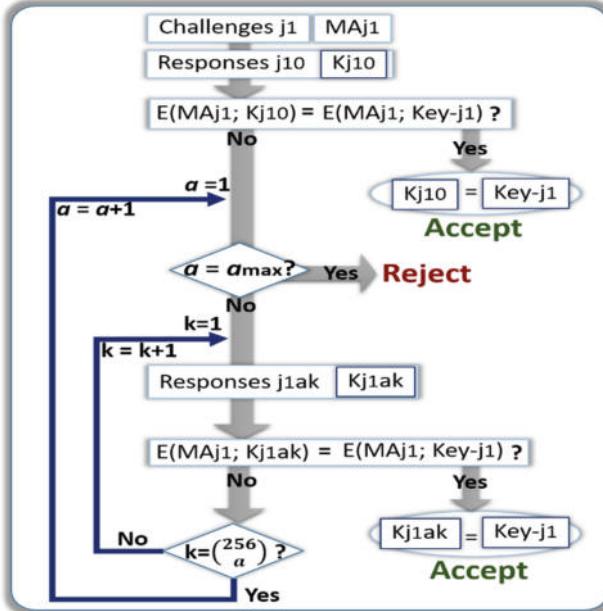


Fig. 6. Response-based key matching algorithm with incremental Hamming distances “ a ”.

both parties do share the same cryptographic key **Key-j₁**, which can be used to protect the communication client-server. In such a protocol, which is based on increasing the hamming distance step by step, the entire burden is placed on the server. The protocol is thereby limited by PUFs with CRP error rates that are low enough, and which have small Hamming distances between challenges and responses.

Multiple Queries The frequency of PUF CRP errors is highly variable. These high CRP rates can be the cause of a bad connection between server and client, a noisy environment, an abrupt temperature variation, or simply the selection of a marginal section of the PUF. Thus, rather than trying to find a match with high Hamming distances, it is faster at some point to stop the process described above, and initiate a new address, address j₂. This revised protocol now places part of the burden on the client devices, by requiring the client to generate a new set of responses, keys, and encrypted messages. As demonstrated in Sect. 3, the CRP error rate varies with each query, but multiple queries reduce the time needed to find matching keys. For example, with a 256-bit long PUF having an average CRP error rate of 0.03%, and limited computing power, 10 s is necessary to find a match with a single query, but only 2 ms are needed with two queries.

Error Correction The time and computing power needed to reduce the errors of the responses is relatively small when the expected error rate, post correction, is not zero. For example, the use of ternary cryptography on PUFs [29–32] is effective in reducing CRP error rates below 0.03% without burdening client devices. The combination of

light correcting methods with increased Hamming distance protocols, and multiple queries make RBC effective, as presented below in Sect. 3.

3 Modelling and Validation of RBC

3.1 Hamming Distance for RBC

In order to demonstrate the practicality of RBC, we need to estimate the length of time needed to generate all response streams with a Hamming distance a and to encrypt them for the purpose of matching the ciphers with $E(MA-j_1, Key-j_1)$. If we assume a 256-bit challenge, then the number of possible streams N_s with Hamming distance of a is given by:

$$N_s = \binom{256}{a} \quad (6)$$

For each stream “ k ” the cipher $E(MA-j_1, Kj_{lak})$ is generated with encryption schemes such as AES, and compared with $E(MA-j_1, Key-j_1)$. Figure 7 is modeling of the time needed to process all streams when “ a ” increases from $a = 0$ to $a = 5$. The time at a given Hamming distance is further reduced with powerful servers. This simple model shows that the time needed exponentially increases with a . On average, the latencies reported in Fig. 7 are reduced by applying smart search methods, and entanglements.

Hamming “ a ”	Numbers of 256-bit streams	Time 1core @ 1GHz	Time 32cores @4GHz
0	$\binom{256}{0}=1$	5 μs	50 ns
1	$\binom{256}{1}=256$	1 ms	10 μs
2	$\binom{256}{2}=32,512$	0.1 s	1 ms
3	$\binom{256}{3}=2,763,520$	10 s	0.1 s
4	$\binom{256}{4}=174,792,640$	10 min	1 s
5	$\binom{256}{5}=8,809,549,056$	10 h	1 min

Fig. 7. Modeling of the time needed to encrypt a group of 256-bit streams with AES, and match it with a reference cipher. The Hamming distance “ a ” varies from 0 to 5.

3.2 Efficiency of RBC at Various CRP Error Rates

Repeating a query is faster than searching for a match with Hamming distance greater than 1. Figure 8 shows the analysis of the efficiency of RBC at different levels of CRP error rates.

256-bit PUF CRP error rate %	Failure rate (FRR) in % to match responses with Hamming distances “ a ”				Queries needed for FRR < 0.1% 0-bit mismatch accepted				Latency for 1 or 2 queries or more FRR < 0.1%						
	$a=2$	$a=3$	$a=4$	$a=5$	$a=1$	$a=2$	$a=3$	$a=4$	$a=5$	PC			Server		
										1Q	2Q	>	1Q	2Q	>
3	96	90	81	69	-	-	-	-	18	-	-	7d	-	18min	
1	36	19	8.4	3.3	-	7	6	3	2	-	20h	0.7s	-	2min	7ms
0.3	7.9	1.9	0.4	0.09	6	3	2	2	1	10h	20s	6ms	1min	0.2s	60μs
0.1	1.7	0.6	0.12	0.03	3	2	2	2	1	10h	0.2s	3ms	1min	2ms	30μs
0.03	0.15	0.01	$1 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	2	2	1	1	1	10s	2ms		0.1s	20μs	
0.01	$3 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$3 \cdot 10^{-6}$	$1 \cdot 10^{-7}$	1	1	1	1	1	1ms			10μs		

Fig. 8. Failure rate (FRR), and latency of RBC at various levels of CRP error rates. ECC is used for the public key generation, and AES to encrypt the authentication messages.

Failure Rates with Several Hamming Distances As shown on the left side of the table of Fig. 8, the false rejection rates (FRR) of RBC, using the Hamming distance method described above in Sect. 2.2 is reduced when the quality of the PUF improves. When the CRP error rates are 1%, or higher, the method is not effective; our implementation of RBC started at 0.1%, and below.

Combination with Queries The mid-section of Fig. 8 shows the respective efficiency of multiple queries, versus increasing “ a ”. Reducing the number of client-server interactions, while important from a security standpoint, must be balanced against the speed at which the server can effectively find a match. For example, as shown in Fig. 8, to authenticate a PUF with a CRP error rate of 1% at a Hamming distance of $a = 2$ requires seven queries to achieve a FRR lower than 0.1%; six queries are needed with $a = 3$, three queries with $a = 4$, and two queries with $a = 5$.

The six columns on the right of Fig. 8 report the modeling of the latency of RBC with an AES scheme as a function of the number of queries needed to get below FRR of 0.1%. The latency is always lower when the application allows multiple queries, as shown in the last 6 columns of Fig. 8. In the example of a CRP error rate of 1%, and FRR < 0.1%, it takes a server 7 ms to authenticate a client device with 7 queries and

$a = 2$, and as much as two minutes with 2 queries and $a = 5$. However, when the CRP error rate is lower, for example 0.03%, a server can authenticate a client device in only one query, at 100 ms with $a = 3$.

3.3 Ternary Cryptography and Ternary PUFs

As demonstrated above, the resistance based cryptography (RBC) is not effective when the Hamming distance “ a ” is too large, because the number of possible streams Ns become out of reach for powerful servers. For example, a CRP error rate of a 256-bit PUF at 3% is already too high for practical implementations. To mitigate this problem, and avoid the need to use error-correcting methods, we designed our implementation with ternary cryptography, and ternary PUFs [29–32]. During enrollment, the challenges/initial responses generated have ternary states, the solid “0”s and “1”s states, and the unstable and marginal fuzzy states “X”. For example, multiple power-off/power-on cycles of an SRAM memory array can sort out the cells with these three states. The generation of ternary PUFs from memory devices is also based on the characterization of a particular parameter of the cells of the array. The bottom 1/3 of the population are the “0”s, the top 1/3 are the “1”s, and the middle third are the “X”s. During response generation, the cells fuzzy states are ignored, which reduces the error rates. As part of the ternary protocol, a mask is transferred to the client device to locate the “good” cells. Multifactor authentication is used to protect the masks from eavesdropping. The ternary protocol that we develop to generate the responses needs less than 800 clock cycles, which is much lower than what is needed to run error-correcting codes.

3.4 Experimental Validation

We analyzed 30 different 32kByte commercial SRAMs, produced by Cypress Semiconductor. The SRAMs with surrounding circuitry are assembled in custom PCB boards. Development boards from Microchip, based on programmable 32-bit RISC microprocessors, are driving the SRAM boards. The microprocessors are programmed with a separate Arduino board. SRAM based PUFs exploit the fact that their cells are designed with flip-flop logic. During the power-off/power-on cycles, the majority of the cells of the SRAM arrays are predictable; they always either flip as a “0”, or as a “1” due to the asymmetries introduced during the manufacturing of the devices; each SRAM device is in general, different from the others. The CRP error rates of SRAM PUFs can be high because some cells are more symmetrical, and flip randomly on both sides. As shown in Fig. 9, the CRP error rates of the SRAM PUFs are reduced with an enrollment that incorporate successive power-off-power-on cycles to remove the erratic cells. During each cycle, and after removal of the bad cells, the error rates of the remaining cells are estimated, and plotted in the graph; after only three cycles the CRP error rate drops below 1%. After 8 cycles the CRP error rate drops below 0.3%; after 27 cycles the error rate drops to 0.1%. In this enrollment protocol, about 50 cycles are needed to generate a ternary PUF with CRP error rates below 0.03%.

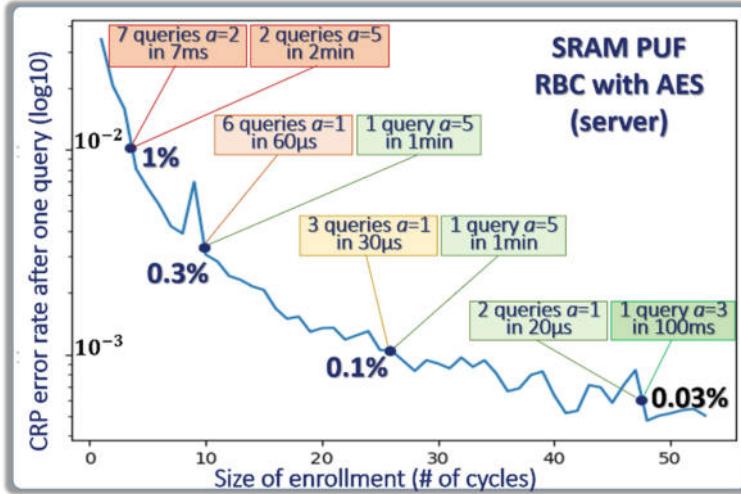


Fig. 9. CRP error rate reduction by masking the fuzzy cells of commercial SRAM. Each cycle: power-off/wait state/power-on/read.

Caution In this experiment, it was very important to impose a complete power-off of the SRAMs. Capacitive and inductive effects can prevent the power to actually switch off. To mitigate the problem, we shorted the SRAM to the ground during these cycles, and imposed a delay before powering on the devices. We also characterized the CRP error rates at various delays, and concluded that waiting times higher than one second did not influence the CRP error rates.

Latency Analysis As presented in section II-B, the RBC matching algorithms are based on increased Hamming distances a , and multiple queries. The latencies needed for RBS matching are plotted in Fig. 9, and this based on the sensitivity analysis presented in Fig. 8, which uses AES. After an enrollment of 50 cycles, RBC authentications with 0.1% FRR are completed in 20 μ s and 2 queries ($a = 1$), or 100 ms and 1 query ($a = 3$) with a 32 core server operating at 4 GHz. A one core PC operating at 1 GHz has a latency of 2 ms and 2 queries ($a = 1$), or 10 s and 1 query ($a = 3$).

Analysis of 100-Cycle Enrolments In the final coding of the prototype demonstrating the RBC protocol, we assumed an enrollment of 100 cycles to sort out the ternary states of the SRAM PUF. The experimental characterization of the CRP error rates of the non-fuzzy cells of the SRAM PUF is shown in Fig. 10. The error rates oscillate around 0.01%, stay consistently (3σ) below 0.02%, and this after 500 successive power-off-power-on cycles to characterize the non-fuzzy cells, which represent about 80% of the 256,000 cells of the SRAM device.

Following the 100-cycle enrollment, we measured the distribution of the Hamming distance between challenges and responses: randomly selected in the SRAM array were one thousand 256 cells having non-fuzzy states. The challenge-response-pairs of these cells were generated to evaluate the Hamming distance a between the 256-bit challenges, and subsequent 256-bit responses. 969 out of 1000 256-bit CRP streams

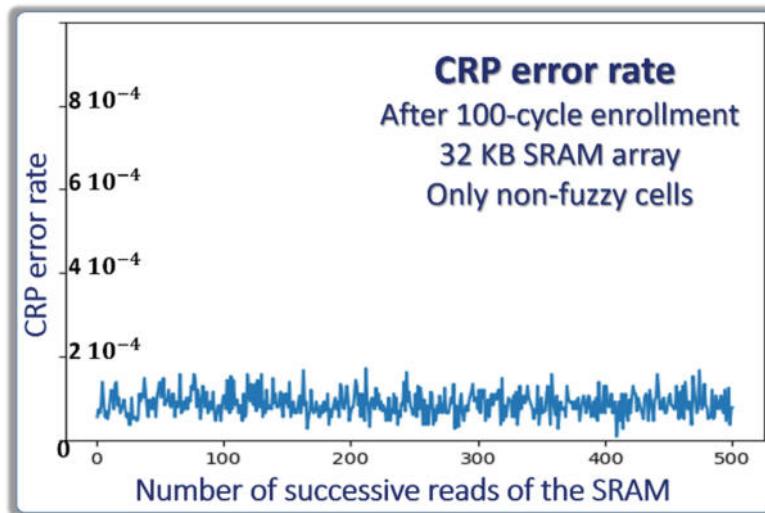


Fig. 10. Characterization of the CRP error rate of T-PUFs after an enrollment of 100 cycles.

had zero mismatch, 30 had one mismatch, and two streams had an Hamming distance of two. None of these 1000 CRP streams had a mismatch of three or higher. This experiment is shown in Fig. 11, with log10 scale for the y-axis.

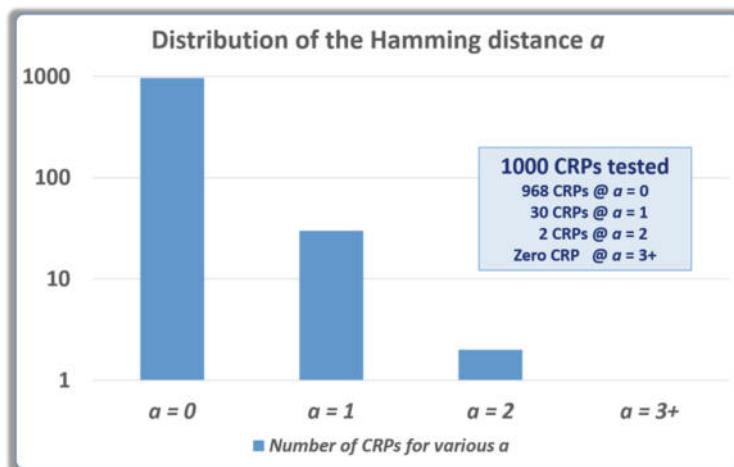


Fig. 11. Experimental evaluation of the Hamming distance a between one thousand 256-bit challenges and 256-bit responses, only non-fuzzy cells of the SRAM.

Based on the modeling shown in Fig. 7, 96.8% of the RBC matching operations take 50 ns for the server, 3% take 10 μ s, and 0.2% take 1 ms. Assuming that the RBC matching algorithm test Hamming distances up to $a = 2$, and reject Hamming distances greater than $a = 3$, the average latencies are in the 2 μ s range.

False Rejection Rates The probability to find one CRP error in a given cell is 0.01%.

The probability to find one CRP error in a 256-bit stream is 2.56%, the probability to find three CRP errors is approximately 16.7×10^{-6} . In the matching algorithm described in this section, if there is no match after trying all streams with a lower than 3, the authentication is considered as negative. Thereby the estimated FRR is equal to 17 part per million, which is acceptable in most implementations.

4 Generalization and Future Work

The authentication of the clients through RBC protocols, as described in Sects. 2 and 3, is initiated by the encryption of the message *Maj1* with the key generated from the responses. The cipher is in fact acting as a feedback loop between the client devices and the server. In this section, this feedback loop is replaced by the public keys, which are generated through asymmetrical schemes from private keys that are extracted from PUF responses. Using the RBC protocol described in Sects. 2 and 3, the server independently generates the same public/private key pair from the initial responses/challenges, and validates the private keys.

4.1 Securing Blockchains

Blockchain technology, in conjunction with SHA-2 hashing function, and digital signature (DSA), for example elliptic curve cryptography (ECC) [33–38], has the potential to secure IoT infrastructure and protect the data flow needed to track transactions for applications such as cryptocurrencies, strategic manufacturing, finance, and commerce. The underlying assumption is that the entire infrastructure of IoT is homogeneous, with each node protected by a crypto-processor handling the hashing and having secure non-volatile memory to store the cryptographic keys.

Malicious side channel attacks as well as physical hijacking of the IoT nodes can expose the private keys, thereby compromising the security of the infrastructure. The secure distribution of public-private key pairs in such an environment can be risky. Without reliable protection of the private keys, the digital signature of the blockchains is vulnerable, and the technology loses its value. We propose to address this vulnerability through the addition of RBC to protect the DSA of the blockchain ecosystem [39], as shown in Fig. 12.

At each transaction, the client device j openly communicates the new blockchain containing the description of the transaction, for example *Block4j*. The client j hashes *Block4j* to generate the message digest $H(B4j)$, encrypts both the blockchain and the message digest with DSA and the private keys, which are generated from the PUF responses, *responses4j*. The client communicates the resulting cipher $D(Block4j-H(B4j))$, and the public key *pub.key4j* generated from the private key.

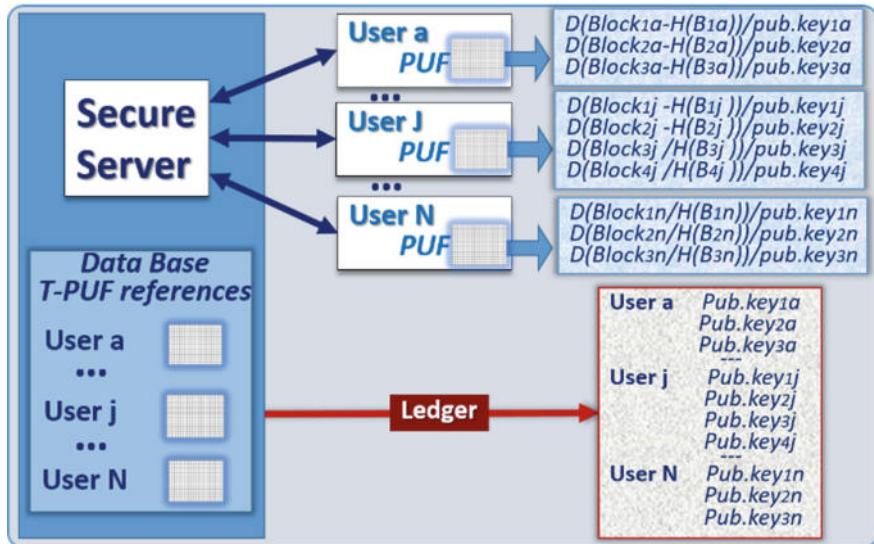


Fig. 12. Block diagram of a blockchain technology protected by RBC. The client devices generate the private keys from the PUF responses, and the public keys, ex. with ECC. The server independently generate a ledger, which validate the public keys.

By using the RBC protocol of Sects. 2 and 3, the server independently extracts the private key from the challenges, **Challenges4j**, and the associated public key. All streams with short Hamming distances from **Challenges4j** are considered, as described in Sect. 2.1. The server keeps a ledger with all the valid public keys. A third party can query this ledger to confirm that the public keys **pub.key4j** communicated by the client **j** related to the blockchain **Block4j**, is valid. Any third party can decrypt the cipher **D** (**Block4j-H(B4j)**), and confirms that the message digests of the blockchains match.

The use of ternary PUFs in this protocol is shown in Fig. 13. In addition to sending a random number, **RN4j**, to find an address within the PUF, the server also sends a mask, **Mask4j**. This allows the client device to skip the fuzzy states, and generate the private key, which results in smaller Hamming distances between the challenges generated by the server and the un-corrected responses.

4.2 Peer to Peer Communication

The scheme described above Sect. 4.1, is in fact a way to distribute, and validate public/private key pairs for PKI networks having a constellation of PUFs. The server periodically resets the scheme, sends handshakes to find new addresses in the PUF arrays of the client devices, with the objective to refresh the public/private key pairs as often as needed. For verification, the server makes the public keys openly available on a public ledger. The scheme use asymmetrical cryptography for peer-to-peer

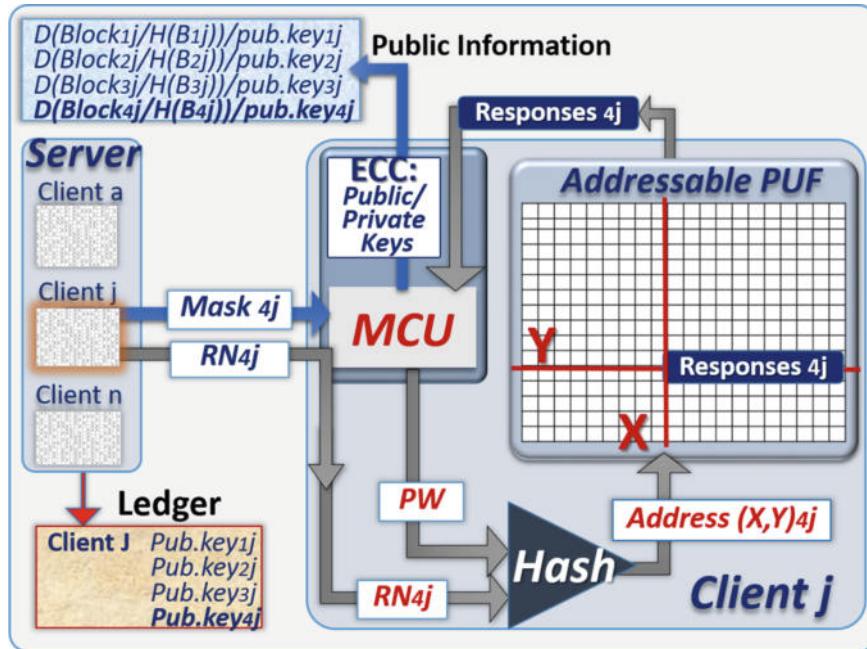


Fig. 13. Blockdiagram of the RBC with ternary PUFs to secure blockchains. As part of the handshake, the server generate a mask to blank the fuzzy cells.

communication as it is done with existing PKIs; the central certificate authority, in this case the server with RBC capability, manages the distribution of the public-private key pairs.

4.3 Impact of RBC on Mobile and IoT Networks

Mobile and IoT networks are vulnerable to several threats that include remote software attacks, as well as physical access vulnerabilities [40]. The RBC protocol, in conjunction with ternary PUF's, presents a hardware based solution to secure IoT devices. One benefit to be highlighted of the work discussed above is that no keys are stored on client devices. This is in contrast to current methods that store keys in flash memory which is vulnerable to a range of physical attacks [41].

Rather, the protocol uses the ternary PUF to generate a key on the server side and client side, independently, each time an authentication and communication event is required. The key is then discarded after use. In addition, the absence of a helper and error correction reduces the size, weight and power computing burden on the client side. This makes RBC accessible to a wider range of resource constrained IoT devices, while also mitigating vulnerabilities associated with such error management systems.

5 Conclusion

The main objective of this research work was to study the possibility to use PUFs for cryptographic key generation without error correction schemes at the client side of the network, therefore to eliminate the burden associated with the transmission of helpers, and the power hungry algorithms needed to correct PUF responses with these helpers. The elimination of error correcting schemes has the potential to significantly decrease the vulnerabilities to side channel attacks such as differential power analysis, which exploit the computing power needed during error correction, and fuzzy extractors.

This work demonstrates, through modeling and experimenting with SRAM PUFs, that uncorrected Response-Based Cryptographic (RBC) methods, in conjunction with our previous work on ternary PUFs, are showing promising parameters:

- The matching algorithm allowing the server to find the response generated by a client device, has typical latencies in the order of microseconds. Both the server and the client device can thereby find independently the same public-private key pairs needed for encryption.
- The False Rejection Rates (FRR) of valid client devices with the matching algorithm are only in the 17 part per million rate. In case of a reject, a new handshake can be initiated to repeat the two-way authentication process of the server, and the client device.
- The enrollment of the SRAM PUF needed 100 cycles power-off-power-on, adding up to a few minutes with our set up. Such an enrollment is only needed once, therefore such process is acceptable for many applications. We are currently working to develop stronger PUF technologies having faster enrollment that mitigate most side-channel analysis.

The RBC methodology benefits from the ternary cryptography, which include the following advantages:

- Using the masking technology, the scheme mitigates man-in-the-middle attack. A third party cannot initiate an effective handshake to generate a valid public-private key pair without knowing the position of the cells with ternary states;
- When lost to the enemy, the ternary PUFs are more difficult to read than binary PUFs. However SRAM based PUFs are rather weak, and need to be replaced by stronger PUFs;
- The generation of one-time public-private pairs generated at each handshake can reduce the vulnerability to side analysis; a potential exposure of the older pairs conveys minimum risks;

This work demonstrated that a client device with RBC can directly generate useable cryptographic keys (0 bit mismatch) for use in encryption and authentication with a server. The applications that can benefit from RBC are: access control, public key infrastructure, digital signatures, one-time pads, and blockchain technology.

We are currently developing matching algorithms that are more powerful, and will accelerate the scheme when the quality of the PUFs is lower. We are also designing stronger ternary PUFs that are more performance than the SRAM PUFs. Finally, these methods can be combined with an error correction schemes to expand the field of use.

Acknowledgements. The authors are thanking the students and faculty from Northern Arizona University, in particular Vince Rodriguez, Brandon Dunn, Julie Heynssens, and Ian Burke. We are also thanking the professionals of the Air Force Research lab of Rome, NY, and Alion science and Technology, who supported this effort.

Disclaimer (a) Contractor acknowledges Government's support in the publication of this paper. This material is based upon work funded by the Information Directorate, under AFRL Contract No. FA8075-14-D-0014-0018. (b) Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of AFRL.

References

1. Cambou, B., Orlowski, M.: Design of PUFs with ReRAM and Ternary States. CISR (2016)
2. Cambou, B., Afghah, F.: Physically unclonable functions with multi-states and machine learning. In: 14th International Workshop on CryptArchi, France (2016)
3. Pappu, R., Recht, B., Taylor, J., Gershenfeld, N.: Physical one-way functions. *Science* **297** (5589), 2026–2030 (2002)
4. Gassend, B., et al.: Silicon physical randomness. In: Proceedings of the 9th ACM Conference on Computer and Communications Security, pp. 148–160, CCS (2002)
5. Gao, Y., et al.: Emerging Physical Unclonable Functions with Nanotechnologies. IEEE, <https://doi.org/10.1109/access.2015.2503432>
6. Herder, C., Yu, M., Koushanfar, F.: Physical unclonable functions and applications: a tutorial. *Proc. IEEE* **102**(8), 1126–1141 (2014)
7. Maes, R., Verbauwheide, I.: Physically unclonable functions: a study on the state of the art and future research directions. In: Towards Hardware-Intrinsic Security (2010)
8. Jin, Y.: Introduction to hardware security. *Electronics* **4**, 763–784 (2015). <https://doi.org/10.3390/electronics4040763>
9. Delavor, M., et al.: PUF Based Solution for Secure Communication in Advanced Metering Infrastructure. ACR Publication (2014)
10. Guajardo, J., Sandeep, S.K., Geert, J.S., Pim, T.: PUFs and PublicKey Crypto for FPGA IP Protection. Field Programmable
11. Plusquellec, J., et al.: Systems and Methods for Generating PUF's from Non-Volatile Cells; WO20151056887A1 (2015)
12. Holcomb, D.E., Burleson, W.P., Fu, K.: Power-up SRAM state as an identifying fingerprint and source of TRN. *IEEE Trans. Comp.* **57**(11) (2008)
13. Maes, R., Tuyls, P., Verbauwheide, I.: A soft decision helper data algorithm for SRAM PUFs. In: 2009 IEEE international symposium on information theory (2009)
14. Christensen, T.A., Sheets II, J.E.: Implementing PUF utilizing EDRAM memory cell capacitance variation. Patent No.: US 8,300,450 B2; 30 Oct 2012

15. Prabhu, P., Akel, A., Grupp, L.M., Yu, W-K S., Suh, G. E., Kan, E., Swanson, S.: Extracting device fingerprints from flash memory by exploiting physical variations. In: 4th International Conference on Trust and Trustworthy Computing, June 2011
16. Chen, A.: Comprehensive Assessment of RRAM-based PUF for Hardware Security Applications. 978-1-4673-9894-7/15/IEDM IEEE (2015)
17. Cambou, B., Afghah, F., Sonderegger, D., Taggart, J., Barnaby, H., Kozicki, M.: Ag conductive bridge RAMs for physical unclonable functions. In: 2017 IEEE International Symposium on Hardware Oriented Security and Trust (HOST), McLean (2017)
18. Zhu, X., Millendorf, S., Guo, X., Jacobson, D.M., Lee, K., Kang, S.H., Nowak, M.M., Fazla, D.: PUFs Based on Resistivity of MRAM B. Cambou; Physically Unclonable Function Based Password Generation Scheme; NAU case D2016-011; Sept 2016
19. Vatajelu, E.I., Di Natale, G., Barbareschi, M., Torres, L., Indaco, M., Prinetto, P.: STT-MRAM-based PUF architecture exploiting magnetic tunnel junction fabrication-induced variability. ACM Trans. (2015)
20. Korenda, A., Afghah, F., Cambou, B.: A secret key generation scheme for internet of things using ternary-states ReRAM-based physical unclonable functions. In: Submitted to International Wireless Communications and Mobile Computing Conference (IWCMC 2018)
21. Taniguchi, M., Shiozaki, M., Kubo, H., Fujino, T.: A stable key generation from PUF responses with a fuzzy extractor for cryptographic authentications. In: 2013 IEEE 2nd Global Conference on Consumer Electronics (GCCE), Tokyo (2013)
22. Delvaux, J., Gu, D., Schellekens, D., Verbauwheide, I.: Helper data algorithms for PUF-based key generation: overview and analysis. IEEE Trans. Comput. Aid. Des. Int. Circuits Syst. **34**(6), 889–902 (2015)
23. Boehm, H.M.: Error Correction Coding for Physical Unclonable Functions: In Austrochip. Workshop in Microelectronics (2010)
24. Hiller, M., Weiner, M., Rodrigues, L., Birkner, M., Sigl, G.: Breaking through Fixed PUF block limitations with differential sequence coding and convolutional codes. In: TrustED'13 (2013)
25. Kang, H., Hori, Y., Katashita, T., Hagiwara, M., Iwamura, K.: Cryptographie key generation from PUF data using efficient fuzzy extractors. In: 16th International Conference on Advanced Communication Technology, Pyeongchang (2014)
26. Chen, T.I.B., Willems, F.M., Maes, R., Sluis, E.v.d., Selimis, G.: A Robust SRAM-PUF Key Generation Scheme Based on Polar Codes. [arXiv:1701.07320 \[cs.IT\]](https://arxiv.org/abs/1701.07320) (2017)
27. Rahman, M.T., Rahman, F., Forte, D., Tehrani poor, M.: An aging-resistant RO-PUF for reliable key generation. IEEE Trans. Emerg. Top. Comput. **4**(3) (2016)
28. Becker, G.T., Wild, A., Güneysu, T.: Security analysis of index-based syndrome coding for PUF-based key generation. In 2015 IEEE International Symposium on Hardware Oriented Security and Trust (HOST), Washington, DC (2015)
29. Cambou, B., Philabaum, C., Duane Booher, D.: Response-based Cryptography with PUFs. NAU case D2018-049 (2018)
30. Cambou, B.: Physically Unlonable Function Generating Systems and Related Methods. US patent disclosure No: 62/204912 (2015)
31. Cambou, B., Flikkema, P., Palmer, J., Telesca, D., Philabaum, C.: Can Ternary Computing Improve Information Assurance?; Cryptography, MDPI, (2018)
32. Cambou, B., Telesca, D.: Ternary Computing to Strengthen Information Assurance. Development of Ternary State based public key exchange. In: IEEE Computer conference on London (2018)
33. Croman, K., Decker, C., Eyal, I., E. Gencer, A., Juels, A., Kosba, A., Miller, A.: On scaling decentralized blockchains. In: Springer International Conference on Financial Cryptography and Data Security, Berlin, Heidelberg (2016)

34. Luu, L., Narayanan, V., Zheng, C., Baweja, K., Gilbert, S., Saxena, P.: A secure sharing protocol for open blockchains. In: ACM SIGSAC Conference on Computer and Communications Security (2016)
35. Eyal, I., Gencer, A.E., Sirer, E.G., Renesse, R.V.: Bitcoin-NG: A Scalable Blockchain Protocol. In NSDI (2016)
36. Dorri, A., Kanhere, S.S., Jurdak, R.: Blockchain in internet of things: challenges and solutions. arXiv preprint arXiv: 1608.05187 (2016)
37. Gervais, A., Karame, G.O., Wüst, K., Glykantzis, V., Ritzdorf, H., Capkun, S.: On the security and performance of proof of work blockchains. In: ACM SIGSAC Conference on Computer and Communications Security (2016)
38. Zheng, Z., Xie, S., Dai, H.-N., Wang, H.: Blockchain challenges and opportunities: A survey. Int. J. Web Grid Serv. 1–25 (2016)
39. Cambou, B.: Digital signature for blockchains with ternary PUFs; invention disclosure. NAU Case **D2018**, 047 (2018)
40. Kamara, S., Fahmy, S., Schultz, E., Kerschbaum, F., Frantzen, M.: Analysis of Vulnerabilities in the Internet Firewall. Comput. Secur. **22**(3) (2003)
41. Prabhu, P., et al.: Extracting device fingerprints from flash memory by exploiting physical variations. In: 4th International Conference on Trust and Trustworthy Computing (2011)



Implementation of Insider Threat Detection System Using Honeypot Based Sensors and Threat Analytics

Muhammad Mudassar Yamin^{1(✉)}, Basel Katt¹, Kashif Sattar²,
and Maaz Bin Ahmad³

¹ Norwegian University of Science and Technology, Gjovik, Norway
 {muhammad.m.yamin, basel.katt}@ntnu.no

² University of Arid Agriculture Rawalpindi, Rawalpindi, Pakistan
kashif@uaar.edu.pk

³ PAF KIET Karachi, Karachi, Pakistan
Maaz@pafkiet.edu.pk

Abstract. An organization is a combination of vision, technology and employees. The wellbeing of organization is directly associated with the honesty of its workers. However, an organization is also threatened by misuse of information from its agents like former employees, current employees, vendors or business associates. These kinds of threats which are posed from within the organization are known as Insider Threats. Many approaches have been employed to detect the Insider Threats in organizations. One of such approaches is to monitor the system functions to detect possible insiders. These approaches raise unnecessary amount of false positive alarm which is then taken care of with the use of evolutionary algorithms. The solution to this Insider Threat detection requires a lot of configuration before implementation in real world scenarios due to different threshold values in different organizations. Insider Threat detection can be done by means of honeypots sensors in a limited and in satisfactory way. The present research proposes a new technique for detecting insiders using encrypted honeypots. This technique complements the existing insider detection systems and improves its performance in terms of decreasing false positive results.

Keywords: Insider threat · System monitoring · Activity detection · Honeypots · Threat analytics

1 Introduction

In this section we discuss about the background study of Insider Threat. Kevin Mitnik once stated, “Companies spend millions of dollars on firewalls, encryption and secure access devices, and its money wasted, because none of these measures address the weakest link in the security chain”. That quote completely gives a good scenario of tension faced by network security professionals in the modern days. The weakest links in the security chain are the individuals who work in secure networks with authorization. They can produce a danger to company if they intentionally or unintentionally

damage networks infrastructure or steal important information. As described by Mallah [1] these kinds of individuals are called insider. Both type of the attacker can damage the vulnerabilities of the system. Therefore, it is very important to implement an efficient and effective security policy for the removal of such threats.

According to Moore [2] the authorized users or employees have access to the confidential data and to the sensitive assets of the company, so there is always a risk that the employees may misuse this data access for any mischievous purpose. An example of insider case is Chelsea Manning who was responsible for the leaking more than 60,000 U.S department of defense documents on WikiLeaks and Edward Snowden who exposed secret NSA documents in public. These two cases are important examples of Insider Threat incidents. In these cases, detection of insiders was very important because of the data breach in the National Security issues of United States and people lives were also endangered. the frequency of Inside attacks during the same years were 29, 20 and 32% respectively.

The Computer Emergency Response Team (CERT) in USA actively researches on Insider Threat and release surveys from time to time regarding Insider Threat detection. The recent surveys published in 2006 by CERT/CC's produced following findings regarding inside and outside threats. The findings revealed that the number of Outside attacks during year 2004, 2005 and 2006 were 71, 80 and 68% respectively.

Early research about the Insider Threat is conducted by Hayden [3] which suggested that it is the need of the hour that the continuous check on the employee activities must be maintained in the company to avoid any discrepancy. The continuous profiling of the user activities help to locate any Insider Threat well before time and the administration can easily avoid, detect and recover this Insider Threat timely. The company should maintain a check on the detailed behavior of its employees to detect the threatening activities of the unauthorized users. The company should study the profile of the trusted users in the light of the organization's policy. After the investigation, if any user is found to be involved in any suspicious activity which is against the policies of the organization, then the user who possesses the suspicious profile is marked as suspicious and onward the more careful watch should be maintained over that user.

The problem is that the current techniques of detecting Insider Threat got many imitations. The limitations are related to the profile checking of the employees because the Insider Threat detection technique developed by researchers [4], can give rise to many false positive alarms during the detection process, the production of false positive alarms is responsible for the unnecessary system lock down of the company. The technique of detecting the Insider Threat gives rise to false positive alarms for the company and the frequency of these false alarms can be cut down by eliminating the human error factor. The human error is thus responsible for the large amount of false positive alarms in the company. It can be explained as follow. The employee checking the main server of the company may get confuse about the action of another employee on the system and he could assume the simple file access log as a bad intention of the employee for the company. He may raise a false positive alarm of Insider Threat detection against that employee. He may not be necessarily right every time. Therefore, it is very important that a mechanism should be developed which would be very

accurate in the raising of real Insider Threat detection alarm and detect only the genuine, real and imminent threat and not raise any false positive alarm.

1.1 Insider Threat

Any company or organization has the resources to locate or control the situation when an outsider (non-representative) tries to steal the organization data physically or electronically. Their rivals constantly threaten the organization that they might steal their data illegally. In most of these cases, when a large amount of sensitive data is stolen from the company, then it is very difficult to find out the culprit because in the majority of cases, the criminal lies within the victim organization. The culprit is most of the times the insider- a specialist who has easy access to all the company's data. That insider could be dangerous because he has access to all the company's assets and secrets. That insider may do this crime for his own satisfaction or he may also be a "spy" of any rival company doing this task to get monetary benefits. The insider may steal the company's data and items to sale this information to another enterprise to strengthen their credibility.

In majority of companies, the employees are given accounts to get easy access to company's data. In case of Insider Threat detection, these accounts are very useful to detect any data breach in the company system. The Insiders (infiltrated employees) got their accounts in the respective company, which allows them to use the PC frameworks of the company without any problem.

Insiders made use of the accounts given to them by the company which allow them legally to use the PC frameworks of the company. In the previous years, this permission allowed the insiders to conduct the execution of these commitments; these permissions could be misused by insiders to harm the business of the company. Insiders have a capacity to get knowledge of the company's assets and they also possess licensed innovation of the systems that are designed to secure company's data. So these access codes make it less challenging and easier for the insider to roam through any security control check which they know. The concrete proximity to the company's data reflects that the insider doesn't need to hack into the hierarchical system through the outside border by navigating firewalls; like which is done in case of buildings, mostly with quick access to the association's inside system. Insider Threats are very difficult to execute, and it needs a lot of efforts on behalf of insider, this statement is given on the basis that the insider has an authorized access to the company's data and resources. The main purpose of insider is to get access to the company's data without the knowledge of the relevant authorities. The insiders then misuse this information to gain their own benefits.

The damage which is caused by this Insider Threat can be of various types. This can be understood with the help of examples of burglary, robbery or extortion. The insider breach in the company's data by introducing a virus, worm or Trojan in the system is similar to the burglary of money from bank. Like the money is stolen from bank with the effort of robber, same is the case of company's data which can be stolen from the data center by the effort of insider. The danger of Insider Threat can be encountered by taking appropriate security measures. These security measures include implementation of ethical policy for Internet surfers, use of different spy product

examining programs, hostile to infection projects, firewalls, and a strong information security check and chronicling administration.

1.2 The Insider

There are a number of different associations like CERT and RAND which perform analysis of Insider Threat and the issues related to the Insider Threat detection [5]. The significance of the term insider has to be dealt in detail for a clear picture now. Insiders are a group of people in a particular company which are involved in any illegal activity in the company and are mostly followed by digital security group and especially by Insider Threat researchers. Many of the researchers suggest a dual way of dealing with the identification of Insider Threat, recommending that the insider may come under the mark of a quantifiable parameter, for example, and an occupation classification. The assumptions which is based on this methodology in this case, resulting out to be somewhat vague and doubtful with the expanding utilization of outsourcing, versatile figuring, contractual workers, and business associations. In the present research; the researchers have made use of a grid approach to describe and characterize the insider in light of access [6]. The researchers have nominated and described an insider as a trusted entity that has given the potential to abuse a security center. The insider is detected and resolved in relation to a set up security arrangement. Insider Threat lies in the entrance and capacity to:

1. Violate a security approach utilizing honest to goodness access, or
2. Violate an entrance control approach by getting unapproved access.

1.3 The Insider Problem

The complexities which are associated with Insider Threat detection are more complex than the written theory. It is much more than management of outer digital threat with reference to objective demographic. The problem arises mostly because of the matters of managing trust, security, and morals, which the researcher investigates later. The Insider Problem can be summed up as the test of protecting an association from interior digital attacks. The problem is that most of the companies do not give this matter due importance and take it for granted. The companies think that the issue of Insider Threat detection is easy to handle but they don't know that this matter can prove to be really dangerous for their company in future.

1.4 Ethics and Trust

One of the important challenges about the insider issue is trust and confidentiality. Whenever any person is doubted to be an insider, there is first some level of trust associated with that specific person. The provision of this trust means that the respective insider is a trusted employee of the company and got access to PC account, access to a confined room and surely possesses access to a framework asset and delicate data of the company as well. The point to be noted over here is that when an

individual is allowed such type of trust in a company, then that person automatically gets the authority to misuse this trust for his own benefits.

1.5 Cyber Attacks

There are three forms by which digital attacks could be performed. These forms are: exfiltration, information defilement, and refusal of administration. These digital attacks can cause a huge amount of destruction to the company. The damages may include harmed believability, uncovered helplessness, and budgetary misfortunes. In order to prevent such digital attacks, first it is very to understand what is a digital attack, how it works and what are the different types of digital attack. After getting knowledge about the digital attack, the investigator can then have a look at real life insider attack cases and get to know that how the attack occurred and what happened on the utilized frameworks utilized during the attack.

1.6 Exfiltration

An exfiltration attack consists of information robbery, IP burglary, or any intentional expulsion of information. This attack is the result of grouping up of insiders with the people of enemy contender groups of the company [7]. The insiders usually hand over touchy or private data to these enemies for financial benefits.

The most infamous and dreadful example of exfiltration in the history is that of a professional FBI operator Robert Hanssen who worked in the secret services of USA. Hanssen worked in the FBI as an agent but actually he was a spy of Soviet Union. He was an insider in FBI. Working in the disguise of FBI agent, Hanssen downloaded a large amount of classified data to encoded information by setting up the equipment and a portable workstation. He further used the portable PC to speak with Soviet insight officers for the exchange of the data that he got. This attack led the Hanssen expose a large amount of extraordinary classified information of USA to Soviet Union. With a specific end goal of acquiring access the data, Hansen got to the FBI computerized records framework. Hanssen hacked the data center of FBI and extracted all the required information. Hanssen constantly put demand the information from the system which was not of his concern. He was an approved client of the database of FBI, therefore his intentions were never doubted. Therefore, Hanssen conducted the attack on the data of organization without raising much suspicion. The conclusion is if FBI has put ample check on the suspicious activities of its employees, then this data breach is not possible to occur.

Another similar case of exfiltration happened in CIA. The culprit was a 16 years old employee of CIA Harold Nicholson [5]. He was employed in CIA as a teacher in a CIA exceptional preparing focus from 1994 to 1996. During his stay at the CIA, Harold transferred a huge amount of classified data to Russia. Harold get hold of the data by hacking into the CIA's venture PC framework and by conducting a vast kind of inquiries on the CIA's databases. These inquiries were out of range of his discretion in the CIA and consisted of information regarding US insight information on Chechnya, which he sold to Russian authorities.

The CIA then took help from the FBI to investigate this matter and Harold also passed through a standard polygraph test successfully. They found out that Harold transported a big number of Top-Secret documents from CIA PCS to Russia onto scrambled plates, movies of film, and also by his PC hard drive. This huge amount of data transfer from the CIA could have raised suspicion if the data traffic on the CIA framework have been continuously monitored. However, due to the absence of any security check, the fact remained unknown that when and how Nicholson hacked the framework of CIA and leaked the classified information.

1.7 Data Corruption

Another type of Insider Threat is that of data corruption. An information defilement attack is described to be as a strange change in the real data or information of the system. It consists of modification and removal of data from the system. This attack is basically meant for digital extortion or in attempt to harm information of the system. When data is changed in a system, then a lot of problems could be produced for the company.

An example of case of data corruption can be studied with reference to programming engineer Chris Harn [5]. Harn was the supervisor of PC system and sever checking at Autotone Systems. The purpose of these servers were to digitalize the daily salary of workers. Harn manipulated his power and authority by changing the data at the framework of Autotone Systems of super-client benefit program and changed the salaries of the worker colleagues for more than 3\$ of the illegal profit of the company. By changing the salaries of the workers, Harn changed the data of company review detail to cover up his corruption of money for the increased salaries of the employees. Harn's access to framework of company allowed him to do this task of data corruption in the system. The bottom line of this discussion is that a huge amount of data corruption was being executed in the database of the company. Eventually, the discrepancy between the wages and salary reviews were balanced in the Autotone's framework. If any strange activity in the database of company have been checked regularly, then we can avoid such incidents.

1.8 Denial of Service

A ridiculous attempt to make a PC information inaccessible to its planned clients is called administration attack. This attack is carried out by bringing changes in the structures of company data, for example, these changes can be produced by the help of malignant code, malware establishment, over-burdening a support, or by another activity that results in damaging a company's data.

Some examples of administration attack/assaults however are significant in the sense that they are not particularly acknowledged or executed on the spot. These attacks are characterized by the attacker induced malware which causes all the data infiltration after a certain period of time. The effect of this induced malware may take minutes, days, months, or even years to show full execution. These attacks show their effect when they are "exploded" or triggered by an occasion or a point in time. This is the reason these attacks are called 'bombs'. These "bombs" can appear in two shapes.

The first of these is a rationale bomb, which is triggered by a framework occasion. Rationale bombs lay dormant and don't execute the effect until a specific framework occasion happens. Unlike rationale bomb the second type of 'bomb', is called a period bomb, which is triggered by a particular minute in time. These bomb attacks are not activated by an occasion. They lay dormant silently until a timeframe is set by the insider.

1.9 Abnormal Activity

A lot of cases of insider attacks like exfiltration, information debasement and foreswearing of administration assaults happen on daily basis in different companies and cause a lot of damage and monetary losses. The implementation of Insider Threat detection procedure in the company can check the irregular activities of insiders/attackers in the company's framework. The insider before initiating his attack can take first step by starting the unusual activities on the system. An efficient Insider Threat detector takes notice of these abnormal and unusual activities of the insider and pin point them in the beginning. An example of this unusual conduct is Hansen's unusual hunt of information, Nicholson's extraordinary huge information exchanges, Harn's information and review record alteration, Cooley's late night access and erasure of \$2.5 million worth of engineering drawings, and Shae's cancellation of review information. These all examples reflect examples of irregular activities that happened during or before the insider attack. A framework that made regular security checks of the company's employees accounts could have distinguished these variations from the normal and identified the attackers. This reason along with the above cases and numerous more case of genuine insider attacks are the inspiration for this present research.

The parameters associated with Insider Threat detection is more complex than external cyber threats. The employee trust level and intention make the situation very complicated, so simple profiling of employees can easily be over looked by employee intention in case of an incident. Which make it very difficult for accurate detection of Insider Threat which raises a lot of false alarms. These false alarms also create mistrust among employee and employer. Researchers are able to reduce the amount of false alarms using evolutionary algorithms. But the problem associated with employee intention remains unfixed. To fix the issue researcher proposes a new technique for detecting Insider Threat using encrypted honey pots. This complements existing Insider Threat detection system and improves its performance.

Following solutions to Insider Threat detection problem is contributed during the research:

1. Classification of Insider Threat based upon attack scenario. The Insider Threat is classified based upon the type of attack the insider can perform in the organization. A seven-factor kernel density estimation is developed for the classification.
2. Development of encrypted honey pots for detection of Insider and deployed on a working network. The honey pot comprises of an actual system which forward all system and network calls to the data stream analyzer.

3. Development of real time data stream analyzer to identify the threat posed by the insider. The real data stream analyzer collects the data from the honeypot sensors and identifies any Insider Threat based upon system and network calls.
4. Reduced the system configuration requirement as compared to the previous system by introducing honeypot sensor which can be any system or virtual machine in a network.

1.10 Layout of Paper

In Sect. 2 we discuss the existing research work related to insider detection system. Section 3 describes the proposed framework and insider detection mechanism for the factor affecting the organizational security. In Sect. 4, we discuss the results achieved by proposed classification and finally we conclude the paper with the summary of research work in Sect. 5.

2 Literature Review

This section consists of complete overview of related literature which is very useful for the evaluation of insider detection system. After giving the complete background and structure of insider detection system, we propose an improved and new framework for the assessment of Insider Threat detection. We discussed the literature related to the usefulness of Insider Threat detection, the literature related to different sensor with virtual machine, the literature with reference to sensors and Honeypot along with the activity of behavior based activity recognition and the effect of false alarm on Insider Threat detection system respectively.

2.1 Perspective Study on Insider Threat Detection System

Mckinney [8] established a by default way of differentiating imposter and insiders in a company by observing the methods of their asset usage. This approach was implemented by creating an “ordinary” profile with respect to the run of the mill client conduct. The setting up of these specific profiles was done by the calculations of The Naïve Bayes machine learning. The performance that divert away from the ordinary profiles are considered as extraordinary and possibly suspicious.

So basically, an irregular behavior of profile observed in the machine raise suspicion against that particular profile. This procedure displayed marvelous results in the companies after their implementation with an exact positive ration of 96.7% and the frequency of false positive alarm was found out to be only 0.4%. This information was collected from the company during the duration of three weeks.

Qiao [9] devised a system of Insider Threat detection by the use of battle Insider Threat by actualizing Subject-Verb-Object (SVO) screens. The working of this method is that the procedure is based on the screening out of clients and procedures, catching data, for example, login times, client name, benefit levels, process IDs, and security levels in the framework of the company.

Basically, this method is very effective in pin pointing the possible suspect of Insider Threat in the company. The method enables the server administrator to log into access sorts, for example, peruses, composes, executes, and the “Item” screen recorded document characteristics, for example, proprietor, size, way, and sort. The information that’s is collected after these operations is then subjected to thorough investigated from a client metadata archive that contains client information, for example, record framework I/O and procedure movement. The threat which may be present are reviewed with the help of the stride/dread threat model which includes spoofing, tampering, repudiation, records disclosure, denial of carrier, escalation of privilege, damage potential, reproducibility, exploitability, affected users, and discoverability. The results which are achieved after the implementation of this methodology can be related to the results of the instrument that is responsible for the gathering of information about safety measures in an isolated place. However, no results were given that give approximation and direct input about the performance of general framework of the company.

Shavlik devised a hidden recognition framework for the detection of information that has been searched on the framework of the company. The detailed information checks include occasion logs, application information, and client behavioral data, for example, the quantity of running projects and writing rate [10]. The Winnow-based calculation was employed for detection of abnormality discovery it produced and preferred recognition rates over the Naïve Bayes calculation. The working of framework is that it recognizes any suspicious activity happening on the framework of the company. This method identifies any strange activity in the company’s system by running the logged action of one client through the profile of another client. The recognition rates of company were found out to be approximately in the mid to upper 90th percentiles with rates as high as 97.4%.

2.2 Sensors with Virtual Machines

Spitzner [11] employed honeypots and honey tokens as a tool to fight exfiltration insider attacks. Honeypots are basically the PC frameworks in the company which are meant to observe the digital activities of the employees and trap and pinpoint any individual who tries to use the company’s data illegally. Honeypots are very helpful in detecting the Insider Threat in the company. Honeypots can instantly detect any irregular activity of employee in the main framework of company and report this irregularity on the spot.

Basically, the movement on a honeypot is thought to be dangerous, and is recorded. Another example of Insider Threat detector are Nectar tokens which are advanced information or data, such a record, a Mastercard number, or a document which are set up to catch any individuals who tries to illegally utilize the data of the company. Till now, no results have been presented regarding the efficiency of this process.

Yu and Chiueh [12] produced the Display-Only File Server (DOFS), it was basically a framework whose purpose was to provide remedy against data theft in 2004. Working as an independent document server, the DOFS studies that if the client had possessed the authority to manipulate a particular asset of the company. If it is found out that the client had the right the access the particular asset of the company then

the server drops the suspicious level against that client. In this way DOFS hinders people from changing or disturbing the original documents of the company.

Pramanik made use of a security approach that was based on the use of structure like Digital Right. Another methodology for the administration for Insider Threat [13] includes the use of an entrance control structure that requires the client to be checked and secured before the client gets access to peruse/compose/overhaul documents. The drawback is that this framework produces unpleasant effects, for example, it creates problem for clients to continuously open the same document they needed many times to study in a single day.

Park made variation in a Role Based Access Control framework to detect Insider Threat [14] by conducting Composite Role-Based Monitoring (CRBM). Taking into account the benefit of the clients, the following method allows or rejects access to records and assets in view of the client given task. The CRBM make use of threefold structure to characterize access benefits: Application, Operating System, and Organization. The access to assets is allowed or stopped according to the results of these three part structures. However the findings of these experiments did not prove to be really fruitful and did not support cases of an Insider Threat framework that can be used effectively to fight and stop any insider attacks more accurately in companies than the previous already implemented methods and procedures.

Symonenko stated a part based technique in which client parts, setting, and semantics are calculated for the detection of Insider Threat [15]. This method makes use of regular dialect handling to decide themes and territories of enthusiasm inside reports.

These factors are then studied in detail to look at respective client appointed themes and ranges of interest by taking into account the connection of a client participation. Bolster Vector Machines (SVMs) and ontologies are used to direct and observe and client themes and interests. The records were then gathered for the utilization of bunching. The limit of groups and the separation from different bunches are put in use to eliminate the threat.

Cathey made use of the bunching calculations with reference to archives, inquiry questions, and significance appraisals in order to understand any irregularity with reference to company's interest [16, 17]. In this method, a collection of records are arranged into groups. A client profile is then established with regard to the utilization of bunches. These report groups are then compared and contrasted with the reports in the client's profile. The significant information which should be known by the organization about the client's profile is narrated by ordinary client seeking activities and consistent terms inside the records are then forwarded to by the client. The threats are pointed out by finding out the abnormalities in a required report and in the client's profile as well. The data obtained after the application of this methodology showed that "abuse identification" rates ranges between 90 and 100% and the ratio of false positive alarm fall somewhere around 12 and 15%.

Aleman-Meza [18] has carried out research on information rupture which occurs inside the setting of Insider Threat. This research makes use of factual, NLP, and machine learning methods to calculate the significance of got to records regarding insider's work. Their focus point is in the national security and in the terrorism space. The data of methods and its drafting records are very relevant, steadily dictated, but in

terms of linguistic clarification, the data appears to be doubtful, insignificant and indecisive. This brief paper has also shown results that happen to be motivating with a test set of 1000 records.

Liu have shown a method for the detection of Insider Threat in an organization. In this method, the operator breaks down the framework calls and framework call data and traits [19]. They made use of the nearest-neighbor machine learning calculation to find out any abnormal activity in the light of framework calls. The tests proved this method to show a high frequency of false-positive rate for the purpose of Insider Threat detection. If we take into account these results as it is, this conclusion have shown intelligent results for interruption identification.

Anderson [20] suggested a technical engineering based method for the detection of Insider Threat by monitoring the documentation and program occasions. That engineering method consists of sensors and substance based recognition in the given setting. The sensors are basically entities that consist of screens conveyed inside applications (e.g. MS Word and Internet Explorer) and different other working framework. These sensors work on the principle that they thoroughly analyze occasions ad activities, for example, document opening, closing, recoveries, and console inputs. Once the activities have been found out, then these activities are forwarded by a material based-directing framework to the classification of element access control and investigation units. The parts of the entrance control can tally with the project and can detect and prevent the doubtful moves from being made.

2.3 Behavior Based Activity Recognition

Nguyen [21] designed a framework to differentiate suspicious insider activities from regular logins by putting a check on the framework calls which are directly associated with the document framework and procedure. This approach studies the calls initiated by the client and also by taking into account the framework calls started by the host framework. The tests have shown that the recorded framework calls begin by the clients reflected a good amount of variation and are not ideal for creating behavioral profiles. The record framework calls were then started again by the working framework, and they showed extraordinary expected performance, and was recommended by the administrators for client behavioral profiles. This research did not show the results about the changes that occurred in the client behavioral profiles. This is beneficial for this research because it gives idea of this concept with respect to the deal that if the desired performance is dependent upon client action or whether the working framework acts randomly, in case of clients as well. The fact is that the framework calls were supposed to generate marvelous results for showing client's performance. After the regular following of all the procedures, 92% of clients have shown a variation of rundown of documents and a reasonable number of them got to records in examples, and 92% of them had a settled rundown of formulated youngster forms. This procedure stressfully reflected all kind of flood attacks along with those in which the methods did not have an altered number of baby procedures.

Another behavior based insider detection model is presented by researchers [22]. The advantages of behavior-based approaches are that it very accurately and remotely finds surprising vulnerabilities which cannot be found without it. They even contribute

substantially to the (partially) automatic discovery of those new attacks on the systems. They are less passionate about operative system-specific mechanisms. They conjointly facilitate discover ‘abuse of privileges’ sorts of attempts that don’t really involve exploiting any security vulnerability. In short, this is often the paranoid approach: Everything that has not been seen historically in the past is dangerous.

Behavior-based intrusion detection techniques assume that associate intrusion are often detected by observant a deviation from traditional or expected behavior of the tactic or the users. The model of traditional or valid behavior is extracted from reference data collected by varied means that. The intrusion detection methodology later compares this model with this activity. Once a deviation is ascertained, associate alarm is generated. In different words, something that doesn’t correspond to a historically in the past learned behavior is believed regarding intrusive. Therefore, the intrusion detection methodology may be complete (i.e. all assaults got to be caught), however its accuracy may be a troublesome issue (i.e. you get many false alarms).

The warning rate being very high is typically cited because the main disadvantage of behavior-based techniques as a result of the complete scope of the behavior of associate data methodology might not be coated within the work of the educational part. Also, behavior alters over with time, introducing the need for periodic on-line preparation of the behavior profile, ensuring either in inaccessibility of the intrusion detection methodology or in further false alarms. The data methodology endure assaults at the same time the intrusion detection methodology is learning the behavior. As a result, the behavior profile contains intrusive behavior, that isn’t detected as abnormal. The behavior Driven hybrid model is represented in Fig. 1.

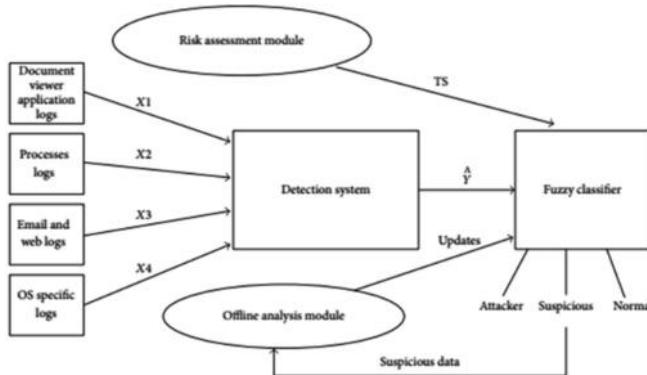


Fig. 1. Insider threat detection and prevention framework [4]

2.4 Effect of False Alarm on Insider Threat Detection System

Hybrid approaches uses two or more different techniques for detecting insiders, the models derived from hybrid approaches possess the characteristics of the derived models. An example of hybrid information security model can be seen in “Using Genetic Algorithm to Minimize False Alarms in Insider Threats Detection of

Information Misuse in Windows Environment," developed by researchers [4]. Which uses evolutionary algorithms with behavior driven models to reduce amount of false positive alarm raised by behavior driven information security models?

Hybrid approaches have many advantages over the previous models but one issue remains a problem. Every organization have a different thresh hold value for detecting the insider to compensate that the insider detection system requires a lot of configuration in a lot of different scenarios to get fruitful results. Without suitable configuration the performance of hybrid insider detection system cannot be satisfactory. This creates the need of system which require little or no configuration in different scenarios.

3 Materials and Methods

3.1 Background

The method used in this research work is stated below. The population of the present research is all the computer systems of TEST BED lab set-up. The sample consists of 50 PC systems of TEST BED which is a complex network of PCs servers in a computer lab of random organization. A huge number of employees, interns, and researchers at TEST BED agreed to be checked for a pre-determined period of time on experimental basis. Randomness was a must component for all participants. However, their part within TEST BED (employee, researcher, intern) was allocated specifically. All experiments were performed on the Windows 7, 8.1 and 10 operating system.

All participants have adequate knowledge that their file system, network, hardware, and process information would be checked and that data would be studied on the basis of the user and system activity on their respective computers. Although the participants got the idea that their digital activities are being checked, we enquired from them that if they would carry on their usual tasks on PC normally the way that they used to conduct if their behavior was not being checked. Issues and problems regarding the accuracy of this supposition are discussed in Chap. 4 along with other ethical and legal issues regarding the present work. The data was gathered form the participants over periods of time that consisted of a duration of few days to a few weeks depending upon the test.

3.2 Honeypot Sensor Setup

The honeypot comprises of an actual running window or linux which operating system which can be running on a physical machine or virtual machine. The honeypot is configured to forward logs to the data stream analyzer. In the Test Bed, a honeypot sensor is placed in the active directory of the network in which all systems are connected in the network used active directory for their system communication over the network so every activity on the system can be logged easily. The malicious system call are forwarded to a read data stream analyzer which is running on a separate system in which the data streams from various sensors is represented graphically based upon the threat level. The data streams contain system logs which helps to clearly identify which malicious activity is going on the system.

3.3 Data Acquisition

For Windows based environment, data collection was conducted by use of a monitoring tool based upon Microsofts.NET framework which is called IBM WinCollect. The function of this tool is to monitor and gather data depending upon the events and states of the operating system and also depends upon the changes in the hardware respectively. The data sets and variables are associated with Install Wincollect agent and configuration console in windows and configure it as shown in screenshot. To send logs on UDP, they create destination in UDP Section and vice versa.

First one has to select which type of events one would like to send like “Local System”, “Security”, “Application” etc.

Here Device Address is the machine IP address. One can Add destination, which is just formed, at the bottom of this window in destination box by clicking on “Add”.

For Linux based machines, rsyslog daemon was used for the acquisitions of logs, the configuration takes place in the following manner.

```
$ sudo nano /etc/rsyslog.conf
Add the following configuration at bottom
$template linuxbox,“<%pri%>%timestamp% 192.168.2.50%syslogtag%%msg%”
## IP of machine
# Use only one UDP or TCP depends on the importance of device
.*.* @192.168.2.50:514;linuxbox ## For UDP
.*.* @@192.168.2.50:514;linuxbox ## For TCP
$ sudo service rsyslog restart
```

3.3.1 Profile Training Phase

The experiment starts with the profile training phase. This profile training phase consists of gathering and processing of all the activities conducted by the participant, processes in hardware, processes in network and in the file system data to create a normal profile. The collected data is then processed with the help of k-means clustering and kernel density estimation algorithms. Each normal profile is represented by the seven kde distributions beneath:

Processor usage profile—kde possibility distribution relies upon the tactics of person data in the stage of profile education.

Memory utilization profile—kde possibility distribution relies upon the approaches of user information inside the level of memory schooling.

Hard drive utilization profile—kde possibility distribution is dependent upon the techniques of person difficult drive data inside the Section of profile training.

Process threads profile—kde probability distribution is depending on the frequency of energetic threads of each process during the level of profile education.

File machine profile—kde possibility distribution relies upon the data acquired from the activity of report device hobby accumulated inside the degree of profile schooling.

Network IP profile—kde opportunity distribution is dependent upon ip addresses and community connection statistics that are acquired in the level of profile education.

Community port profile—kde probability distribution is dependent on the records of network port usage which is obtained in the degree of community schooling.

3.3.2 Nominal Behavior Probability

A KDE can be described as a collection of probability distribution. In the view of preset research work, when the probability distributions are associated with the data of normal profile, then the end result is that the user behavior is normal. Thus the user behavior can be termed as nominal behavior probability.

In Fig. 2, we illustrated the participant process profile. It can be seen from the Figure that in the stage of profile training, the frequency of use of CPU normally fell between 15 and 40%. Following the creation of the frequency of use of processor falls between 15 and 40%. The value of this percentage shows comparatively huge probability of nominal behavior. The values of the use of processor outside the range of 15–40% result in comparatively less probabilities of nominal behavior.

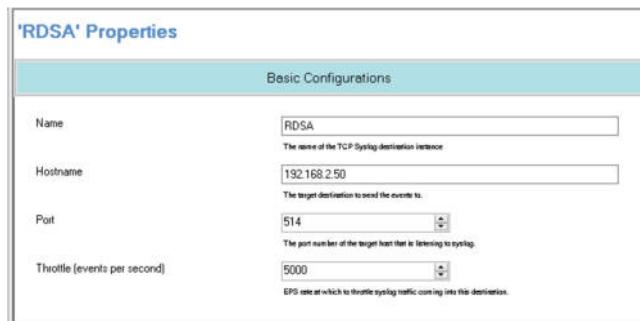


Fig. 2. Honey pot sensor configuration using IBM win get

3.4 Real Data Stream Analyzer

The real data stream analyzer is an Ubuntu based system which is placed on a separate machine. The real data stream analyzer performs the following functionalities.

3.4.1 Sensor Data Parser

This module parse incoming data streams from honeypot sensors in a readable format. The data generated from the honey pot sensor are in form of big Java Sting Object Notation strings which is very difficult to read. The module separates the strings in form of source IP, destination IP, event name, user name, Date, Time and many other important are factors represented in a readable form. After collecting all the relevant data of the event, the data is forwarded to the threat analyzer to identify potential Insider Threat. The threat analyzer is a separate module and is placed on the same server.

3.4.2 Threat Analyzer

The threat analyzer analyzes the parse data on three different levels. Each level is designed to filter out any chance of a false alarm. It only sends alert notification when it is unable to detect event activity an any level details of which is explained in the algorithm below:

Algorithm 1: False Positive Insider Threat Detection

- Step1: If user event activity is normal then no alarm
- Step2: Else check system running software and application events
- Step3: If user activity is according to software and application events then No Alarm
- Step4: Else check system hardware changes events
- Step5: if hardware changes are allowed to user then No Alarm
- Step6: Else check user privileges
- Step7: if user modifies files as per user privileges then No Alarm
- Step8: Else raise the alarm.

The algorithm is simple enough to be easily implementable and is effective enough to detect any false alarm generated by user system activity. The three level classification of user system activity first detect that if the event is software or system service generated which is previously identified in profile training phase then it ignores the event. Next if any hardware changes occur in the system which generated suspicious events should be analyzed and if the hardware changes events fall in normal user behavior generated events then it ignores the events. Finally it checks the changes in data made by the users. If the data modified by the user is according to the given user privileges then it ignore those events. Any suspicious events that are generated by abnormal thread, read, writes, network logins, data transfer, hardware changes, and privileges changes are detected in real time for the identification of Insider Threat. The three-level hierachal classification of proposed threat analyzer algorithm is shown in Fig. 3. The detected Insider Threat is displayed in a user-friendly web based GUI which is only accessible to system administrators details of which is given in the next Section.

3.4.3 Threat Representation in GUI

A web-based GUI is developed for the representation of the insider. The GUI is built with Kibana frame work which is used in the representation of big data. As the sensor generated logs ranges in GB's per hour so accurate representation of important event logs was necessary. The data is represented in two formats first in the form of an event graph which is update with time and then the logs represented in a customizable manner so system admin can choose what type of information they wanted to be displayed first. All logs are stored in a database which are retrievable by log search functionality and can be used for investigation if an incident happens. Figures 4 and 5 show the actual GUI of web platform.

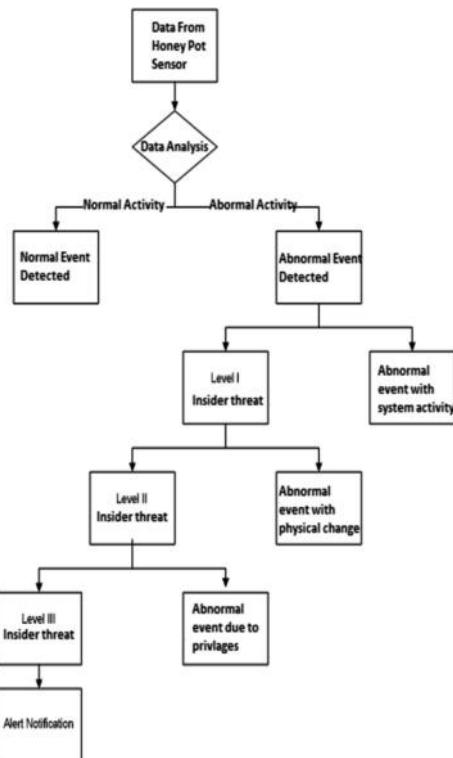


Fig. 3. Three level classification of threat analyzer algorithm

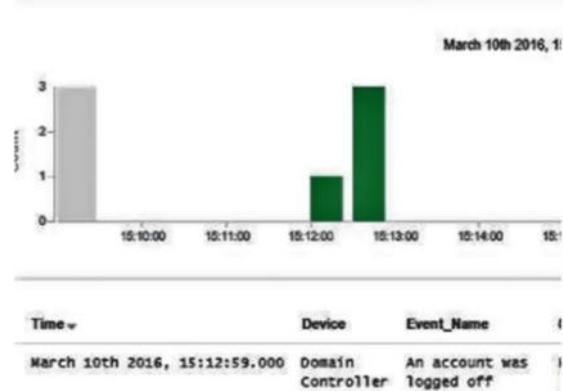


Fig. 4. Representation of system logs in web based GUI



Fig. 5. Representation of system logs in web based GUI

3.5 Test Cases

In the given test cases, the conduct of no participant was observed to be threatening. This result was expected by the researcher as the sample participants involved in the study were fully aware of the fact that their behavior is being continuously monitored. By considering this factor, the researcher also carried out tests by prompting ordinary insider assaults with the objective of detecting any irregular conduct during the simulated attack. The descriptive Section below shows the different experimental evidence that were collected at the end of each research. The researcher would also discuss the objectives and issues related to every experiment in this Section.

3.5.1 Typical Behavior

This test made use of contrast of normal behavior of users to their profile of normal behavior. At the beginning of the experiment, the participants were directed to create a normal behavior profile. Afterwards, their activities were observed in the view of probabilities of nominal behavior while they were performing their normal day to day jobs. However, this practice does not necessarily relate to an insider attack but in a way, it somehow give worthy understanding of the desired probabilities for nominal behavior which is supposed to be normal.

3.5.2 Thread Bomb

A thread bomb is largely a type of assault of denial of carrier. On this experimental check, a thread bomb is added into an application together with notepad or net explorer and it starts off evolved to generate both boundless quantity of threads and a pre-calculated variety of threads. This process appears after the level of profile education. Like diverse denial of offerings attack, this particular attack is also much unexpected and has the ability to shut down all the functions of the host machinery making it cripple, stopping its functions and making the host vulnerable to various forms of

assaults. It is therefore very important to detect this attack and take precautionary measures earlier before it damages the system of the host and makes it disable. In order to conduct this test, the researcher initiated an attack that exploited Notepad and created two thousand threads.

The abnormal activity is quickly detected by the Insider Threat detection system and displayed on the real data stream analyzer.

3.5.3 Abnormal File System Activity

Instances of cybercrime exfiltration and data corruption consists of admittance to the working of the file system. In the cases of real life exfiltration, the hacker sometimes access directories that they are not allowed to access and which comes out of range of their work ethics. The researcher conducted the following experiments for the detection of abnormal activities in the file system of TEST BED.

Abnormal File Deletions Prior to the creation of profiles of normal users, the researcher formulated random 40 MB directories within the programs of users, and formulated 32 directory in the system. After the formulation of normal profile, it was observed that many participants methodologically deleted a group of files from these respective directories. The participants narrated that they rarely execute tasks within the boundary of system 32 and in program directory as they make use of machines with least user privileges (LUP).

Abnormal File System Activity After formulating a normal profile, each participant formulated a varied version of changed, deleted, and modified files in the drives that they usually do not access.

3.5.4 Abnormal Network Activity

In the given experiments, the researcher has attempted to detect irregular endpoint-to-endpoint community interest through investigating the IP addresses of network connections. The researcher has also attempted to look at the ordinary port get right of entry to. The subsequent experiments have been performed after the degree of profile training:

New Network Connection: Every participant can pay everyday visits to an overseas internet site that they generally do now not go to.

New Port: Each participant tried to connect up with the remote servers on ports that they normally do now not connect.

To simulate the issue the researcher, launch a simulated brute force attack for remote logging in on a machine present in the network.

3.5.5 Abnormal Process Activity

The activity of abnormal process can be a precursor that dangerous application, malicious code, or any other form of malware is there in the system. As, it has been narrated earlier that the insider attacks on the organization can prove to be very harmful and dangerous if they are left undetected and uninvestigated. The researcher has

invented different experiments that could be really beneficial in detecting the abnormal process activity. The following tests were executed after the stage of profile training:

New Program: Each participant began to utilize the program that they did not apply in the stage of profile training.

3.5.6 Abnormal Hardware States

A typical hardware states can be a sign of a range of denial-of-service and records corruption assaults. The following assessments were depended on to be helpful in finding strange strategies and hardware changes:

Processor: Each participant conducted and utilized programs to improve the work load of the processor.

Hard Drive Test 1: After the completion of stage of profile training, every participant downloaded at least 10 GB of data into the disk drive.

Hard Drive Test 2: Earlier than the final testing of level of profile training, each participant downloaded 15 GB of random information into the disk drive of employee system. After training segment, the 15 GB of information downloaded with the aid of the participant on their systems would be deleted.

4 Results and Discussion

The proposed framework which has been explained in the Sect. 3, shows feature extraction of different daily threat activity with insider data from embedded sensors of networks. The honeypot sensor detects the abnormal activity of an individual and the sensor show the threat posed by the individual. The Weka machine learning toolbox is used for feature extraction of raw data from system calls generated from honeypots. The framework is used for 3-level hierarchical classification.

In this section we have evaluated the performance of experiment and subsequent result for Insider Threat detection with activity classification and hierarchical detection of insider, an experimental study of 50 users have been carried out with mentioned dataset in this research as shown in Table 1. In the present experiment, raw data is collected with the help of network sensors. The raw data of honeypots is transformed to segment of 10 s called sample data. Then raw data is used for further feature extraction of different classes in daily life activity. The feature classification is used with Insider Threat value for further 3 classes i.e. Class1, Class2 and Class3. The class1 contains normal user behavior in the network like with Insider Threat value ranging from 0 to 1. The Class2 contains suspicious user behavior like accessing unauthorized data with Insider Threat value ranging from 2 to 5. Similarly the Class3 contains insider attacker behavior like manipulating unauthorized data with Insider Threat value ranging from 6 to 10.

Table 1. Insider threat alarms due to application software and system services

User	Scenario	User activity	Initial category	Final category	False alarm (%)
6	Abnormal process activity	System administration service was running	Attacker	Attacker	100
12	Thread bomb	Data compression software was running	Attacker	Attacker	100
18	Abnormal process activity	System administration service was running	Attacker	Attacker	100
24	Abnormal network activity	Torrent Client was running on system	Attacker	Attacker	100
25	Abnormal process activity	System administration service was running	Suspicious	Attacker	100
48	Thread bomb	Data compression software was running	Attacker	Attacker	100

4.1 Validation of Honeypot Sensor

The validation of honeypot sensor is done by generating a single Insider Threat in the TEST BED of 50 systems and then confirming its detection by the insider detection system. The standard deviation and mean of acquisition of Insider Threat values from normal user behavior generated and manually generated attacks from sensor detected data is calculated.

4.2 False Positive Detection at Level 1

The detection of false positive alarm is based on user activity factor. Insider Threat detection percentage are checked in accordance with system activity i.e. from normal to suspicious; the abnormal Insider Threat is detected as false positive alarm due to different.

Application software and system services running on system which raised false alarms as shown in Table 1.

4.3 False Positive Detection at Level 2

The abnormal insider is effected by the physical system changes like USB connection, peripheral device removal etc. The level 2 classification is used to detect false positive Insider Threat detection percentage due to physical changes in system as shown in Table 2.

4.4 False Positive Detection at Level 3

The increase of Insider Threat detection percentage is due to the activities of a privileged user who has the write of read, write and modify the system files this caused unnecessary false Insider Threat alarm which can be seen in Table 3.

Table 2. Insider threat false alarm due to hardware changes

User	Scenario	User activity	Initial category	Final category	False alarm (%)
9	Abnormal hardware state	USB disk attached on system	Suspicious	Attacker	100
14	Abnormal hardware state	Data moved from one drive to another within the system	Attacker	Attacker	100
16	Abnormal hardware state	USB disk attached on system	Suspicious	None	100
34	Abnormal hardware state	Data moved from one drive to another within the system	Attacker	Attacker	100

Table 3. Insider threat false alarm due to activities of privileged users

User	Scenario	User activity	Initial category	Final category	False alarm (%)
49	Abnormal file system activity	File is deleted by a privileged user	Suspicious	Attacker	100

4.5 Actual Abnormal Insider Threat Value

Table 4 is used to detect true positive result based on Insider Threat detection percentage. By classification of a 3-level system, true positive Insider Threat value is detected as shown in Table 4.

Table 4. Real insider threat detected

User	Scenario	User activity	Initial category	Final category	False Alarm (%)
5	Thread bomb	Simulated thread bomb launched on system	Attacker	Attacker	0
8	Abnormal network activity	Simulated port scanning was performed from the system	Attacker	Attacker	0
19	Thread bomb	Simulated thread bomb launched on system	Attacker	Attacker	0
28	Abnormal process activity	Simulated malware was introduced into the system	Attacker	Attacker	0
29	Abnormal process activity	Simulated malware was introduced into the system	Attacker	Attacker	0
30	Thread bomb	Simulated thread bomb launched on system	Attacker	Attacker	0
33	Abnormal network activity	Simulated port scanning was performed from the system	Suspicious	Attacker	0
37	Abnormal file system activity	File is copied by a under privileged user	Attacker	Attacker	0
38	Abnormal process activity	Simulated malware was introduced into the system	Attacker	Attacker	0
42	Thread bomb	Simulated thread bomb launched on system	Attacker	Attacker	0
44	Abnormal hardware state	Data copied to external hard drive from the system	Suspicious	Attacker	0
46	Abnormal process activity	Simulated malware was introduced into the system	Attacker	Attacker	0

4.6 Result Comparison with Existing System

Table 4 shows the difference in results between proposed and existing system for one-hour operation which is graphically represented in Fig. 7. Table 6 shows the difference in results for 24 h of operation which is graphically represented in Fig. 8. It is clearly seen that the proposed system is working more reliably as compared to the existing system. Table 7 shows the difference of results in proposed and existing systems for the period of 10 days which is graphically represented in Fig. 9. From the gather data of 10 days total number of malicious events was found to be 4963 for existing system and 3874 in proposed system. It is then calculated average improvement in proposed system is 21.9% as compared to existing system.

This research made use of the procedure of Insider Threat detection with the use of honeypot sensors. In local networks, honeypots sensors are used to detect and relay the information of Insider Threat with the activity recognition of an insider. The Real Data Stream analysis plays the role of hub, processor and sensor, to transmit the preprocessed data received from honeypot sensor. The classification of data from honeypot sensor is processed by using the Insider Threat rate classifier in Kabana toolbox for activity recognition.

The real time monitoring of system calls by using honeypot is low cost and less complex, because the system calls of measurement from each system is not required to be implanted and monitored continuously. Results have shown that the present framework, activity classification and false positive Insider Threat value detection mechanism is producing accurate results. The zero mean error is used to count Insider Threat value with the real time accuracy.

In near future, we aim to extend his work with data efficiency and other relative vital sign for measurement and analysis of Insider Threat.

Figure 6 shows the actual number of warnings generated by the proposed Insider Threat detection system. In the graph it can be seen that out of nearly eight thousand events only thirty-eight are found to be actual warning. This is achieved by three level hierachal classification of user generated events to avoid false alarm and optimize system performance.

Figure 7 shows the data of Table 5 in a graphical manner, in which effects of different events generated by users are measured against both proposed and existing systems. Events generated by thread bombs, abnormal network activity, file read and write operations, hardware changes and system process activity are carefully measured to calculate the performance improvement in proposed system as compared to existing system.

Figure 8 shows the data of Table 6 in a graphical manner, in which performance of existing and proposed systems are compared for a period of one day with respect of events generated to compare the insider detection performance of both systems. 16% improvement was measured in a single day of comparison between both systems.

Figure 9 shows the data of Table 7 in a graphical manner, in which average performance of Insider Threat detection system of existing and proposed system were compared for the period of ten days. It was calculated that average performance increase in proposed system is 21.9% as compared to existing system.



Fig. 6. Actual number of warnings generated by insider threat detection system

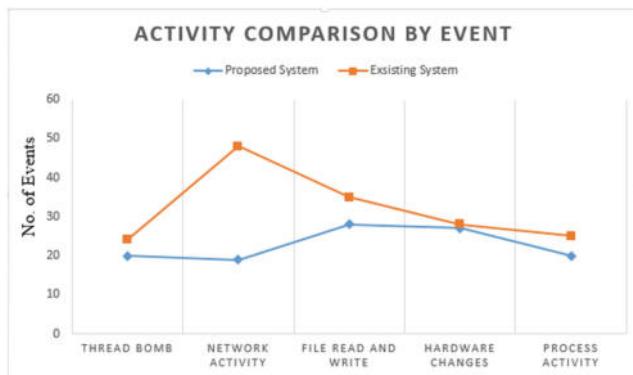


Fig. 7. Activity comparison of proposed and existing system with respect to events

Table 5. Event activity comparison between proposed and existing system in an hour

Events activity	Proposed system	Existing system
Thread bomb	20	24
Network activity	19	48
File read and write	28	35
Hardware changes	27	28
Process activity	20	25

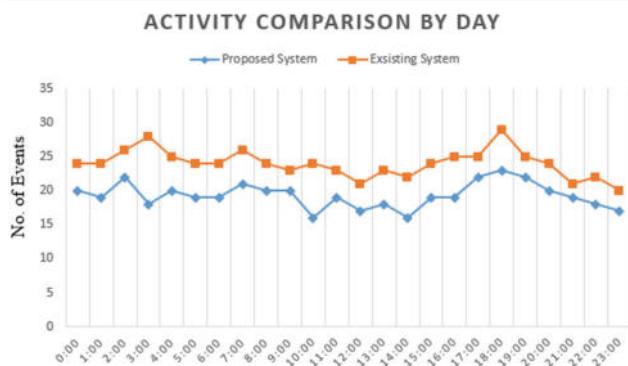


Fig. 8. Activity comparison of proposed and existing system with respect to time

Table 6. Event activity comparison between proposed and existing system in a day

Time	Proposed system	Existing system
0:00	20	24
1:00	19	24
2:00	22	26
3:00	18	28
4:00	20	25
5:00	19	24
6:00	19	24
7:00	21	26
8:00	20	24
9:00	20	23
10:00	16	24
11:00	19	23
12:00	17	21
13:00	18	23
14:00	16	22
15:00	19	24
16:00	19	25
17:00	22	25
18:00	23	29
19:00	22	25
20:00	20	24
21:00	19	21
22:00	18	22
23:00	17	20

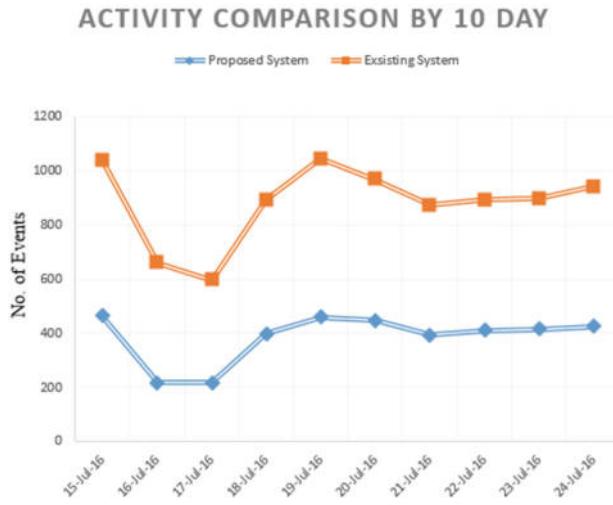


Fig. 9. Activity comparison of proposed and existing system with respect to 10 day

Table 7. Event activity comparison between proposed and existing system in 10 day

Date	Proposed system	Existing system
15-July-16	463	576
16-July-16	220	440
17-July-16	216	380
18-July-16	398	496
19-July-16	458	587
20-July-16	446	523
21-July-16	396	476
22-July-16	410	482
23-July-16	416	483
24-July-16	424	520

5 Conclusion

Section 1 states the background and significance of Insider Threat detection. Insider Threats are very common in big multinational companies and their rival companies can easily destroy other company with the help of an insider. An insider is basically a part of the parent company and works for the benefit of rival company. An insider can damage the data of company with the help of exfiltration, data corruption, and abnormal activities and by denial of services technique. Therefore, it is very important that companies should be equipped with adequate Insider Threat detection system to ensure safety from the Insider Threat. Insider Threat can be checked by the use of sensors

and honeypots who can check the activities of employees on the main framework and raise an alarm if any abnormality is found out in the behavior of employees.

Section 2 is about the literature review of term Insider and Insider Threat detection. We defined the term Insider in detail and quoted many examples and cases of Insider assaults in various organizations. The bottom line of this chapter is that Insider Threat is a very recent and contagious type of cybercrime in the modern era and therefore, proper safety measures should be adopted to cope with the bad and harmful effects of this problem.

In Sect. 3, we proposed the framework and design for the research work. The researcher has taken sample of 50 PCs in a TEST BED lab system. The research has continuously monitored the activities of employees within a specific duration. The nominal behavior of sample population was studied before and after the profile training phase. Basically, we implanted a honeypot on network and that honeypot collected data of any person who try to illegal access any irrelevant website. The honeypots have the ability to track both the IP address and the physical address of the attacker.

In Sect. 4, we presented the validation of Honeypot sensors with reference to the detection of Insider Threats. The Insider Threats have been validated and classified on three levels respectively. The research has graphically represented the validation of Insider Threat detection with reference to honeypots in the form of Tables and illustrations. These Tables and their values have been extracted after the study of sample population of TEST BED participants; activities on PC frameworks. The improvement in proposed Insider Threat detection system is carefully calculated and average improvement of 21.9% was found as compared to existing system.

Finally, we presented the findings of the results after the quantification of collected data. We gave a brief review of insider and Insider Threats to the readers. Consequently, we pointed out various cases of insider attacks in different organizations and concluded that these insider attacks could have been prevented if the companies have implemented in them a proper insider detection system. Then, we mentioned some old techniques of detecting Insider Threats in organizations that made use of Honeypots. In previous times, Honeypot sensors were used to be installed at every single PC and most of the honeypots detect false positive alarms about the insiders. This created a lot of complexities. Therefore, the researcher took motivation from this point and tried to improve this process of Insider Threat detection by associating the honeypot on the main network of PC server. We then made software and the purpose of this software was that the honeypot keep a check on every activity of each PC on the system. Whenever any suspicious activity was found out by the honeypot, then the system of that suspicious employees locked automatically. This technique is produced for the detection of Insider Threat. It is expected that in future more research would be done to improve the system of Insider Threat detection system by optimizing the amount of data generated from each honey pot to save disk space and to avoid unnecessary network activity. Kernel Density Estimation algorithms can also be improved to make the insider detection system appropriate for external threats detection and to perform threat intelligence.

References

1. Mallah, G.A., Shaikh, Z.A.: A platform independent approach for mobile agents to monitor Network vulnerabilities. *WSEAS Trans. Comput.* **4**(11), 1672–1677 (2005)
2. Moore, A.P., Cappelli, D.M., Caron, T., Shaw, E., Trzeciak, R.F.: Insider theft of intellectual property for business advantage: a preliminary model. In: Proceedings of the 1st International Workshop on Managing Insider Security Threats (MIST2009), Purdue University, West Lafayette, USA (2009)
3. Hayden, M.: The insider threat to US government information systems (No. NSTITSAM-INFOSEC/1-99). National Security Agency/Central Security Service Fort George G Meade Md (1999)
4. Ahmad, M.B., Akram, A., Asif, M., Ur-Rehman, S.: Using genetic algorithm to minimize false alarms in insider threats detection of information misuse in windows environment. *Mathematical Problems in Engineering* (2014)
5. Legg, P.A., Buckley, O., Goldsmith, M., Creese, S.: Automated insider threat detection system using user and role-based profile assessment. *IEEE Syst. J.* **1**–10 (2015)
6. Bishop, M.: The insider problem revisited. In: Proceedings of the 2005 workshop on New security paradigms, pp. 75–76. ACM (2005)
7. Grobauer, B., Schreck, T.: Towards incident handling in the cloud: challenges and approaches. In: Proceedings of the 2010 ACM workshop on Cloud computing security workshop, pp. 77–86. ACM (2010)
8. McKinney, S., Reeves, D.S.: User identification via process profiling. In: Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies, p. 51. ACM (2009)
9. Qiao, H., Peng, J., Feng, C., Rozenblit, J.W.: Behavior analysis-based learning framework for host level intrusion detection. In: 14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems (ECBS'07), pp. 441–447. IEEE (2007)
10. Shavlik, J., Shavlik, M., Fahland, M.: Evaluating software sensors for actively profiling Windows 2000 computer users. In: Fourth International Symposium on Recent Advances in Intrusion Detection (2001)
11. Spitzner, L.: Honeypots: catching the insider threat. In: Computer Security Applications Conference, 2003. Proceedings. 19th Annual, pp. 170–179. IEEE (2003)
12. Yu, Y., Chiueh, T.C.: Display-only file server: a solution against information theft due to insider attack. In: Proceedings of the 4th ACM workshop on Digital rights management, pp. 31–39. ACM (2004)
13. Pramanik, S., Sankaranarayanan, V., Upadhyaya, S.: Security policies to mitigate Insider Threat in the document control domain. In: Computer Security Applications Conference, 2004. 20th Annual, pp. 304–313. IEEE (2004)
14. Park, J.S., Ho, S.M.: Composite role-based monitoring (CRBM) for countering insider threats. In: International Conference on Intelligence and Security Informatics, pp. 201–213. Springer, Berlin (2004)
15. Ali, G., Shaikh, N.A., Shaikh, Z.A.: Towards an automated multiagent system to monitor user activities against Insider Threat. In: International Symposium on Biometrics and Security Technologies, 2008. ISBAST 2008, pp. 1–5. IEEE (2008)
16. Cathey, R., Ma, L., Goharian, N., Grossman, D.: Misuse detection for information retrieval systems. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, pp. 183–190. ACM (2003)

17. Ma, L., Goharian, N.: Query length impact on misuse detection in information retrieval systems. In: Proceedings of the 2005 ACM Symposium on Applied Computing, pp. 1070–1075. ACM (2005)
18. Aleman-Meza, B., Burns, P., Eavenson, M., Palaniswami, D., Sheth, A.: An ontological approach to the document access problem of Insider Threat. In: International Conference on Intelligence and Security Informatics, pp. 486–491. Springer, Berlin (2005)
19. Liu, A., Martin, C., Hetherington, T., Matzner, S.: A comparison of system call feature representations for insider threat detection. In: Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop, pp. 340–347. IEEE (2005)
20. Anderson, K., Carzaniga, A., Heimbigner, D., Wolf, A.: Event-based document sensing for Insider Threats. University of Colorado, Computer Science Technical Report CU-CS-968-04 (2004)
21. Nguyen, N.T., Reiher, P.L., Kuenning, G.H.: Detecting insider threats by monitoring system call activity. In: IAW, pp. 45–52 (2003)
22. Ahmad, M.B., Akram, A., Islam, H.: Implementation of a behavior driven methodology for insider threats detection of misuse of information in windows environment. Information **16** (11), 8121 (2013)



Subject Identification from Low-Density EEG-Recordings of Resting-States: A Study of Feature Extraction and Classification

Luis Alfredo Moctezuma^(✉) and Marta Molinas

Department of Engineering Cybernetics, Norwegian University of Science and Technology, 7491 Trondheim, Norway

luisalfredomoctezuma@gmail.com, marta.molinas@ntnu.no

Abstract. A new concept of low-density electroencephalograms-based (EEG) Subject identification is proposed in this paper. To that aim, EEG recordings of resting-states were analyzed with 3 different classifiers (*SVM*, *k-NN*, and *naive Bayes*) using Empirical Mode Decomposition (EMD) and Discrete Wavelet Transform (DWT) for feature extraction and their accuracies were estimated to compare their performances. To explore the feasibility of using fewer channels with minimum loss of accuracy, the methods were applied to a dataset of 27 Subjects (From 5 sessions of 30 instances per Subject) recorded using the EMOTIV EPOC device with 1 set of 14 channels and 4 subsets (8, 4, 2 and 1 channel) that were selected using a *greedy* algorithm. The experiments were reproduced using fewer instances each time to observe the evolution of the accuracy using both; fewer channels and fewer instances. The results of this experiments suggest that EMD compared with DWT is a more robust technique for feature extraction from brain signals to identify Subjects during resting-states, particularly when the amount of information is reduced: e.g., using *Linear SVM* and 30 instances per Subject, the accuracies obtained using 14 channels were 0.91 and 0.95, with 8 channels were 0.87 and 0.89 with EMD and DWT respectively but were reversed in favor of EMD when the number of channels was reduced to 4 channels (0.76 and 0.74), 2 (0.64 and 0.56) and 1 channel (0.46 and 0.31). The general observed trend is that, *Linear SVM* exhibits higher accuracy rates using high-density EEG (0.91 with 14 channels) while *Gaussian naive Bayes* exhibits better accuracies when using low-density EEG in comparison with the other classifiers (with EMD 0.88, 0.81, 0.76 and 0.61 respectively for 8, 4, 2 and 1 channel). The findings of these experiments reveal an important insight for continuing the exploration of low-density EEG for Subject identification.

Keywords: Biometric security · Subject identification · Electroencephalograms (EEG) · Resting-states · Empirical Mode Decomposition (EMD) · Discrete Wavelet Transform (DWT)

1 Introduction

To protect places and/or information where privileges are required, organizations use security systems. To achieve this, different measures have been proposed, ranging from security-guards/smart-cards to fingerprint/face-recognition [1,2].

The use of security systems has been increasing not only in organizations but also in low-cost portable devices (e.g., mobile phones, tablets and personal computers). Due to the increasing vulnerabilities to skip the authentication and authorization process of current traditional/biometric security systems [2], there is a growing interest in exploring new biometric measures. With this trend, the use of brain signals to create biometric markers using different neuro-paradigms also has emerged as a robust alternative to the above mentioned vulnerabilities. Brain signals can be used as a basis for the design of biometric markers since they satisfy the requirements of *universality, permanence, collectability, performance, acceptability, and circumvention* [1]. Brain signals are more reliable and secure because biometric markers obtained from EEG-recordings from human brain activity will be almost impossible to duplicate since the brain is highly individual [3].

To promote the concept of “low-cost” affordable devices to record brain signals using different neuro-paradigms, a popular/non-invasive technique using Electroencephalography (EEG) is the well known Brain-Computer Interface (BCI).

Empirical Mode Decomposition (EMD) [4] and Discrete Wavelet Transform (DWT) [5–7] have been applied to transform and analyze brain signals while different mental tasks are performed. Both, EMD and DWT, have shown to be effective in decomposing non-stationary/non-linear time series. But, EMD has the advantage that it does not need the definition of any mother function or pre-processing to improve the signal-to-noise ratio [4]. On the other hand, DWT needs a pre-processing stage to fit the appropriate mother function depending to the task/neuro-paradigm used.

Biometric systems based on EEG-recordings can be separated into states: task-related-state and resting-states. In task-related state different ways have been used to stimulate the brain, for example in [8] Visual Counting and geometric figure Rotation (visual stimulation of images) were used. Another way presented in [9] consisted in mental composition of letters, or as in [10], imagining random digit numbers were presented, among others method and techniques that can be found in [11,12]. However, persons with certain diseases (e.g., Amyotrophic Lateral Sclerosis, Attention Deficit Disorder, etc.) [13,14] cannot perform some tasks and the use of the above stimuli are not feasible.

The use of EEG signal from resting-states has been reported for example in [7] where a method based on Morlet Wavelet and *Linear SVM* was tested using a dataset of 40 Subjects (192 instances per Subject) and the signal was captured with a sample rate of 256 Hz from 64 channels. In that work, resting-states with the lengths of 300, 60 and 30 s were used obtaining accuracies of 1.00, 0.96 and 0.72. However, the use of 300, 60 or even 30 s of brain signal

length is computationally costly and for a real-time application, fast recognition capabilities with limited information will be essential.

In [15], a method based on Convolutional Neural Networks using the raw signal as input (without pre-processing and without feature extraction) was presented. The dataset was obtained using BCI2000 from 64 channels with a sample rate of 160 Hz. from 10 Subjects and 55 instances of 1 s of duration per subject. Three different experiments were presented: Using resting-states with Open-Eyes/Closed-Eyes/both, and the accuracies obtained were 0.88, 0.86 and 0.82 respectively.

Although the use of resting-states as a biometric marker has been reported by several researchers [7, 11, 15], the possibility of using fewer channels or fewer instances has not been explored so far. As mentioned earlier in the paper, the use of 64 channels does not support the concept of a flexible, low-cost portable EEG-device as presented in [16]. The biometric systems currently adopted by the industry/market use about 5 instances or even fewer to add a new person (e.g., fingerprint, voice/face recognition, retinal scans) and in the research on biometric systems based on EEG, 192 or 55 instances per Subject, which is not practical for a real implementation.

In more recent works, Subject identification methods based on *imagined speech* using DWT [5] and EMD [4] for feature extraction, were presented and the results obtained suggest that EEG of *imagined speech* can be a good candidate as a biometric marker. In the present work, a new conceptual proposition using resting-states (unconstrained rest) in conjunction with fewer EEG channels/instances, is explored. Resting-states can be a valuable biometric marker when the population is large (e.g., in an airport, big enterprises or government organizations) because of its self-reliance and inherent independence from training.

In the following, the new concept of EEG with a reduced number of channels/instances will be presented followed by the proposition of using resting-states (resting-states without restrictions) in conjunction with this new EEG concept as a flexible and affordable portable recognition/authentication system.

2 Towards a Low-Density EEG Concept: FlexEEG

In order to realize a low-cost and flexible solution for subject identification from brain signals, a new EEG concept is envisioned and presented in this paper. This new EEG concept will be based on a design with a reduced number of channels and the use of wireless dry electrodes to support portability and ease of use. While a laboratory setting and research-grade EEG equipment ensure a controlled environment and high-quality multiple-channel EEG recording, there are applications, situations, and populations for which this is not suitable. Conventional EEG is challenged by high cost (i.e., computationally costly), high-density, immobility of equipment and the use of inconvenient conductive gels. One consequence of high-density EEG is that interpretation in real-time is not available today. Technological advancements in dry sensor system have opened avenues of

possibilities to develop wireless and portable EEG systems with dry electrodes to reduce many of these barriers.

In [16] a new EEG concept of portable (non-invasive) dry single-channel or low-density EEG system, was introduced. While being portable and relying on dry-sensor technology, it will be expected to produce recordings of comparable quality to a research-grade EEG system but with wider scope and capabilities than conventional lab-based EEG equipment. In short, a single more intelligent EEG sensor could defeat high-density EEG. Through this new concept, the range of applications of EEG signals will be expanded from clinical diagnosis and research to health-care, to better understanding of cognitive processes, to learning and education, and to today hidden/unknown properties behind ordinary human activity and ailments (e.g., resting-states, walking, sleeping, complex cognitive activity, chronic pain, insomnia, etc.).

The proposition of real-time Subjects identification using low-density EEG recordings of resting-states will benefit from an EEG device that can offer the flexibility and capabilities envisioned in the FlexEEG concept. The combination of resting-states brain signals with a flexible EEG design with a reduced number of channels will make possible to materialize low-cost and seamless Subject identification within the reach of everyone.

3 Methods

In this section, the methods used with the aim of Subject identification are described in brief. The idea of using resting-states for Subjects identification is motivated by the fact that resting-states are typically used to analyze problems relative to the Subject internal state of mind [13], and this suggests the existence of unique patterns pertaining to the Subject.

According to [17], a stable resting-state (even called resting-state activity) does not necessarily exist, because spontaneous changes in regional neuronal firing occur even when the organism is apparently in rest-state. Also, spontaneous activation can change local blood flow, cause low-frequency blood oxygenation level-dependent signal fluctuations [18]. In other words, the brain is never really at rest [19], and the term only refers to the absence of goal-directed neuronal action with the integration of information of external environment and the Subject internal state, that could be a starting point to discuss why the Subject identification task can work, which in this paper will be done experimentally.

The methods were applied to a dataset of resting-states from the low-cost EMOTIV EPOC device using 14 channels, 8 (P7, P8, O1, O2, F7, F8, T7 and T8), 4 (F7, F8, T7 and T8), 2 (T7 and T8) and 1 channel (T7) that were placed according to the 10–20 international system [20]. Subsets of channels were selected using a *greedy* Algorithm [21] as a first attempt to move towards the FlexEEG Concept [16]. This is done in order to analyze the evolution of the accuracy using each time fewer channels, as it is explained later. Additionally, for the set and each subset of channels, experiments were reproduced using 30, 20, 10, 5 and 3 instances per subject, to observe the evolution of accuracy and to

obtain a first approximation of the necessary instances to create a competitive system to the current biometric systems used in industry.

In the following subsections, the methods used are explained in brief, including the dataset description, the logic for channels selection, the feature extraction and classification techniques.

3.1 Dataset Description

To test the new conceptual proposition, EEG signals obtained from the low-cost EMOTIV EPOC device with a sample rate of 128 Hz and 14 channels, were used. The dataset consists of brain signals from 27 subjects while imagining 33 repetitions of five imagined words in Spanish, where each repetition was separated by a resting-state. The protocol for acquisition is shown in Fig. 1 and is described in [22].

The imagined words were recorded in 5 different sessions (not consecutively one after the other), that allows the use of resting-states between instances of words and from 5 different sessions. The mean size of the resting-states in the dataset is [3] s.

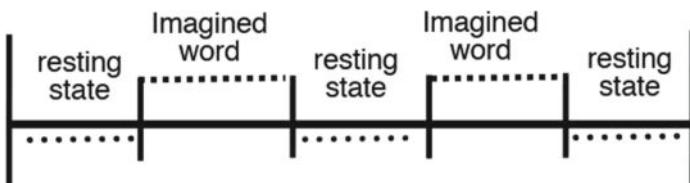


Fig. 1. Protocol for EEG signal acquisition using EMOTIV EPOC [6, 22]

The 14 recorded electrodes as shown in Fig. 2 were placed according to the 10–20 international system [20].

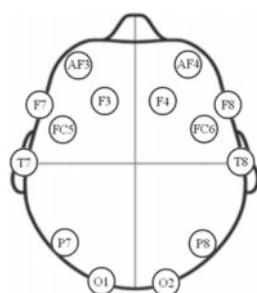


Fig. 2. 10–20 International system for 14 channels

Algorithm 1 Greedy algorithm for channels selection

```

1: procedure CH_SELECTION(subjects)      ▷ 27 Subjects, 14 channels per each one.
2:   sj  $\leftarrow$  len(subjects)
3:   ch  $\leftarrow$  len(sj[0])
4:   ch_selected  $\leftarrow$  []
5:   while ch  $>$  1 do                  ▷ Stop if there are no channels to remove.
6:     ch_combinations  $\leftarrow C_{ch,1}$           ▷  $k$ -combinations,  $C_{n,k}$ 
7:     accuracies = []
8:     for ch_combination in ch_combinations do
9:       accuracies  $\leftarrow$  accuracies  $\cup$  classifier(subjects, ch_combination)
10:    end for
11:    highest_accuracy  $\leftarrow$  max(accuracies)
12:    ch  $\leftarrow$  ch_combinations[highest_accuracy]           ▷  $ch \leftarrow ch - 1$ 
13:    ch_selected  $\leftarrow$  ch_selected  $\cup$  ch
14:   end while
15:   return ch_selected                         ▷ Channels selected
16: end procedure

```

3.2 Channel Reduction Criteria for Low-Density EEG

A first step towards the low-density EEG concept discussed in [16] is the channel reduction approach applied based on the information provided for a given neurophysiological task. With this starting point, restrictions of fixed electrodes, the use of a single design for different tasks and its implications for the device portability are discussed. The logic based on the *greedy* algorithm was applied for channels reduction a first step to understand how many channels are needed to obtain sufficient information to detect the relevant activity of the brain.

Channel Selection: In the Algorithm 1 the *greedy* procedure presented in [21] has been adopted to remove channels in a step-by-step manner. The idea is to obtain the combinations removing 1 channel at a time (k -combinations: $k = 1$) and selecting the subset with the highest accuracy (local maximum). Then the procedure is repeated with the subset obtained while the length of the subset is still greater than 1 channel.

3.3 Feature Extraction

Two methods were used for this purpose to compared their capabilities. The first method used is based on the EMD algorithm for which the relevant Intrinsic Mode Functions (IMFs) were decided based on Minkowski distance [23]. Then, for each IMF 4 features were computed: *Instantaneous/Teager energy distribution* and *Higuchi/Petrosian Fractal Dimension*. The flowchart for feature extraction from a given channel is shown in Fig. 3 and is detailed in [4].

For the second method used, the DWT, first the Common Average Reference (CAR) was applied to improve the signal-to-noise ratio and then the biorthogonal

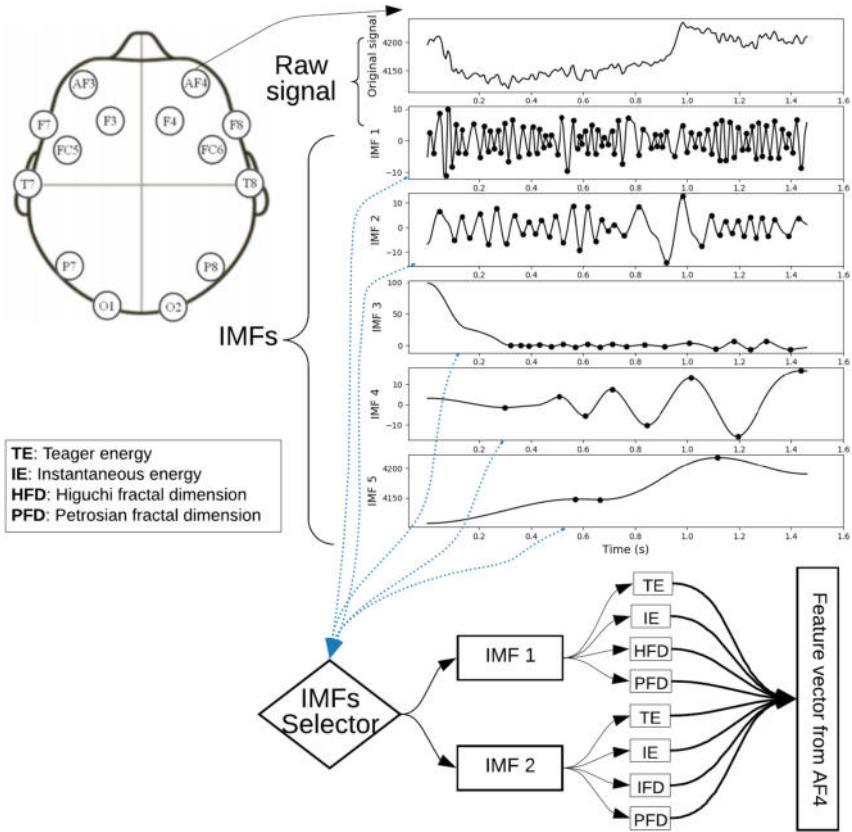


Fig. 3. Flowchart summarizing the feature extraction procedure using EMD

2.2 (bior2.2) DWT with 4 levels of decomposition was computed. Then, for each level of decomposition the *instantaneous energy* was obtained. The flowchart describing the method is shown in Fig. 4 and is detailed first in [6] and then used for Subject identification using *imagined speech* in [5].

3.4 Classification

The classification procedure was performed using *SVM* (with the kernels: *Linear*, *Sigmoid* and *Radial Basis Function*), *Gaussian naive Bayes* and with k -*NN* ($k = 1, 2, 3, 4$).

To estimate the accuracy and thus evaluate the performance of the methods, $\{10, 10, 10, 5, 3\}$ -folds cross-validation were respectively used for each experiment using respectively 30, 20, 10, 5 and 3 instances per Subject.

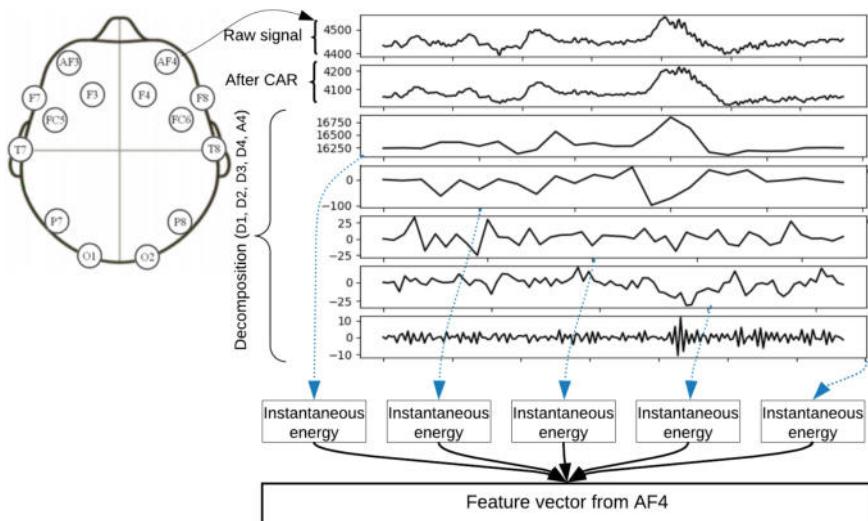


Fig. 4. Flowchart summarizing the feature extraction using bior2.2 DWT

3.5 Experiment Setup

The method used to reduce the number of EEG channels provides a general outlook about which channels contain more information and the minimum number of channels for which the loss of accuracy is minimum. However, because the analysis in this paper was carried out between 5 different sessions and using fewer instances; the channels selected in the first sessions/experiment were not the same when using fewer instances or even when using DWT and EMD.

In order to ensure fairness of comparison, the same channels need to be used with EMD and DWT in all sessions and with a different number of instances. Therefore, only the channels that were common to all the experiments were selected. Then, with these subsets of channels, the experiments were repeated with the common channels, to obtain a fair comparison using EMD and DWT.

4 Results

The idea behind the use of fewer channels is to understand and observe the evolution of the accuracy when using DWT and EMD. On the other hand, the method used for channel selection provides a general outlook about which are the channels that contain more information for the Subject identification task and the neuro-paradigm used (In this case: resting-states). According to the method applied for channel selection and after the analysis of the common channels (common channels between instances-used/sessions/methods), the experiments were repeated using 14 channels, 8 (P7, P8, O1, O2, F7, F8, T7 and T8), 4 (F7, F8, T7 and T8), 2 (T7 and T8) and 1 channel (T7).

In Fig. 5 the average accuracies obtained from 5 different sessions with *Linear SVM* and using the set and subsets of channels, are shown.

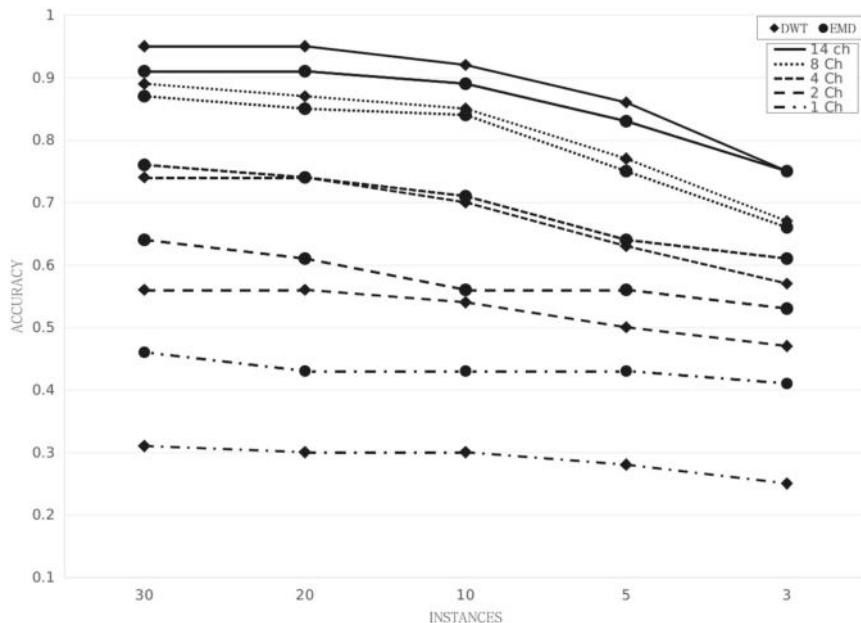


Fig. 5. Average accuracies obtained from 5 sessions using different number of channels with *Linear SVM*

When 14 and 8 channels were used, the highest accuracies were reached with DWT even when using 3 instances. However using 4, 2 and 1 channels the highest accuracies were reached using EMD. For this task, the evolution of accuracy is clear and easy to understand that EMD can better represent the signal even when the information is reduced. At the same time, these results show that DWT is a more robust method for transforming the brain signals and for feature extraction when the amount of information is higher, which in this context is achievable with a high-density EEG device.

For example, using 14 channels and 30 instances the accuracies obtained with DWT and EMD were 0.95 and 0.91, but using 1 channel and 3 instances the accuracies obtained with DWT and EMD were instead reversed (0.25 and 0.41 respectively) in favor of EMD.

The results presented in Fig. 6 confirm the observed property above that suggests that the method based on EMD can represent well the brain signal obtaining high accuracy rates.

In Fig. 7 the results obtained using *Gaussian naive Bayes*, are shown. These results still confirm the high accuracy rates using EMD, but also with an apparently random behavior when information is low (using 5 and 3 instances per Subject).

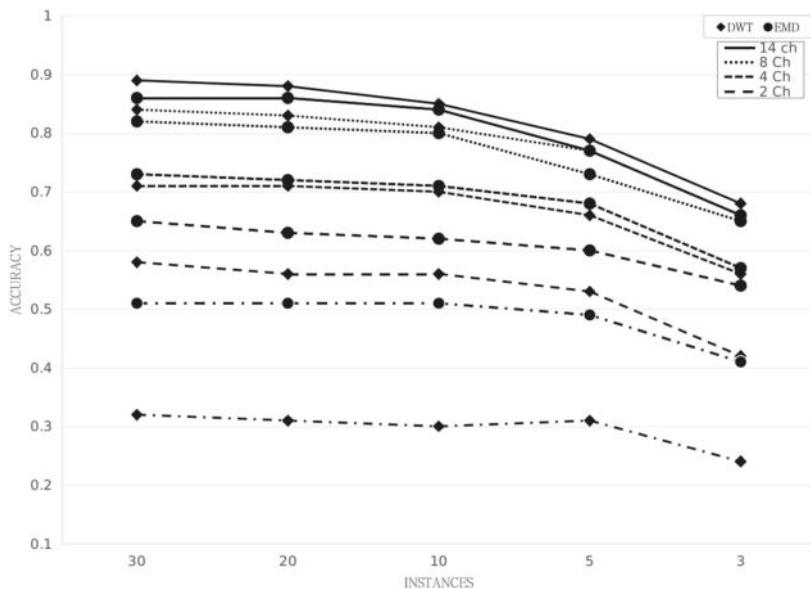


Fig. 6. Average accuracies obtained from 5 sessions using different number of channels with β -NN

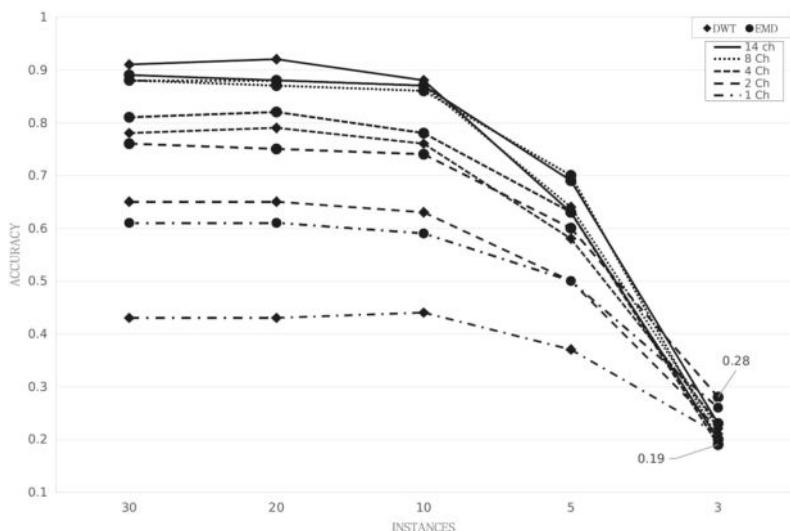


Fig. 7. Average accuracies obtained from 5 sessions using a different number of channels with *Gaussian naive Bayes*

A possible response for the *naive Bayes* behavior was presented in [24], where the authors defend the idea that optimal performance is presented in two extreme cases (completely independent features and functionally dependent features) and the performance is worst between these extremes.

In Fig. 8, the evolution of average accuracies obtained with 20 instances are shown as an example to understand that *SVM* exhibits the highest accuracy using 14 channels but when the number of channels is reduced the *naive Bayes* classifier exhibits higher accuracy rates using EMD compared with the other two classifiers. An important revelation from these result is also that all three classifiers exhibit better accuracies under low-density EEG conditions only when using EMD for feature extraction. In general, using 14 and 8 channels the highest accuracies were obtained using DWT for feature extraction while the highest accuracies using fewer channels (4, 2, 1 channels) were obtained using EMD for feature extraction.

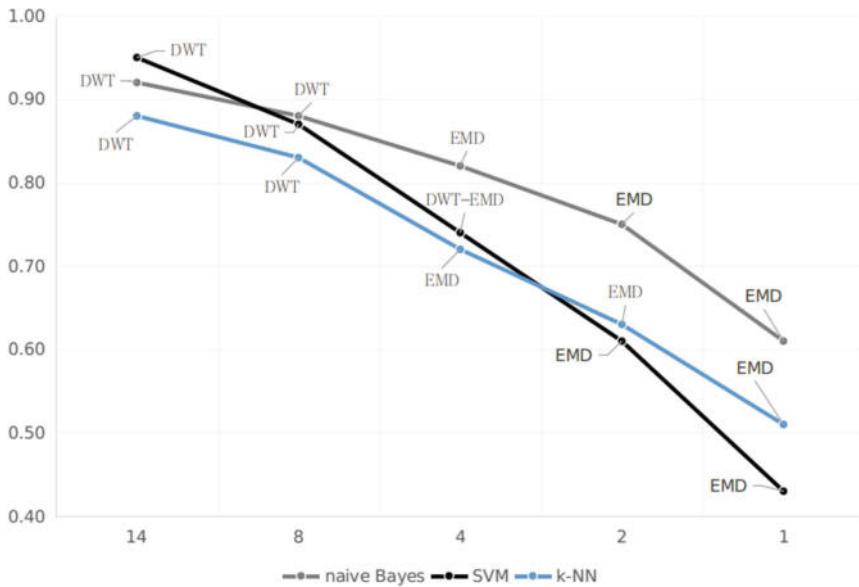


Fig. 8. Evolution of average accuracy using 20 instances with *Linear SVM*, *k-NN* and *Gaussian naive Bayes*

In Tables 1, 2 and 3 the accuracies per session with *Linear SVM*, *3-NN* (*k-NN*) and *Gaussian naive Bayes* using EMD and DWT with a different number of channels and a different number of instances, are presented in detail. The average accuracies (Avg.) and standard deviation (Std) between sessions, also are presented for each subset of channels and using fewer instances.

Inspecting the results in the tables, the following obvious questions come to mind: *What does exactly mean the fluctuation of accuracies? Why sometimes*

Table 1. Accuracies obtained with *Linear SVM* in 5 different sessions using EMD and DWT

Session	Ch	EMD					DWT					
		30	20	10	5	3	Avg.	30	20	10	5	Avg.
1	14	0.89	0.92	0.87	0.83	0.79	0.86±0.05	0.94	0.95	0.94	0.88	0.76 0.89±0.08
	8	0.86	0.87	0.86	0.75	0.69	0.81±0.08	0.90	0.90	0.87	0.84	0.71 0.84±0.08
	4	0.77	0.78	0.79	0.69	0.69	0.74±0.05	0.77	0.79	0.76	0.68	0.68 0.74±0.05
	2	0.65	0.65	0.59	0.62	0.61	0.62±0.03	0.57	0.59	0.58	0.53	0.47 0.55±0.05
	1	0.47	0.48	0.46	0.50	0.47	0.47±0.01	0.32	0.34	0.33	0.26	0.20 0.29±0.06
2	14	0.91	0.91	0.88	0.79	0.81	0.86±0.05	0.95	0.96	0.92	0.87	0.80 0.90±0.07
	8	0.87	0.84	0.84	0.77	0.77	0.82±0.05	0.89	0.88	0.85	0.79	0.71 0.82±0.08
	4	0.76	0.73	0.68	0.63	0.57	0.68±0.07	0.76	0.74	0.72	0.71	0.60 0.71±0.06
	2	0.63	0.57	0.59	0.66	0.52	0.59±0.05	0.61	0.60	0.60	0.54	0.53 0.58±0.04
	1	0.44	0.41	0.42	0.50	0.45	0.44±0.04	0.30	0.28	0.33	0.28	0.33 0.30±0.03
3	14	0.93	0.92	0.91	0.84	0.81	0.88±0.05	0.96	0.96	0.93	0.86	0.75 0.89±0.09
	8	0.89	0.87	0.84	0.80	0.69	0.82±0.08	0.89	0.86	0.85	0.74	0.64 0.80±0.10
	4	0.74	0.71	0.63	0.64	0.57	0.66±0.07	0.70	0.67	0.62	0.58	0.47 0.61±0.09
	2	0.62	0.60	0.52	0.54	0.47	0.55±0.06	0.53	0.51	0.45	0.47	0.39 0.47±0.06
	1	0.48	0.42	0.35	0.37	0.27	0.38±0.08	0.31	0.29	0.26	0.30	0.20 0.27±0.04
4	14	0.92	0.92	0.90	0.86	0.67	0.85±0.11	0.94	0.93	0.88	0.82	0.65 0.85±0.12
	8	0.84	0.84	0.84	0.74	0.59	0.77±0.11	0.86	0.83	0.81	0.72	0.59 0.76±0.11
	4	0.74	0.70	0.66	0.66	0.60	0.67±0.05	0.70	0.72	0.66	0.60	0.53 0.64±0.08
	2	0.60	0.61	0.57	0.54	0.51	0.57±0.04	0.54	0.56	0.55	0.51	0.47 0.52±0.04
	1	0.43	0.39	0.49	0.46	0.33	0.42±0.06	0.28	0.25	0.26	0.22	0.19 0.24±0.04
5	14	0.93	0.89	0.87	0.82	0.69	0.84±0.09	0.95	0.93	0.94	0.86	0.76 0.89±0.08
	8	0.88	0.86	0.83	0.69	0.53	0.76±0.15	0.90	0.90	0.89	0.76	0.69 0.83±0.10
	4	0.80	0.79	0.76	0.61	0.60	0.71±0.10	0.78	0.75	0.72	0.58	0.59 0.69±0.09
	2	0.68	0.63	0.53	0.49	0.56	0.58±0.08	0.56	0.56	0.50	0.46	0.52 0.52±0.04
	1	0.48	0.45	0.42	0.38	0.51	0.45±0.05	0.35	0.33	0.33	0.31	0.31 0.33±0.02
Avg.	14	0.91	0.91	0.89	0.83	0.75		0.95	0.95	0.92	0.86	0.75
	8	0.87	0.85	0.84	0.75	0.66		0.89	0.87	0.85	0.77	0.67
	4	0.76	0.74	0.71	0.64	0.61		0.74	0.74	0.70	0.63	0.57
	2	0.64	0.61	0.56	0.57	0.53		0.56	0.56	0.54	0.50	0.47
	1	0.46	0.43	0.43	0.44	0.41		0.31	0.30	0.30	0.28	0.25
Std	14	±0.02	±0.01	±0.02	±0.02	±0.07		±0.01	±0.01	±0.02	±0.02	±0.05
	8	±0.02	±0.01	±0.01	±0.04	±0.10		±0.02	±0.03	±0.03	±0.05	±0.05
	4	±0.02	±0.04	±0.07	±0.03	±0.05		±0.04	±0.04	±0.05	±0.06	±0.08
	2	±0.03	±0.03	±0.03	±0.07	±0.06		±0.03	±0.04	±0.06	±0.03	±0.06
	1	±0.02	±0.03	±0.05	±0.06	±0.10		±0.03	±0.04	±0.04	±0.03	±0.07

is the accuracy highest with fewer channels/instances?. An important insight from this paper is that more data is not necessarily more information, and that irrelevant data from certain channels can affect the performance of classifiers depending on the chosen task. Hence the channels selection approach depending on task takes relevance.

4.1 Discussion

The experiments presented were carried out with fewer channels and instances. However, to create a real machine-learning-based model will be necessary to select the best instances to use in a real application. To select those instances the greedy algorithm can be used, but depending on the task and the Subjects, the instances can differ. This means that the selection of instances is Subject-dependent.

In addition, the findings in the experiments suggest that EMD-based method can be used for feature extraction. However, for a certain task, it will be necessary

Table 2. Accuracies obtained with 3-NN in 5 different sessions using EMD and DWT

Session	Ch	EMD						DWT					
		30	20	10	5	3	Avg.	30	20	10	5	3	Avg.
1	14	0.86	0.87	0.88	0.80	0.64	0.81±0.10	0.89	0.89	0.88	0.83	0.76	0.85±0.06
	8	0.84	0.84	0.84	0.79	0.63	0.79±0.09	0.85	0.85	0.84	0.82	0.75	0.82±0.04
	4	0.79	0.79	0.79	0.75	0.64	0.75±0.06	0.75	0.75	0.76	0.73	0.68	0.73±0.03
	2	0.68	0.68	0.67	0.66	0.60	0.66±0.03	0.60	0.57	0.60	0.56	0.40	0.55±0.08
	1	0.51	0.51	0.53	0.55	0.48	0.52±0.03	0.36	0.35	0.33	0.29	0.23	0.31±0.06
2	14	0.85	0.85	0.83	0.72	0.67	0.78±0.08	0.90	0.90	0.84	0.81	0.72	0.83±0.07
	8	0.82	0.81	0.81	0.74	0.72	0.78±0.05	0.84	0.83	0.80	0.81	0.69	0.80±0.06
	4	0.72	0.69	0.67	0.60	0.48	0.63±0.10	0.74	0.73	0.71	0.70	0.60	0.70±0.08
	2	0.65	0.62	0.63	0.62	0.44	0.59±0.08	0.62	0.61	0.61	0.60	0.44	0.57±0.08
	1	0.48	0.49	0.48	0.48	0.40	0.46±0.04	0.34	0.33	0.36	0.42	0.25	0.34±0.06
3	14	0.87	0.88	0.83	0.78	0.72	0.81±0.07	0.89	0.88	0.86	0.80	0.65	0.81±0.10
	8	0.82	0.82	0.77	0.73	0.69	0.77±0.06	0.84	0.83	0.80	0.74	0.55	0.75±0.12
	4	0.71	0.72	0.69	0.69	0.53	0.67±0.08	0.66	0.65	0.63	0.62	0.45	0.60±0.09
	2	0.63	0.59	0.58	0.55	0.53	0.58±0.04	0.52	0.52	0.48	0.50	0.37	0.48±0.06
	1	0.52	0.51	0.44	0.40	0.32	0.44±0.08	0.29	0.27	0.22	0.30	0.20	0.26±0.05
4	14	0.87	0.85	0.82	0.78	0.63	0.79±0.10	0.86	0.83	0.80	0.73	0.55	0.75±0.13
	8	0.79	0.76	0.76	0.70	0.56	0.71±0.09	0.80	0.79	0.78	0.71	0.55	0.73±0.11
	4	0.70	0.68	0.64	0.67	0.57	0.65±0.05	0.69	0.70	0.68	0.58	0.45	0.62±0.11
	2	0.62	0.60	0.58	0.58	0.51	0.58±0.04	0.59	0.57	0.58	0.51	0.36	0.52±0.09
	1	0.49	0.49	0.55	0.52	0.41	0.44±0.05	0.26	0.26	0.22	0.25	0.21	0.24±0.02
5	14	0.87	0.85	0.85	0.76	0.67	0.80±0.09	0.89	0.89	0.85	0.78	0.73	0.83±0.07
	8	0.83	0.81	0.82	0.71	0.63	0.76±0.09	0.85	0.85	0.84	0.80	0.71	0.81±0.06
	4	0.75	0.74	0.76	0.69	0.63	0.71±0.06	0.72	0.72	0.74	0.69	0.63	0.70±0.04
	2	0.68	0.66	0.62	0.58	0.60	0.63±0.04	0.58	0.56	0.54	0.47	0.51	0.53±0.04
	1	0.53	0.55	0.54	0.49	0.44	0.51±0.05	0.35	0.36	0.36	0.30	0.31	0.34±0.03
Avg.	14	0.86	0.86	0.84	0.77	0.66		0.89	0.88	0.85	0.79	0.68	
	8	0.82	0.81	0.80	0.73	0.65		0.84	0.83	0.81	0.78	0.65	
	4	0.73	0.72	0.71	0.68	0.57		0.71	0.71	0.70	0.66	0.56	
	2	0.65	0.63	0.62	0.60	0.54		0.58	0.56	0.56	0.53	0.42	
	1	0.51	0.51	0.51	0.49	0.41		0.32	0.31	0.30	0.31	0.24	
Std	14	±0.01	±0.01	±0.02	±0.03	±0.04		±0.01	±0.03	±0.03	±0.04	±0.09	
	8	±0.02	±0.03	±0.03	±0.03	±0.06		±0.02	±0.02	±0.03	±0.05	±0.09	
	4	±0.04	±0.04	±0.06	±0.05	±0.07		±0.04	±0.04	±0.05	±0.06	±0.10	
	2	±0.03	±0.04	±0.04	±0.04	±0.07		±0.04	±0.03	±0.05	±0.05	±0.06	
	1	±0.02	±0.02	±0.05	±0.06	±0.06		±0.04	±0.05	±0.07	±0.06	±0.04	

to fix some problems related to the EMD method itself: the well-known mode-mixing problem and explore and test possible solutions available in the state-of-the-art to improve the methodology presented in this paper [26, 27].

5 Conclusions

In this paper, a comparison of EMD and DWT for Subject identification using EEG recordings of resting-states, was presented. In addition, the new FlexEEG Concept [16] was introduced and two of its main challenges were tackled: the use of fewer EEG channels and fewer instances to obtain a competitive/unhackable biometric system. In this paper and as a first attempt to reduce the number of channels, the *greedy* algorithm, was tested. The results obtained are promising and show that resting-states can be effectively used as a biometric marker for subject identification. The experiments conducted on a dataset obtained with an EMOTIV EPOC EEG device were reproduced by gradually reducing the

Table 3. Accuracies obtained with *Gaussian Naive Bayes* in 5 different sessions using EMD and DWT

Session	Ch	EMD					DWT					
		30	20	10	5	3	Avg.	30	20	10	5	3
1	14	0.88	0.89	0.90	0.78	0.23	0.74±0.29	0.92	0.93	0.90	0.63	0.15 0.71±0.34
	8	0.87	0.88	0.91	0.80	0.20	0.73±0.30	0.89	0.90	0.90	0.62	0.20 0.70±0.31
	4	0.82	0.84	0.85	0.68	0.24	0.69±0.26	0.84	0.85	0.82	0.55	0.15 0.64±0.30
	2	0.77	0.78	0.80	0.65	0.28	0.66±0.22	0.70	0.71	0.73	0.59	0.25 0.60±0.20
	1	0.61	0.63	0.64	0.54	0.29	0.54±0.14	0.47	0.49	0.51	0.41	0.23 0.42±0.11
2	14	0.87	0.85	0.84	0.70	0.32	0.72±0.23	0.92	0.93	0.90	0.64	0.31 0.74±0.27
	8	0.87	0.85	0.87	0.72	0.16	0.69±0.31	0.90	0.89	0.89	0.66	0.28 0.72±0.27
	4	0.79	0.81	0.75	0.63	0.16	0.63±0.27	0.79	0.82	0.80	0.61	0.32 0.67±0.21
	2	0.75	0.74	0.72	0.62	0.25	0.62±0.21	0.70	0.70	0.70	0.57	0.25 0.58±0.19
	1	0.60	0.57	0.58	0.52	0.25	0.51±0.14	0.42	0.43	0.46	0.45	0.21 0.39±0.10
3	14	0.92	0.89	0.87	0.73	0.24	0.73±0.28	0.93	0.94	0.91	0.61	0.25 0.73±0.30
	8	0.89	0.87	0.83	0.70	0.23	0.70±0.28	0.89	0.90	0.87	0.62	0.27 0.71±0.27
	4	0.82	0.80	0.76	0.64	0.17	0.64±0.27	0.75	0.74	0.69	0.56	0.25 0.60±0.21
	2	0.75	0.72	0.68	0.60	0.27	0.60±0.20	0.60	0.59	0.53	0.49	0.17 0.48±0.18
	1	0.60	0.59	0.51	0.46	0.21	0.47±0.16	0.42	0.37	0.36	0.33	0.19 0.33±0.09
4	14	0.90	0.89	0.88	0.69	0.19	0.71±0.30	0.90	0.89	0.86	0.66	0.19 0.70±0.30
	8	0.89	0.88	0.84	0.68	0.17	0.69±0.30	0.86	0.86	0.85	0.62	0.19 0.68±0.29
	4	0.81	0.81	0.75	0.62	0.21	0.64±0.25	0.75	0.75	0.74	0.54	0.20 0.60±0.24
	2	0.75	0.74	0.73	0.58	0.31	0.62±0.19	0.62	0.61	0.59	0.42	0.21 0.49±0.17
	1	0.60	0.61	0.60	0.46	0.25	0.51±0.15	0.39	0.41	0.39	0.31	0.16 0.33±0.10
5	14	0.89	0.88	0.84	0.54	0.20	0.67±0.30	0.88	0.90	0.85	0.62	0.11 0.67±0.34
	8	0.88	0.89	0.85	0.58	0.23	0.69±0.29	0.85	0.87	0.85	0.67	0.15 0.68±0.31
	4	0.83	0.83	0.80	0.58	0.19	0.64±0.28	0.76	0.76	0.74	0.62	0.21 0.62±0.23
	2	0.78	0.77	0.76	0.57	0.28	0.63±0.21	0.62	0.63	0.60	0.44	0.20 0.50±0.18
	1	0.63	0.65	0.62	0.51	0.31	0.54±0.14	0.45	0.46	0.46	0.38	0.27 0.40±0.08
Avg.	14	0.89	0.88	0.87	0.69	0.23		0.91	0.92	0.88	0.63	0.20
	8	0.88	0.87	0.86	0.70	0.20		0.88	0.88	0.87	0.64	0.22
	4	0.81	0.82	0.78	0.63	0.19		0.78	0.79	0.76	0.58	0.23
	2	0.76	0.75	0.74	0.60	0.28		0.65	0.65	0.63	0.50	0.22
	1	0.61	0.61	0.59	0.50	0.26		0.43	0.43	0.44	0.37	0.21
Std	14	±0.02	±0.02	±0.02	±0.09	±0.05		±0.02	±0.02	±0.03	±0.02	±0.08
	8	±0.01	±0.01	±0.03	±0.08	±0.03		±0.02	±0.02	±0.02	±0.02	±0.06
	4	±0.01	±0.02	±0.04	±0.04	±0.03		±0.04	±0.05	±0.05	±0.04	±0.06
	2	±0.02	±0.02	±0.04	±0.03	±0.02		±0.05	±0.05	±0.08	±0.08	±0.03
	1	±0.02	±0.03	±0.05	±0.04	±0.04		±0.03	±0.04	±0.06	±0.06	±0.04

number of channels and instances, in the first attempt towards a low-cost EEG-based subject identification. The relatively high accuracies obtained with fewer channels indicate a promising potential towards the FlexEEG concept.

When the methods based on EMD and DWT were compared, the difference in performance increases when using less information. EMD shows a robust and powerful property for feature extraction especially when fewer instances and fewer channels are used. This finding suggests that combining brain signals during resting-states with the use of EMD and the *Gaussian naive Bayes* classifier for low-density EEG, can materialize in a valuable biometric system based on a low-cost EEG-device.

A limitation of the method is that the use of the proposed method will require to find the smallest number of channels and instances to obtain similar accuracies as with high-density EEG. In general, a drastic fall of accuracy is observed from 10 to 5 instances and from 2 (or even 4) to 1 channel. In the

future, alternative approaches will be tested for channel selection to improve the performance obtained in these experiments.

Since the focus of this work is the use of EEG-recordings in real-time applications, it is necessary to analyze the computational complexity of the algorithms used to process a signal of size (N): In the case of DWT is $\mathcal{O}(N \log_2 N)$ [25], and for EMD $\mathcal{O}(N \log N)$ [28]. Recently, the real-time implementation of EMD has been reported by some authors [29,30], and among the challenges anticipated, techniques to distribute the computation and memory, and considerations about the benefits of cloud computing [5,31] have been discussed.

Future efforts will be directed towards the use of Multivariate Empirical Mode Decomposition [32] which is aimed at multichannel data analysis, but can also be explored for channel selection taking into account the findings of this work.

Acknowledgements. This work was supported by Enabling Technologies - NTNU, under the project “David versus Goliath: single-channel EEG unravels its power through adaptive signal analysis - FlexEEG”.

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. *IEEE Trans. Circ. Syst. Video Technol.* **14**(1), 4–20 (2004)
2. Jain, A.K., Ross, A., Uludag, U.: Biometric template security: challenges and solutions. In: Signal Processing Conference 13th European, pp. 1–4. IEEE (2005)
3. Valizadeh, S.A., Liem, F., Mérillat, S., Hänggi, J., Jäncke, L.: Identification of individual subjects on the basis of their brain anatomical features. *Sci. Rep.* **8**(1), 5611 (2018)
4. Moctezuma, L.A., Molinas, M.: EEG-based subjects identification based on biometrics of imagined speech using EMD. In: Submitted to The 11th International Conference on Brain Informatics (BI 2018) (2018)
5. Moctezuma, L.A., Molinas, M., García, A.A.T., Pineda, L.V., Carrillo, M.: Towards an API for EEG-based imagined speech classification. In: International Conference on Time Series and Forecasting (2018)
6. Moctezuma, L.A.: Distinción de estados de actividad e inactividad lingüística para interfaces cerebro computadora. Thesis project of Master Degree (2017)
7. Nishimoto, T., Azuma, Y., Morioka, H., Ishii, S.: Individual identification by resting-state EEG using common dictionary learning. In: International Conference on Artificial Neural Networks, pp. 199–207. Springer, Cham (2017)
8. Ashby, C., Bhatia, A., Tenore, F., Vogelstein, J.: Low-cost electroencephalogram (EEG) based authentication. In: 2011 5th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 442–445 (2011)
9. Palaniappan, R.: Electroencephalogram signals from imagined activities: a novel biometric identifier for a small population. In: International Conference on Intelligent Data Engineering and Automated Learning, pp. 604–611. Springer, Berlin, Heidelberg (2006)
10. Jayaratne, I., Cohen, M., Amarakeerthi, S.: BrainID: development of an EEG-based biometric authentication system. In: 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1–6 (2016)

11. Jayarathne, I., Cohen, M., Amarakeerthi, S.: Survey of EEG-based biometric authentication. In: 2017 IEEE 8th International Conference on Awareness Science and Technology (iCAST), pp. 324–329 (2017)
12. Del Pozo-Banos, M., Alonso, J.B., Ticay-Rivas, J.R., Travieso, C.M.: Electroencephalogram subject identification: a review. *Expert Syst. Appl.* **41**(15), 6537–6554 (2014)
13. Elman, L.B., McCluskey, L.: Clinical features of amyotrophic lateral sclerosis and other forms of motor neuron disease. Up-to-date, p. 23. Wolters Kluwer Health, Waltham (2012)
14. Feller, T.G., Jones, R.E., Netsky, M.G.: Amyotrophic lateral sclerosis and sensory changes. *Virginia Med. Mon.* **93**(6), 328 (1966)
15. Ma, L., Minett, J.W., Blu, T., Wang, W.S.: Resting state EEG-based biometrics for individual identification using convolutional neural networks. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2848–2851 (2015)
16. Molinas, M., Van der Meer, A., Skjærvold, N.K., Lundheim, L.: David versus Goliath: single-channel EEG unravels its power through adaptive signal analysis - FlexEEG. Research project (2018)
17. Xiong, J., Ma, L., Wang, B., Narayana, S., Eugene, E.P., Egan, G.F., Fox, P.T.: Long-term motor training induced changes in regional cerebral blood flow in both task and resting states. *Neuroimage* **45**(1), 75–82 (2009)
18. Golovanov, E.V., Yamamoto, S., Reis, D.J.: Spontaneous waves of cerebral blood flow associated with a pattern of electrocortical activity. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **266**(1), R204–R214 (1994)
19. Mantini, D., Perrucci, M.G., Del Gratta, C., Romani, G.L., Corbetta, M.: Electrophysiological signatures of resting state networks in the human brain. *Proc. Nat. Acad. Sci.* **104**(32), 13170–13175 (2007)
20. Jasper, H.: Report of the committee on methods of clinical examination in electroencephalography. *Electroencephalogr. Clin. Neurophysiol.* **10**, 370–375 (1958)
21. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to algorithms. In: Greedy Algorithms. MIT press, Cambridge (2001)
22. Torres-García, A.A., Reyes-García, C.A., Villaseñor-Pineda, L., Ramírez-Cortís, J.M.: Análisis de señales electroencefalográficas para la clasificación de habla imaginada. *Revista mexicana de ingeniería biomédica* **34**(1), 23–39 (2013)
23. Boutana, D., Benidir, M., Barkat, B.: On the selection of intrinsic mode function in EMD method: application on heart sound signal. In: 2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL), pp. 1–5 (2010)
24. Rish, I., Hellerstein, J., Thathachar, J.: An analysis of data characteristics that affect naive Bayes performance. *IBM TJ Watson Research Center* **30**, (2001)
25. Averbuch, A.Z., Zheludev, V.A.: Construction of biorthogonal discrete wavelet transforms using interpolatory splines. *Appl. Comput. Harmonic Anal.* **12**(1), 25–56 (2002)
26. Gao, Y., Ge, G., Sheng, Z., Sang, E.: Analysis and solution to the mode mixing phenomenon in EMD. In: Congress on Image and Signal Processing, CISIP'08, vol. 5, pp. 223–227 (2008)
27. Fosso, O.B., Molinas, M.: Method for Mode Mixing Separation in Empirical Mode Decomposition. arXiv preprint [arXiv:1709.05547](https://arxiv.org/abs/1709.05547) (2017)
28. Wang, Y.-H., Yeh, C.-H., Young, H.-W.V., Hu, K., Lo, M.-T.: On the computational complexity of the empirical mode decomposition algorithm. *Phys. A Stat. Mech. Appl.* **400**, 159–167 (2014)

29. Fontugne, R., Borgnat, P., Flandrin, P.: Online empirical mode decomposition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4306–4310 (2017)
30. Faltermeier, R., Zeiler, A., Keck, I.R., Tom, A.M., Brawanski, A., Lang, E.W.: Sliding empirical mode decomposition. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2010)
31. Mahmudova, S.: Analysis of biometric authentication methods of users in clouds. *Int. J. Adv. Eng. Technol.* **1**(5), 14–17 (2017)
32. Rehman, N., Mandic, D.P.: Multivariate empirical mode decomposition. *Proc. R. Soc. Lond. A* **466**(2117), 1291–1302 (2010)



Early Detection of Mirai-Like IoT Bots in Large-Scale Networks through Sub-sampled Packet Traffic Analysis

Ayush Kumar^(✉) and Teng Joon Lim

National University of Singapore, Singapore 119077, Singapore
ayush.kumar@u.nus.edu, eletlj@nus.edu.sg

Abstract. The widespread adoption of Internet of Things has led to many security issues. Recently, there have been malware attacks on IoT devices, the most prominent one being that of Mirai. IoT devices such as IP cameras, DVRs and routers were compromised by the Mirai malware and later large-scale DDoS attacks were propagated using those infected devices (bots) in October 2016. In this research, we develop a network-based algorithm which can be used to detect IoT bots infected by Mirai or similar malware in large-scale networks (e.g. ISP network). The algorithm particularly targets bots scanning the network for vulnerable devices since the typical scanning phase for botnets lasts for months and the bots can be detected much before they are involved in an actual attack. We analyze the unique signatures of the Mirai malware to identify its presence in an IoT device. Further, to optimize the usage of computational resources, we use a two-dimensional (2D) packet sampling approach, wherein we sample the packets transmitted by IoT devices both across time and across the devices. Leveraging the Mirai signatures identified and the 2D packet sampling approach, a bot detection algorithm is proposed. We use testbed measurements and simulations to study the relationship between bot detection delays and the sampling frequencies for device packets. Subsequently, we derive insights from the obtained results and use them to design our proposed bot detection algorithm. Finally, we discuss the deployment of our bot detection algorithm and the countermeasures which can be taken post detection.

Keywords: Internet of Things · IoT · Malware · Mirai · Botnet · Bot Detection

1 Introduction

The Internet of things (IoT) [1] refers to the network of low-power, limited processing capability sensing devices which can send/receive data to/from other devices using wireless technologies such as RFID (Radio Frequency Identification), Zigbee, WiFi, Bluetooth, 3G/4G etc. IoT devices are being deployed in a number of applications such as wearables, home automation, smart grids,

environmental monitoring, infrastructure management, industrial automation, agricultural automation, healthcare and smart cities. Some of the popular platforms for IoT are Samsung SmartThings (consumer IoT for device management) and Amazon Web Services IoT, Microsoft Azure IoT, Google Cloud Platform (enterprise IoT for cloud storage and data analytics). The number of IoT devices deployed globally by 2020 is expected to be in the range of 20–30 billion [2]. The number of devices has been increasing steadily (albeit at a slower rate than some earlier generous predictions), and this trend is expected to hold in the future.

IoT devices are being increasingly targeted by hackers using malware (malicious software) as they are easier to infect than conventional computers for the following reasons [3–5]:

- There are many legacy IoT devices connected to the Internet with no security updates.
- Security is given a low priority within the development cycle of IoT devices.
- Implementing conventional cryptography in IoT devices is computationally expensive due to processing power and memory constraints.
- Many IoT devices have weak login credentials either provided by the manufacturer or configured by users.
- IoT device manufacturers sometimes leave *backdoors* (such as an open port) to provide support for the device remotely.
- Often, consumer IoT devices are connected to the Internet without going through a firewall.

In a widely publicized attack, the IoT malware *Mirai* was used to propagate the biggest DDoS (Distributed Denial-of-Service) attack on record on October 21, 2016. The attack targeted the Dyn DNS (Domain Name Service) servers [6] and generated an attack throughput of the order of 1.2 Tbps. It disabled major internet services such as Amazon, Twitter and Netflix. The attackers had infected IoT devices such as IP cameras and DVR recorders with Mirai, thereby creating an army of bots (botnet) to take part in the DDoS attack. Apart from Mirai, there are other IoT malware which operate using a similar brute force technique of scanning random IP addresses for open ports and attempting to login using a built-in dictionary of commonly used credentials. BASHLITE [7], Remaiten [8], Hajime [9] are some examples of these IoT malware.

Bots compromised by Mirai or similar IoT malware can be used for DDoS attacks, phishing and spamming [10]. These attacks can cause network downtime for long periods which may lead to financial loss to network companies, and leak users' confidential data. McAfee reported in April 2017 [11] that about 2.5 million IoT devices were infected by Mirai in late 2016. Bitdefender mentioned in its blog in September 2017 [12] that researchers had estimated at least 100,000 devices infected by Mirai or similar malware revealed daily through telnet scanning telemetry data. Further, many of the infected devices are expected to remain infected for a long time. Therefore, there is a substantial motivation for detecting these IoT bots and taking appropriate action against them so that they are unable to cause any further damage.

As pointed out in [13], attempting to ensure that all IoT devices are secure-by-construction is futile as there will always be insecure devices (with patched and unpatched vulnerabilities) connected to the Internet due to the scale and diversity of IoT devices and vendors. Moreover, considering the lack of full-fledged operating systems, low power requirements, resource constraints and presence of legacy devices, it is practically unfeasible to deploy traditional host-based detection and prevention mechanisms such as antivirus, firewalls for IoT devices. Therefore, it becomes imperative that the security mechanisms for the IoT ecosystem are designed to be network-based rather than host-based.

In this research, we propose a network-based algorithm which can be used to detect IoT bots infected by Mirai-like malware (which use port-based scanning) in large-scale networks. Bots scanning the network for vulnerable devices are targeted in particular by our algorithm. This is because the scanning and propagation phase of the botnet life-cycle stretches over many months and we can detect and isolate the bots before they can participate in an actual attack such as DDoS. If the DDoS attack has already occurred (due to a botnet), detecting the attack itself is not that difficult and there are already existing methods both in literature and industry to defend against such attacks. Moreover, our algorithm is practical in terms of utilization of computational resources (such as CPU processing power, memory). For example, ISP (Internet Service Provider) network operators can use the proposed algorithm to identify infected IoT devices in their network. The operators can then take suitable countermeasures such as blocking the traffic originating from IoT bots and notifying the local network administrators. Actions that can be taken post bot detection are further discussed in a later section. The major contributions of this paper are listed below:

1. We have analyzed the traffic signatures produced by Mirai malware infecting IoT devices through testbed experiments. Further, we have identified specific signatures which can be used to positively detect the presence of Mirai and similar malware in IoT devices. These signatures are similar to the observations reported in [14] based on their analysis of the Mirai source code.
2. We have proposed an algorithm to detect Mirai-like IoT malware bots in large-scale networks. The algorithm is based on a novel two dimensional sampling approach where the device packets are sampled across time as well as across the devices.

The rest of the contents of this paper are organized as follows. In Sect. 2, we review few prominent works on detecting botnets exploiting CnC communication features and intrusion detection systems for IoT. Subsequently, in Sect. 3, we explain the operation of Mirai, extract important features from the traffic generated by Mirai bots in a testbed and present a detailed analysis of those features towards detecting Mirai-like bots. Section 4 formulates the optimization problem resulting from detection of IoT bots in large-scale networks along with the constraints imposed by limited computational resources followed by the proposed bot detection algorithm. The algorithm is numerically evaluated and the results are presented in Sect. 5. Finally, in Sect. 6, the implementation of the

proposed IoT bot detection algorithm in a real-world network is discussed as well as the mitigating actions that can be taken post detection.

2 Related Work

There are several works in the literature on detecting botnets using their CnC communication features. We list a few prominent ones in this section. The authors in [15] present machine-learning based classification methods to detect CnC traffic of IRC (Internet Relay Chat) botnets by differentiating between IRC and non-IRC traffic and then differentiating between bot and real IRC traffic. Bothunter [16] builds a *bot infection dialog model* using the network communication flows between internal hosts and external entities during successful bot infections. Three bot-specific sensors are constructed based on the dialog model and correlation is performed between inbound intrusion/scan alarms and the infection dialog model to generate a consolidated report. Spatio-temporal similarities between bots in a botnet in terms of bot-CnC coordinated activities are captured from network traffic and leveraged towards botnet detection in a local area network in Botsniffer [17]. In BotMiner [18], the authors have proposed a botnet detection system which clusters similar CnC communication traffic and similar malicious activity traffic, and uses cross cluster correlation to detect bots in a monitored network. It is claimed to be independent of CnC protocol and structure with no requirement of a priori knowledge about the botnets. A system for detecting covert P2P (Peer-to-Peer) botnets has been proposed in [19]. After extracting the statistical CnC communication features for P2P botnets, the botnet detection system utilizes them to distinguish between legitimate and malicious P2P traffic.

There has also been some research on intrusion detection and anomaly detection systems for IoT. A whitelist-based intrusion detection system for IoT devices (Heimdall) has been presented in [20]. Heimdall is based on dynamic profile learning and is designed to work on routers acting as gateways for IoT devices. The authors in [21] propose an intrusion detection model for IoT backbone networks leveraging two-layer dimension reduction and two-tier classification techniques to detect U2R (User-to-Root) and R2L (Remote-to-Local) attacks. In a recently published paper [22], deep-autoencoders based anomaly detection has been used to detect attacks launched from IoT botnets. The method consists of extraction of statistical features from behavioral snapshots of normal IoT device traffic captures, training of a deep learning-based autoencoder (for each IoT device) on the extracted features and comparison of the reconstruction error for traffic observations with a threshold for normal-anomalous classification. The proposed detection method was evaluated on Mirai and BASHLITE botnets formed using commercial IoT devices.

While a number of above anomaly detection works leverage ML (machine learning)-based approaches, there are several issues associated with them [23]. One of the major issues is the occurrence of false positives. Even a small percentage of false positives, e.g. 1% which is considered acceptable in academic

research on anomaly detection, can lead to thousands of alerts per day based on the traffic volume processed [24]. Both false positives and false negatives have costs (e.g. financial expenses for an organization) associated with them, with the cost associated with false negatives typically being much higher. Second, many research works on anomaly detection using ML fail to explain why a particular ML algorithm would perform well in the system under consideration. Third, many ML algorithms are suitable for offline batch operations rather than low-latency real-time detection. Finally, instead of starting with the premise of using ML approach for a detection task which is a common flaw in anomaly detection research, one should carry out a neutral evaluation of all the available tools for the task and then decide on the most appropriate one.

Our work addresses a few important gaps in the literature when it comes to distinguishing between legitimate and botnet IoT traffic. First, almost all the works cited above on detecting botnets using their CnC communication features [15–18] utilize all the packets transmitted by all the devices in a monitored network for a specific time period towards designing a botnet detection solution. This approach is highly impractical if the resulting solution is to be deployed for IoT devices in real world networks. The reason is that processing all the packets for all devices in a large network would require a lot of computational resources as illustrated in Sect. 4.1. Second, our focus is not only on detecting bots employing IRC-based CnC communications as done in [15]. The bot detection algorithm proposed by us in Sect. 4.1 is independent of the bot-CnC communication protocol. Third, we do not aim to detect botnets (networks of bots) but instead, individual bots. Therefore, we don't require computationally expensive clustering algorithms as used in [17, 18].

Fourth, we do not extract CnC communication features and use them to identify bot-CnC communications as done in [17–19]. This is because we aim to detect bots infected by Mirai-like IoT malware, towards which much simpler features can be used as discussed in Sect. 3.3. Fifth, unlike [22], we aim to detect IoT bots much before the actual attack, during the scanning phase itself as explained in Sect. 4. Finally, most of the above cited works use quantifiers such as detection rate and false positive rates to evaluate the performance of their proposed botnet detection solutions. Instead, we use a quantity called average detection delay (defined in Sect. 4.1) for the performance evaluation of our proposed bot detection solution since the features used by our solution eliminate the possibility of inaccurate detections or false positives. To the best of our knowledge, there are no existing papers on detecting IoT bots compromised by Mirai or its variants which exhibit port-based SYN scanning behavior.

3 Mirai Traffic Analysis

Detecting IoT devices compromised by Mirai-like malware requires us to analyze the packet traffic generated by those devices and extract some features to aid us in detection. In this section, we begin with a brief description the operation of Mirai to make the readers familiar with some of the related terms. Later, we

present a testbed that we use to emulate IoT devices, infect them with Mirai and capture the packet traffic generated from them. Finally, we present the extracted features and analyze them in detail with respect to identifying Mirai bots.

3.1 Mirai Operation

The Mirai [25] setup consists of three major components: *bot*, *scanListen/loading* server, and the *CnC* (Command-and-Control) server. The *CnC* server also functions as a MySQL [26] database server. User accounts can be created in this database for customers who wish to hire DDoS-as-a-service. The operation of Mirai is illustrated in Fig. 1. Once an IoT device is infected with Mirai (and becomes a bot), it first attempts to connect to the listening *CnC* server by resolving its domain name and opening a socket connection. Thereafter, it starts scanning the network by sending SYN packets to random IP addresses and waiting for them to respond. This process may take a long time since the bot has to go through a large number of IP addresses. Once it finds a vulnerable device with a TELNET port open, it attempts to open a socket connection to that device and emulates the TELNET protocol. Then it attempts to login using a list of default credentials and if working credential is found, it reports the IP address of the discovered device and the working TELNET login credentials to the listening *scanListen* server. The *scanListen* server sends that information to the *loader* which again logs in to the discovered device using the details received from the *scanListen* server. Once logged in, the *loader* downloads the Mirai bot binary to that device and the new bot connects to the *CnC* server and starts scanning the network.

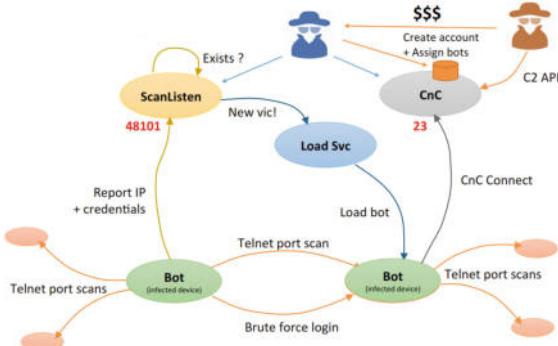


Fig. 1. Operation of various components of Mirai (Source Radware [27])

3.2 Testbed Description

The testbed shown in Fig. 2 was configured on an isolated computing cluster. Each cluster node has two Intel Xeon E5-2620 processors, 64 GB DDR4 ECC

memeory and runs Ubuntu 14.04 LTS standard image. The testbed consists of a local authoritative DNS server, a CnC (Command-and-Control) server and a server for scanListen and loading utility, all connected to a single LAN. The IoT gateways are connected to the above LAN through routers and behind the gateways are QEMU [28]-emulated IoT devices (Raspberry Pi). We chose this gateway-IoT device topology since it is used in a number of IoT deployments (such as IP cameras, smart lighting devices, wearables etc.). The testbed also includes few non-IoT devices (PCs) to reflect real-world networks. As per our information, this is the first controlled testbed to simulate the true behavior of Mirai malware. It can be modified to add more nodes, study a different network topology and test more advanced versions or derivatives of Mirai malware.

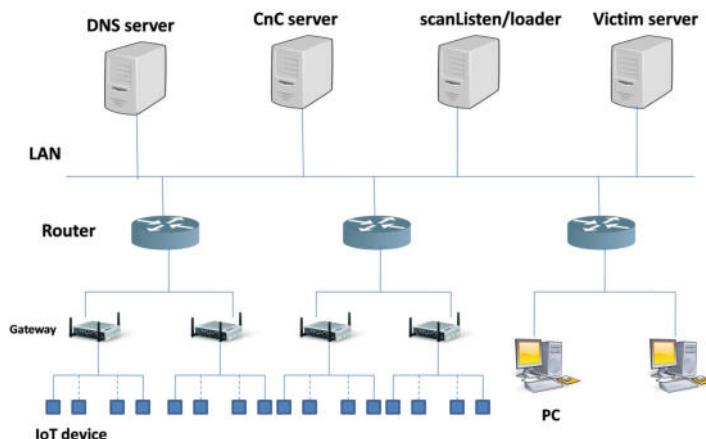


Fig. 2. Testbed used to simulate Mirai behavior

3.3 Mirai Traffic Features

We infected the emulated IoT devices in our testbed with Mirai and captured a total of 1,583,623 packets transmitted by the devices. An analysis of the captured packets reveals the following features/signatures:

- The scanning packets are all TCP SYN (synchronization) packets.
 - The destination port numbers of scanning packets are distributed as ~90% port 23 and ~10% port 2323. No other port numbers are observed.
 - There is a periodic exchange of keep alive messages (PSH+ACK) between the bot and the CnC server. PSH refers to a push message and ACK refers to acknowledgement.

Both ports 23 and 2323 are assigned for TELNET applications [29, 30]. The TELNET [31] protocol is used for bidirectional byte-oriented communication. In

the most widely used implementation of TELNET, a user with a terminal and running a TELNET client program, accesses a remote host running a TELNET server by requesting a connection to the remote host and logging in by providing its credentials. The most common application of TELNET is for configuring network devices such as routers. Now, IoT devices operate by continuously transmitting sensed data to and receiving commands from cloud servers through a gateway over a secure communication channel without external human input [32]. We claim that an IoT device is unlikely to be used to access or configure another device using TELNET, and therefore in the absence of malware infection, IoT devices should not open TELNET connections to any other device.

To verify our claim that uninfected IoT devices are not expected to open TELNET connections, the following experiment was conducted. We configured a Raspberry Pi 3 (Model B+) to act as a gateway and connected it to several real-world IoT devices such as IP cameras (D-Link), motion sensors (D-Link), smart bulbs (Philips Hue), smart switches (WeMo) and smart plugs (TPLink). We left the devices connected for a long time and for each device type mentioned above, we captured around 10,000 packets per device at the gateway interface. Later, the captured packets were analysed using Wireshark [33] and no SYN packets with destination ports 23 or 2323 were found. Thus, if a SYN packet from an IoT device with destination port number 23 or 2323 is received, it is sufficient evidence to conclude with certainty that the IoT device is infected with a Mirai-like malware. The above experiment also help us to rule out *false positives*, if any at all, if we use the identified scanning traffic signatures, which is a substantial advantage when it comes to practical intrusion detection.

The third Mirai signature related to keep-alive messages is not required since the port-scanning signatures is sufficient for detection with certainty. We may require the third signature to detect more advanced malware which do not use TELNET port-based scanning. It needs to be emphasized here that the TELNET port-scanning signatures can be used to identify not only bots infected by Mirai but also other Mirai-like malware such as BASHLITE, Remaiten, Hajime etc. which employ similar TELNET port brute forcing technique.

4 Mirai-Like IoT Malware Bot Detection

The bot scanning traffic analyzed in the previous section cannot be detected using simple firewalls. Since IoT devices are usually resource-constrained, they do not have firewalls installed on them. Moreover, network-level firewalls (protecting computers in a LAN/WAN/intranet) are not configured to block TELNET traffic which in most cases may be legitimate. In this section, we formulate the optimization problem arising out of detecting IoT bots in large-scale networks with the accompanying computational resource constraints. Further, we propose an algorithm for bot detection based on our analysis.

4.1 Formulation of Optimization Problem

Even though receiving a SYN packet with destination port number 23 or 2323 in its TCP (Transmission Control Protocol) header is sufficient to identify the transmitting IoT device as infected, we cannot strip off the TCP headers and check the encapsulated TCP flags and destination port numbers for all the packets transmitted by all the IoT devices in a network as this would require a lot of computational resources (both processing power and memory) from the bot detection device that captures and processes the IoT device packets. To give an example, the total number of IoT devices being used in the U.S. stands at 715 million [34]. Given that there are 12 major ISPs operating in U.S. [35], assuming equal number of IoT devices being used in each ISP network yields 59.58 million devices per ISP. IEEE 802.15.4 standard [36] which forms the basis of most IoT communication protocols allows a peak data rate of 250 kbit/s. The peak total IoT device data rate for an ISP network can thus be estimated as 14,895 billion bits/s (IoT devices are considered to be *always ON* once installed).

To send or receive 1 bit/s of TCP/IP, 1Hz of processing speed is required as a general rule of thumb. However, as shown in [37], this rule doesn't always hold and the Hz/bps ratio increases upto 6–7 times for small data transfers (payload size of the order of \approx 64 bytes) as compared to larger transfers. In fact, the maximum payload size allowed by IEEE 802.15.4 link headers is 81 bytes. Further, the Hz/bps ratio increases when one goes to higher CPU speeds. Now, socket receive processing can take upto 23% of the total processing required for TCP processing for small transfer sizes. Hence, the processing speed required for socket operations of TCP flag and destination port lookup at just 10% of the estimated peak total IoT device data rate for an ISP network can be calculated as 2398 GHz. This translates to a requirement of nearly 480 additional 2.5 GHz dual-core processors for a single ISP, just for detecting IoT bots. This represents a significant investment from ISP companies. Moreover, as the number of devices is increasing steadily with time, the above investment is only bound to grow.

Therefore, for our bot detection problem, we propose to sample only a fraction of the IoT devices per unit time for TCP processing. This will reduce the number of IoT packets that need to be processed by the TCP stack, bringing down the the computational resources required. However, this approach has the drawback that we may miss the scanning packets due to the sub-sampling operation. This leads to the formulation of the following optimization problem to detect infected devices.

Our objective in this optimization problem is to minimize the cost associated with the delay in detecting a compromised device. We define *average detection delay* (T_D) as the average time between the first occurrence of a scanning packet and the positive conclusion that the originating device is infected. Now, some IoT devices in a network are easier to infect with malware than others. Therefore, we split the IoT devices into two categories: *vulnerable* and *non-vulnerable* devices.

Vulnerable devices are the devices which are easier to get successfully infected with Mirai-like malware and added to the botnet. The devices other than vulnerable ones are non-vulnerable devices. For example, personal IoT devices installed at homes can be deemed as vulnerable since they are less likely to be behind a firewall (host-level firewalls not feasible on IoT devices due to resource constraints) and more likely to have their TELNET ports open (often owners buy cheap devices in which the manufacturer has left TELNET port open for remote configuration etc.). IoT devices installed in enterprise/industrial/government networks can be categorized as non-vulnerable since most likely, they would be behind a network-level firewall (blocking access to insecure TELNET connections) and they are much less likely to have to have their TELNET ports open (due to organizational IT security policies).

We define the *sampling frequency for an IoT device* as the fraction of the time when that device is selected for monitoring for possible infection. We also define the *sampling matrix*, Σ as a matrix with columns representing devices and rows representing the packets transmitted by those devices. An element of Σ is equal to 1 when the corresponding packet has been sampled and equal to 0 when the corresponding packet has not been sampled.

Further, our optimization problem imposes the following constraints that need to be satisfied:

- The sampling frequency for a vulnerable device (f_n^v) should be greater than the sampling frequency for a non-vulnerable device (f_n^{nv}). This is because vulnerable devices are more likely to be attacked than non-vulnerable devices and hence they need to be more frequently monitored.
- The total number of vulnerable and non-vulnerable devices selected within a certain time period ($\rho_v f_n^v T + \rho_{nv} f_n^{nv} T$) should not exceed a maximum number ($f_n^{max} T$), where ρ_v and ρ_{nv} are the fractions of total number of devices that are vulnerable and non-vulnerable respectively. This is to limit the utilization of computational resources for if the total number of selected devices is more than an upper bound, it may require significant amounts of processing power defeating the purpose of packet sub-sampling.
- The maximum number of vulnerable devices selected at any time should have an upper bound (N_v^{max}). Similarly, the maximum number of non-vulnerable devices selected at any time should have an upper bound (N_{nv}^{max}). This is again to place a bound on computational resources utilization.
- After a certain number of sampling time units (T), every device (in the set of all devices, Ω_N) should be covered by the sampling process. This is to ensure that every device is checked for malware infection within a certain time duration or else few devices which are infected may be missed by the sampling process.

We propose to minimize the cost associated with the average detection delay while satisfying the above constraints as follows:

$$\begin{aligned}
 & \underset{\Sigma, f_n^v, f_n^{nv}}{\text{minimize}} \quad \alpha T_D(f_n^v, f_n^{nv}, Y_v, Y_{nv}) \\
 & \text{subject to} \quad f_n^v > f_n^{nv} \\
 & \quad \rho_v f_n^v + \rho_{nv} f_n^{nv} < f_n^{max} \\
 & \quad \max[N_v\{\Sigma\}] < N_v^{max} \\
 & \quad \max[N_{nv}\{\Sigma\}] < N_{nv}^{max} \\
 & \quad \bigcup_{t=t_{start}}^{t_{start}+T} dev_set(\Sigma, t) = \Omega_N
 \end{aligned}$$

where α is defined as the cost incurred by the bot detection algorithm due to a unit average detection delay, $N_v\{\cdot\}$, $N_{nv}\{\cdot\}$ denote the number of vulnerable and non-vulnerable devices selected in Σ at any point of time, Y_v is the set of vulnerable devices, Y_{nv} is the set of non-vulnerable devices, and $dev_set(\cdot)$ is a function that outputs the set of devices sampled in Σ at a time t . It is to be noted that the above optimization problem is a combinatorial one and it is computationally hard to find an optimal solution [38]. Hence, we devise a method to numerically solve the optimization problem. The results obtained from the numerical analysis are explained in Sect. 5. Based on our findings through the formulation of optimization problem, we have proposed an algorithm for detecting IoT bots (shown in Algorithm 1) which is practical in terms of lower number of packets that need to be monitored for infected device detection. The values for f_n^v and f_n^{nv} to be used while designing our algorithm will be discussed in our numerical analysis.

5 Evaluation of Proposed Algorithm

In this section, we analyze the behavior of average detection delay for vulnerable and non-vulnerable devices with varying sampling rates. A few important background details are presented below:

- The set of attacked devices, Φ is selected based on the assumed probability model for malware attack on vulnerable and non-vulnerable devices. For example, we can assume the probability of attack on vulnerable devices within a given time duration (N_p packets' transmission) as p_1 and that on non-vulnerable devices as p_2 .
- The sampling matrix, Σ used in our evaluation has a staggered structure and may be visualized as in Fig. 3. Since the sampling frequency for vulnerable devices is greater than that for non-vulnerable devices, the portion of Σ containing packets transmitted by vulnerable devices has a more dense distribution of 1s than that for non-vulnerable devices. The structure of the matrix also ensures that every device is sampled after a certain number of sampling time units as required by one of the constraints in the optimization problem presented in Sect. 4.1.

Algorithm 1 IoT Bot Detection Algorithm

```

1: Initialize  $\Sigma$ , NUM_PKTS, t.
2: for  $pktcnt = 1$  to NUM_PKTS do
3:   if  $src\_dev(recv\_pkt) \notin list\_dev$  then
4:      $add\_dev\_to\_list(src\_dev(recv\_pkt), list\_dev)$ 
5:   end if
6:    $add\_pkt\_to\_buf(recv\_pkt, dev\_buf(src\_dev(recv\_pkt)))$ 
7:    $pktcnt=pktcnt+1$ 
8: end for
9: while TRUE do
10:    $sel\_dev\_set=dev\_set(\Sigma,t)$ 
11:   for  $i = 1$  to length( $sel\_dev\_set$ ) do
12:      $sampled\_pkts(t,:)=dev\_buf(sel\_dev\_set(i), CURRENT\_PKT)$ 
13:   end for
14:   for  $j = 1$  to length( $sampled\_pkts(t,:)$ ) do
15:     if Check_TCP_flag( $sampled\_pkts(t,j)$ ) = SYN &
        Check_dst_port( $sampled\_pkts(t,j)$ ) = 23 OR 2323 then
16:        $Bot\_detected(src\_dev(sampled\_pkts(t,j))) = TRUE$ 
17:     end if
18:   end for
19:    $t=t+1$ 
20: end while

```

1	0	1	0	1	0	1	0	1	.	.
0	1	0	1	0	1	0	1	0	.	.
1	0	1	0	1	0	1	0	1	.	.
0	1	0	1	0	1	0	1	0	.	.
1	0	1	0	1	0	1	0	1	.	.
0	1	0	1	0	1	0	1	0	.	.
.
.
-	-	-	-	-	-	-	-	-	-	-
1	0	0	0	1	0	0	0	1	.	.
0	1	0	0	0	1	0	0	0	.	.
0	0	1	0	0	0	1	0	0	.	.
0	0	0	1	0	0	0	1	0	.	.
1	0	0	0	1	0	0	0	1	.	.
0	1	0	0	0	1	0	0	0	.	.
.

Vulnerable devices

Non-vulnerable devices

Fig. 3. Sampling matrix example

- We form a *scanning* matrix with size as (number of IoT devices) \times (number of packets transmitted). The matrix uses 0 to represent a normal IoT device packet and 1 to represent a malware scanning packet. Only the devices in Φ would have 1s in their corresponding rows in the scanning matrix.
- The elements where the scanning and the sampling matrices are both 1 represent detected scanning packets. This is because the matching elements would only be present where the scanning packet transmitted by an attacked device has been selected by the sampling process.

Moreover, we need to form a statistical model for scanning packet arrivals in the scanning matrix. Towards this, we used one of our emulated IoT devices and established a video streaming server to simulate the operation of an IP camera (IoT device used in Mirai attack on Dyn). Another emulated IoT device acted as a client connected to the video stream. The other emulated devices were configured to have their TELNET port number 23 open and listening for connections. Subsequently, we infected the video streaming device with Mirai and captured the transmitted packets at its gateway interface using Wireshark. Our observations from the packet capture are listed below:

- The video streaming packets are transmitted almost continuously. The transmission is interrupted only by bot-CNC server communication packets, scanning packets and some other types of packets such as ARP (Address Resolution Protocol).
- The bot scanning packets are sometimes transmitted within short intervals and at other times they are transmitted far apart as shown in Fig. 4.

Based on the above empirical observations, we model the scanning packet arrivals as a Poisson process, i.e., the inter-packet arrival times for scanning packets are exponentially distributed with the average packet arrival rate calculated from the testbed measurements. At all other times, we assume that normal IoT traffic is transmitted, again based on above observations.

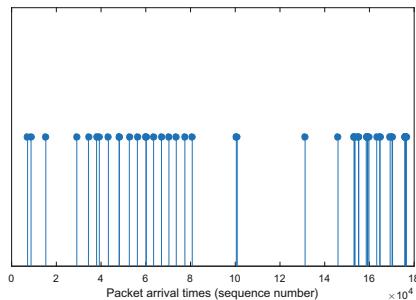


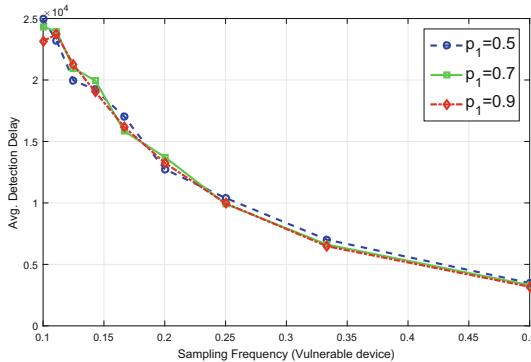
Fig. 4. Arrival times of scanning packets

The values assumed for the various parameters in our analysis are shown in Table 1.

The plot for average detection delay versus sampling frequency for different values of attack probability on vulnerable devices (p_1) is shown in Fig. 5. The detection delay values are averaged over all the detected devices as well as over a number of trial runs (1000). The units of average detection delay are in *number of packets elapsed* while the units of sampling frequency are in *per packet elapsed*. It can be observed that the average detection delay decreases almost exponentially with increasing sampling frequency. This behavior can be intuitively explained

Table 1. Parameter values assumed in numerical analysis

Parameter	Value
N_v^{max}	40
N_{nv}^{max}	80
f_n^{max}	0.5
N_p	50
Total no. of IoT devices	100
% age of vulnerable devices	40
No. of packets transmitted per device	100,000
Avg. rate of arrival of scanning packets (per packet elapsed)	3386

**Fig. 5.** Average detection delay versus sampling frequency plot for vulnerable devices

as follows. Increasing the sampling frequency means that the vulnerable devices are sampled much more frequently, which in turn increases the likelihood of sampling the scanning packets transmitted by infected vulnerable devices. Once a scanning packet is sampled, it can be positively concluded that the corresponding source device is infected as discussed in Sect. 3.3. Hence, an increase in the likelihood of sampling scanning packets should lead to a decrease in the average detection delay as defined in Sect. 4.1. Further, it can also be noted from the plot that increasing the sampling frequency beyond a certain value (e.g. ‘0.33’ for $p_1 = 0.5$) leads to slower reduction in average detection delay. This suggests that while designing the proposed Algorithm 1, the sampling frequency for vulnerable devices should be selected towards the upper half of the range of available values but not too high since higher sampling frequencies will not result in more benefit in terms of decrease in average detection delay. Instead, sampling frequencies which are too high may lead to greater consumption of computational resources.

One may observe that the average detection delay values decrease slightly as the attack probability increases. This is expected since an increase in attack prob-

ability means that more number of vulnerable devices are likely to be infected, thus increasing the likelihood of sampling the scanning packets transmitted by those infected devices resulting in a decrease in average detection delay. Lastly, the plots for the three attack probabilities, $p_1 = 0.5, 0.7, 0.9$, are quite close to each other, suggesting that changes in attack probability do not affect the average detection delay versus sampling frequency behavior significantly.

In Fig. 6, we have illustrated the distribution of average detection delays for vulnerable devices for a sampling frequency of 0.2 and attack probability of 0.6 using a histogram. The distribution closely fits an exponential distribution with a mean of ≈ 52 , suggesting that the probability of achieving higher and higher average detection delays for vulnerable devices decreases almost exponentially. Vulnerable devices are sampled at a relatively higher frequency and also have a higher probability of being infected than non-vulnerable devices. Therefore, scanning packets can be detected with lower delays in most trials, resulting in higher probability for lower values and lower probabilities for higher values of average detection delays.

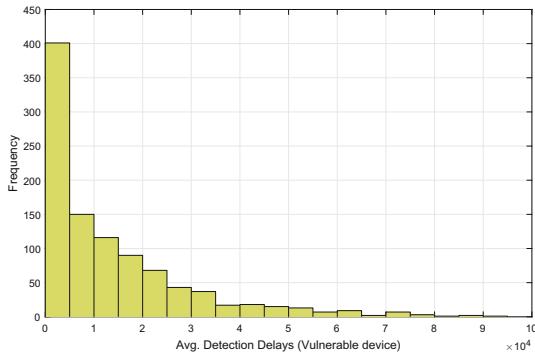


Fig. 6. Histogram of average detection delays for vulnerable devices

In Fig. 7, we have presented the plot for average detection delay versus sampling frequency for different values of attack probability on non-vulnerable devices (p_2). The plot behavior is somewhat irregular near lower sampling frequencies. For higher sampling frequencies, the average detection delay can be observed to decrease almost linearly with increasing sampling frequency. The intuitive explanation for the decreasing behavior is similar to the one given above for vulnerable devices. While designing the proposed Algorithm 1, a sampling frequency for non-vulnerable devices which is too high may lead to lower average detection delay but the corresponding increase in processing power and memory requirements may not be desirable since non-vulnerable devices are not expected to be compromised easily. A sampling frequency which is too low on the other hand, may increase the average detection delay significantly in the unexpected scenario when some of the non-vulnerable devices are compromised. Therefore,

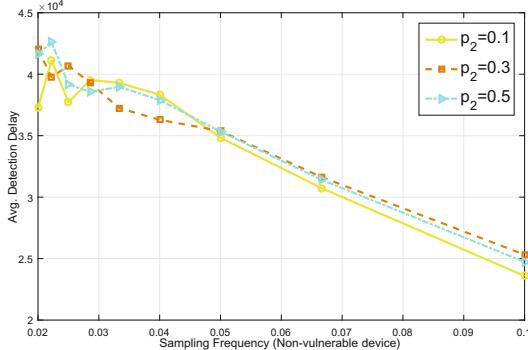


Fig. 7. Average detection delay versus sampling frequency plot for non-vulnerable devices

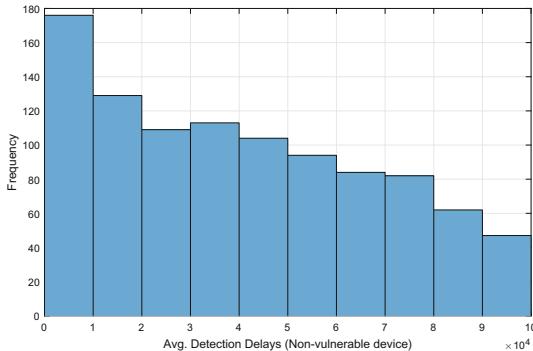


Fig. 8. Histogram of average detection delays for non-vulnerable devices

the algorithm designers may have to settle for a sampling frequency which falls in the mid of the range of available values. Figure 8 shows the distribution of average detection delays for non-vulnerable devices for a sampling frequency of 0.025 and attack probability of 0.2 using a histogram. The distribution assumes the highest values for average detection delays between ‘0 and 10,000’. Thereafter, values taken by the distribution decrease slowly with increasing average detection delays.

6 Implementation of IoT Bot Detection Algorithm

As mentioned earlier in Sect. 4.1, our proposed algorithm for bot detection has to be run on some special bot detection devices within a given network. These *sentinel* (monitoring) devices should have enough processing power and memory to run the bot detection algorithm for a large number of IoT devices. The sentinel devices can be placed higher up the network hierarchy, but below the core

network routers, to make maximum use of the sub-sampling approach employed by our proposed algorithm. We propose that a sentinel device should monitor *only the IoT devices connected to a few access network routers*, which implies that an ISP network would require multiple sentinel devices to monitor all the IoT devices in that network.

We also need to processs only IoT device packets at the sentinel devices, whereas the network traffic consists of IoT as well as non-IoT traffic (PCs, smart-phones etc.) The authors in [39] distinguish between traffic generated by IoT and non-IoT devices from a single TCP session by analyzing user-agent HTTP property for smartphones and single-session binary classifiers for PCs. A classification accuracy of 100% for smartphones and false positive, negative rates of 0.003 each for PCs were claimed to be achieved. We can use their methods to distinguish between IoT and non-IoT device packets using a single session worth of packets. Further, once we identify a device as belonging to IoT or non-IoT type, we can continue to use this information in the future as the device type is not expected to change.

It is assumed that the ISPs already have access to the information regarding vulnerable and non-vulnerable devices. As explained earlier in Sect. 4.1, IoT devices installed in home environments can be regarded as vulnerable while the devices installed in enterprise/industrial/government networks can be deemed as non-vulnerable. The routers can be configured to forward copies of received packets as well as the corresponding source device IP addresses to the sentinel devices. The incoming packets at sentinel devices can be arranged and stored in buffers according to their source devices. Figure 9 shows a prospective network deployment for our proposed algorithm illustrating the path of an IoT device

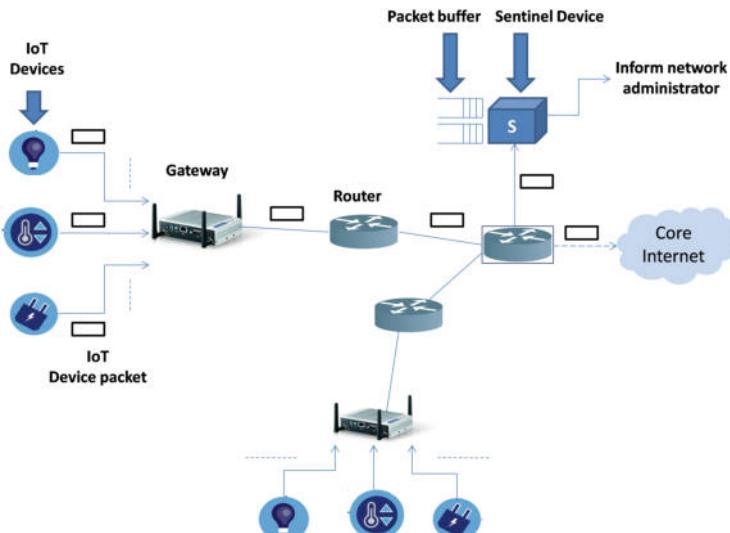


Fig. 9. Prospective network deployment for proposed bot detection solution

packet as it originates from an IoT device, passes through IoT gateway, network routers and sentinel devices. We expect the firmware running on sentinel devices to be upgradeable so that in future, if more advanced bot detection algorithms are designed (e.g. for IoT malware which do not rely on port based scanning), the corresponding software updates can be easily pushed to the sentinel devices. We cannot run our algorithm on existing core network routers since they strip a packet only until its IP (Internet Protocol) header whereas the destination port numbers are encapsulated within the TCP header.

Once the bots are detected by our proposed algorithm, the next step is to take mitigating actions to prevent the bots from spreading further damage. The network administrator can block the entire traffic originating from bots and bring them back online only after it is confirmed that the malware has been removed from those IoT devices. The concerned ISP can inform the device owners and ask them to secure their device (by using strong usernames/passwords, placing the device behind a firewall etc.). Another defense mechanism is that instead of blocking all the traffic, the bot can be allowed communications with a few secure domains for remediation of malware infection. This strategy has been mentioned as part of the bot remediation techniques [40] recommended for ISPs by IETF (Internet Engineering Task Force). The bot can also be placed under continuous monitoring and all other communication except that required for the underlying IoT device to function can be denied. Finally, security personnel can exploit bugs in the bot binary to disinfect them remotely.

7 Future Work

We are developing a software prototype of the proposed bot detection algorithm [41] which will be evaluated on an extended version of our Mirai testbed (Fig. 2) emulating a real-world network of connected IoT and non-IoT devices, gateways, routers and the proposed *sentinel* devices. It is not possible to test our algorithm with a network of physical devices as we would require hundreds of thousands of IoT devices in addition to gateways, routers and other networking equipment to replicate real-world large-scale networks. Further, we are also looking at the optimal placement of *sentinel* devices in a network by performing a cost analysis. In the future, we would like to develop solutions for detecting IoT bots infected with malware exploiting software vulnerabilities to hack the devices and add to the botnet. For instance, Linux.Darlloz, Reaper and Amnesia malware [42–44] use HTTP (Hyper Text Transfer Protocol)-based exploits to perform code injection and arbitrarily execute code on remote devices bypassing authentication. It should be noted here that the packet sub-sampling approach proposed in this paper is likely to be a part of the bot detection solution devised for such advanced malware. Finally, some malware may try to evade detection, e.g. by attempting to hide their scanning activity. It would be an interesting problem to detect such evasive IoT malware.

8 Conclusion

In this paper, we proposed an algorithm for detecting IoT devices infected by Mirai or similar malware. The bot detection algorithm uses Mirai traffic signatures and a two-dimensional sub-sampling approach. Leveraging measurements taken from a testbed constructed to simulate the behavior of Mirai, we studied the relationship between average detection delays and sampling frequencies for vulnerable and non-vulnerable devices. Based on our analysis of the plots, we made suggestions regarding the process of selection of sampling frequencies while designing our proposed algorithm. Subsequently, the deployment of our bot detection algorithm within a real-world network was discussed where we proposed using special *sentinel* devices to run the algorithm. Prospective actions which can be taken after detection of bots were also mentioned. Finally, we identified few interesting problems stemming out of this research which we would like to work upon in the future.

Acknowledgements. The authors would like to thank Dr. Liang Zhenkai (SoC, NUS) for helping us with some of the initial ideas used in this paper and Dr. Min Suk Kang (SoC, NUS) for providing comments on our manuscript. We would also like to appreciate the National Cybersecurity R&D Lab, Singapore for allowing us to use their testbed to collect important data which has been used in our work. This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Corporate Laboratory@University Scheme, National University of Singapore, and Singapore Telecommunications Ltd.

References

1. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tutorials* **17**(4), 2347–2376 (2015)
2. Nordrum, A.: Popular internet of things forecast of 50 billion devices by 2020 is outdated. <https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated>
3. Frustaci, M., Pace, P., Aloisio, G., Fortino, G.: Evaluating critical security issues of the IoT world: present and future challenges. *IEEE Internet of Things J.* **PP**(99), 1–1 (2017)
4. Lin, J., Yu, W., Zhang, N., Yang, X., Zhang, H., Zhao, W.: A survey on internet of things: architecture, enabling technologies, security and privacy, and applications. *IEEE Internet Things J.* **4**(5), 1125–1142 (2017)
5. Yang, Y., Wu, L., Yin, G., Li, L., Zhao, H.: A survey on security and privacy issues in internet-of-things. *IEEE Internet of Things J.* **4**(5), 1250–1258 (2017)
6. Krebs, B.: Hacked cameras, DVRs powered today's massive internet outage. <https://krebsonsecurity.com/2016/10/hacked-cameras-dvrs-powered-todays-massive-internet-outage/> (2016)
7. Cimpanu, C.: There's a 120,000-strong IoT DDoS Botnet lurking around. <http://news.softpedia.com/news/there-s-a-120-000-strong-iot-ddos-botnet-lurking-around-507773.shtml>

8. Constantin, L.: Your Linux-based home router could succumb to a new Telnet worm, Remaiten. <https://www.computerworld.com/article/3049982/security/your-linux-based-home-router-could-succumb-to-a-new-telnet-worm-remaiten.html>
9. Grange, W.: Hajime worm battles Mirai for control of the Internet of Things. <https://www.symantec.com/connect/blogs/hajime-worm-battles-mirai-control-internet-things>
10. Arghire, I.: IoT Botnet used in website hacking attacks. <https://www.securityweek.com/iot-botnet-used-website-hacking-attacks>
11. Beek, C.: Mirai Botnet creates army of IoT Orcs. <https://securingtomorrow.mcafee.com/mcafee-labs/mirai-botnet-creates-army-iot-orcs/>
12. Ilascu, I.: Mirai code still runs on many IoT devices. <https://www.bitdefender.com/box/blog/iot-news/mirai-code-still-runs-many-iot-devices/>
13. Yu, T., Sekar, V., Seshan, S., Agarwal, Y., Xu, C.: Handling a trillion (unfixable) flaws on a billion devices: rethinking network security for the internet-of-things. In: Proceedings of the 14th ACM Workshop on Hot Topics in Networks, HotNets-XIV, pp. 5:1–5:7, New York, NY, USA. ACM (2015)
14. Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J.A., Invernizzi, L., Kallitsis, M., Kumar, Lever, C., Ma, Z., Mason, J., Menscher, D., Seaman, C., Sullivan, N., Thomas, K., Zhou, Y.: Understanding the mirai botnet. In: 26th USENIX Security Symposium (USENIX Security 17), Vancouver, BC, pp. 1093–1110. USENIX Association (2017)
15. Livadas, C., Walsh, R., Lapsley, D., Strayer, W.T.: Using machine learning techniques to identify Botnet traffic. In: Proceedings. 2006 31st IEEE Conference on Local Computer Networks, pp. 967–974, Nov 2006
16. Gu, G., Porras, P., Yegneswaran, V., Fong, M.: Bothunter: detecting malware infection through ids-driven dialog correlation. In: 16th USENIX Security Symposium (USENIX Security 07), Boston, MA. USENIX Association (2007)
17. Gu, G., Zhang, J., Lee, W.: BotSniffer: detecting Botnet command and control channels in network traffic. In: Network and Distributed System Security Symposium (NDSS) (2008)
18. Gu, G., Perdisci, R., Zhang, J., Lee, W.: BotMiner: clustering analysis of network traffic for protocol- and structure-independent Botnet detection. In: Proceedings of the 17th Conference on Security Symposium, SS’08, Berkeley, CA, USA, pp. 139–154. USENIX Association (2008)
19. Zhang, J., Perdisci, R., Lee, W., Luo, X., Sarfraz, U.: Building a scalable system for stealthy P2P-Botnet detection. IEEE Trans. Inf. Forensics Secur. **9**(1), 27–38 (2014)
20. Habibi, J., Midi, D., Mudgerikar, A., Bertino, E.: Heimdall: mitigating the internet of insecure things. IEEE Internet Things J. **4**(4), 968–978 (2017)
21. Pajouh, H.H., Javidan, R., Khayami, R., Ali, D., Choo, K.K.R.: A two-layer dimension reduction and two-tier classification model for anomaly-based intrusion detection in IoT backbone networks. IEEE Trans. Emerg. Topics Comput. **PP**(99), 1–1 (2016)
22. Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Breitenbacher, D., Shabtai, A., Elovici, Y.: N-baiot: network-based detection of IoT botnet attacks using deep autoencoders. CoRR, abs/1805.03409 (2018)
23. Sommer, R., Paxson, V.: Outside the closed world: on using machine learning for network intrusion detection. In: 2010 IEEE Symposium on Security and Privacy, pp. 305–316 (2010)

24. Gates, C., Taylor, C.: Challenging the anomaly detection paradigm: a provocative discussion. In: Proceedings of the 2006 Workshop on New Security Paradigms, NSPW '06, New York, NY, USA, pp. 21–29. ACM (2007)
25. Koliاس, C., Kambourakis, G., Stavrou, A., Voas, J.: DDoS in the IoT: Mirai and other Botnets. Computer **50**(7), 80–84 (2017)
26. MySQL: The world's most popular open source database. <https://www.mysql.com/>
27. Ron Winward: Mirai: Inside of an IoT Botnet. https://www.nanog.org/sites/default/files/1_Winward_Mirai.The.Rise.pdf (2017)
28. Bellard, F.: QEMU, a fast and portable dynamic translator. In: Proceedings of the Annual Conference on USENIX Annual Technical Conference, ATEC '05, Berkeley, CA, USA, pp. 41–41. USENIX Association. <https://www.qemu.org/> (2005)
29. IANA: Service Name and Transport Protocol Port Number Registry. <https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml?&page=1>
30. Wanner, R.: What is happening on 2323/TCP? <https://isc.sans.edu/forums/diary/What+is+happening+on+2323TCP/21563/>
31. Postel, J., Reynolds, J.: Telnet protocol specification. <https://tools.ietf.org/html/rfc854> (1983)
32. Google Cloud: Overview of Internet of Things. <https://cloud.google.com/solutions/iot-overview>
33. Wireshark: Network protocol analyzer. <https://www.wireshark.org/>
34. Statista: Installed base of IoT consumer devices by category in the United States in 2017 (in million units). <https://www.statista.com/statistics/757717/iot-consumer-product-installed-base-in-the-us-by-category/> (2018)
35. Wikipedia: Broadband Providers in the United States. https://en.wikipedia.org/wiki/Internet_in_the_United_States#Broadband_providers
36. IEEE standard for low-rate wireless networks. IEEE Std 802.15.4-2015 (Revision of IEEE Std 802.15.4-2011), pp. 1–709, Apr 2016
37. Foong, A.P., Huff, T.R., Hum, H.H., Patwardhan, J.R., Regnier, G.J.: TCP performance re-visited. In: 2003 IEEE International Symposium on Performance Analysis of Systems and Software. ISPASS 2003, pp. 70–79 (2003)
38. Karp, R.M.: Reducibility among Combinatorial Problems, pp. 85–103. Springer US, Boston, MA (1972)
39. Meidan, Y., Bohadana, M., Shabtai, A., Guarnizo, J.D., Ochoa, M., Tippenhauer, N.O., Elovici, Y.: Profiliot: a machine learning approach for IoT device identification based on network traffic analysis. In: Proceedings of the Symposium on Applied Computing, SAC '17, pp. 506–509, New York, NY, USA. ACM (2017)
40. Livingood, J., Mody, N., O'Reirdan, M.: Recommendations for the remediation of bots in ISP networks. <https://tools.ietf.org/html/rfc6561>
41. Kumar, A.: Software Repository for Mirai-Like IoT Malware Detection Algorithm. <https://github.com/ayush47-github/IoT-bot-detection>
42. Zheng, C., Xiao, C., Jia, Y.: New IoT/Linux malware targets DVRs, Forms Botnet. <https://researchcenter.paloaltonetworks.com/2017/04/unit42-new-iotlinux-malware-targets-dvrs-forms-botnet/>
43. Hayashi, K.: Linux worm targeting hidden devices. <https://www.symantec.com/connect/blogs/linux-worm-targeting-hidden-devices>
44. Radware. Reaper Botnet. <https://security.radware.com/ddos-threats-attacks/threat-advisories-attack-reports/reaper-botnet/>



Desktop Browser Extension Security and Privacy Issues

Steven Ursell^(✉) and Thaier Hayajneh

Fordham Center for Cybersecurity,
Fordham University, New York, NY, USA
sursell@fordham.edu, thayajneh@fordham.edu

Abstract. Since their introduction in the 1990's, users have adopted internet browsers as a convenient method of interacting with computers and servers whether collocated with the user or located across the planet. As browsers have become more sophisticated, additional capabilities have been made available to users through browser extensions. When written by trusted agents, these browser extensions provide safeguards for users, but browser extensions can also be written so that a user's data can be extracted and used for purposes the user would never agree to. This paper began with the exploration of extensions in four popular browsers: Safari, Firefox, Chrome, and Internet Explorer (Edge) and the author explored the security and privacy practices inherent within the extensions, but only two of these browsers will be examined in this paper. Safari is eliminating all extensions outside of its tightly controlled delivery system beginning with the debut of its new operating system in September 2018 and Internet Explorer is being replaced by Edge, which is also tightly controlled by Microsoft. Presumably, Safari and Edge extensions will be secure once the developers submit the code and it is reviewed before the extensions are published. Because there are literally thousands of browser extensions it is not possible to examine all of them in a single paper, but it is the intent of the author to establish an evaluation framework so browser extensions can be objectively scored.

Keywords: Extension · Malware · Security · Privacy

1 Introduction

It can be argued that the popularity of the internet began with the adoption of internet browsers. In 1994 Netscape was introduced as a browsing tool and then Microsoft introduced Internet Explorer (IE) in 1995. Microsoft began to bundle IE with its operating system, and by 2003 IE was being used by 95% of all internet users. Other browsers were introduced; Firefox in 2003, Safari in 2003, and Chrome in 2008. It should be noted that while Chrome is the browser most recently introduced, it now is the preferred desktop browser with a 66% market share as of June 2018 [1].

Microsoft enabled browser extensions to be used with IE in 1999, Firefox in 2004, while Safari and Chrome introduced extensions for use in 2010. With the release of Safari 12.0 on 17 September 2018, all legacy extensions were removed from the Apple store and from user computers [2]. At this time there are only 49 Safari extensions available for download. Future extensions will be evaluated by Apple experts and the extensions will be made available to users after the evaluation. For the sake of simplicity, the terms browser extensions and add-on are referred to as extensions in this paper. Additionally, in 2018 Firefox Quantum replaced Firefox, but the name Firefox will be used when referring to Firefox Quantum. Browsers are capable of performing nearly all the tasks a user needs to complete, but there are often additional tasks that users want to accomplish that would not make sense to include in a browser. For example, a user might wish to read websites in a familiar language, but no browser contains translation capabilities for all languages. To solve this shortcoming the user can install an extension that provides real-time translation capabilities. Maybe the user doesn't wish to be tracked across the web. An extension is available to prevent tracking. Legitimate and well written extensions can assist users to be more creative and safer when access the internet, but poorly written extensions, or extensions written to take advantage of uninformed users can easily be selected instead.

Because extensions are designed to work within browsers, extensive privileges are required. The misuse of privileges can cause browser instability as well as open up many security vulnerabilities [3]. It is important to make a distinction between extensions and plug-ins because security vulnerabilities are very different. While extensions enable the browser to carry out additional tasks when accessing web pages, plug-ins that have been used to introduce security vulnerabilities are Adobe Flash and Oracle Java. One of the many approaches to overcoming security measures is for attackers to embed malicious files in websites and then the plug-ins run the malicious file when the browser encounters the website [4].

Reference [5] provides useful definitions that can be used when discussing malware and many of them apply to the discussion of faulty or malicious extensions. Espionage, sabotage, and hijacking were listed as the main approaches malware uses to attain the goals of the attacker. While sabotage is not usually a goal of the extension attacker, though system sabotage is often collateral damage. Espionage is the attempted theft of user data and personal information as well as the theft of system data required to spoof accounts. Hijacking refers to either the repurposing of a system so that the attacker can use the system, or the creation of zombie machines that can be used to attack other machines or facilitate the creation of a communications relay system that can forward messages or data. Other terms are equally descriptive and useful, but all malicious behavior can be described as being harmful to the user.

Poorly written or malicious extensions can leak data and privacy information such as browsing history, credit card numbers, contacts, and passwords [4]. Other information that can be exfiltrated are user credentials, email messages, and

organizational infrastructure [6]. Exfiltration attempts are rarely detected in real-time. Most successful data exfiltration events are not revealed to the target for extended amounts of time and this delay allows attackers to exploit the stolen information.

If a researcher asked the typical user to describe the difference between an operating system, browser, and a word processing tool the typical user would not be able to answer the question. (“Typical user” in this paper refers to someone who does not normally read about computer technology other than what is published in general news sources like the NY Times or The Washington Post. Typical users have little technical knowledge except for the knowledge gained from taking a computer safety course at work or school.) Typical computer users and skilled computer users approach the internet in substantially different ways. The typical user turns on the computer and begins interacting with it in a similar manner that a driver starts a car, while the mechanic starts the car and monitors its health while using the car to get from one place to the other. Experienced users are usually more aware of the dangers the internet presents to those who interact with it. Some extensions disguise themselves in order to attract additional and unsuspecting users. It can be assumed that few experienced users would install a Firefox language translation package if it was known that operating system commands could be executed by that package [3]. Typical users would not understand the consequences of allowing access to the operating system and this is an obvious reason why typical users need to be protected from making uninformed decisions.

This paper will present related work in Sect. 2. This overview contains examples of extension research for Chrome and Firefox browsers. Section 3 discusses extension security and privacy concerns and Sect. 4 presents an approach to categorizing those extension concerns. Section 5 provides a discussion of the top 5 challenges future researchers will have to overcome before a standardized solution can be provided to the user. Section 6 provides recommendations for future research and finally, Sect. 7 concludes the paper.

2 Related Work

The rapidly rising popularity of extensions among users dramatically increased the number of extensions being created, and not all of the extensions were developed by established software vendors. Solving security issues was often not part of the development process when extensions began to be developed, and privacy data was exposed when poorly written extensions were paired with browsers. While some design errors occurred in the extension development process, some extensions did not accidentally expose user information, some extensions deliberately collected user information. And though researchers started to examine the security flaws of extensions at least 10 years ago major security issues have not been resolved.

In 2011 researchers Bandhakavi et al. [7] examined Firefox extension security issues by focusing on JavaScript code and provided programming guidance that

would help avoid unnecessary privilege escalation. Vetting extensions (VEX) was their solution to the problem of analyzing thousands of extensions. Through static analysis and by analyzing extension code and extension strings, VEX's systematic approach did not entirely rely on humans. VEX had the ability to analyze thousands of extensions in a single pass, but false positives often were reported in the analysis. Human vetting of Firefox extensions was the final step in the VEX analysis model. VEX was not a proposed solution for browsers other than Firefox.

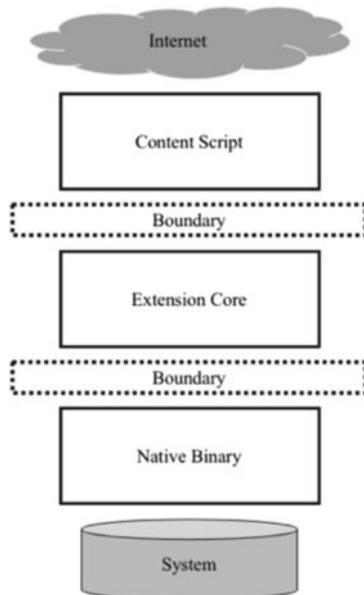


Fig. 1. Chrome extension separation framework

Nearly 50,000 Chrome Web Store extensions were analyzed by Kapravelos et al. [8] in 2014 by examining extension execution and the resulting network activity. Their system was named Hulk and it was designed to be dynamic instead of static. Hulk was used to look for Chrome extensions with malicious intent and not just poorly designed extensions. Kapravelos listed four categories of malicious behavior: ad manipulation, affiliate fraud, information theft, and online social network (OSN) abuse. By observing the interaction of extensions and web pages, Hulk's algorithms were able to label extensions as malicious, suspicious, or benign. Using extensive URL lists, Hulk attempted to trigger the functionality of the extensions it was investigating; however, some extensions would only activate if the correct web page content was encountered. To solve this content issue, the researchers created "HoneyPages" that would present the conditions needed for extensions to trigger. HoneyPages provided a controlled environment that

allowed the researchers to determine what an extension might be looking for during its execution. To provide greater granularity, the researchers used a fuzzer to trigger event handlers on over 1,000,000 URLs.

While many Chrome extensions have been found to be malicious, the Chrome browser itself was created with security as one of its goals [9]. Google partnered with the University of California to create the Chrome extension framework. Part of the rationale for creating the Chrome extension framework was a survey of 25 Firefox extensions and the subsequent discovery of the extensive use of elevated privileges to enable the extensions. The Chrome extension framework is based on using the least amount of system privilege, separation of extension components (see Fig. 1) and isolating those components from each other by initiating separate system processes for each of them [10]. One of the initial processes used was the Netscape Plug-in API (NPAPI); however, the NPAPI interface eventually was exploited by attackers and it was removed from the Chrome extension framework in 2013 [11].

Research into the security implications of Firefox extensions began at least as early as 2008 [12]. The first known published paper that addressed malicious extensions was written by Beaucamps and Reynaud and presented at a 2008 information security symposium in France. Their paper is notable because it included only one citation [13] and it was included within the paper itself as a warning to attendees that ignoring extension security needed to be addressed before a security disaster happened. The dangers of system-wide access, extension viral behavior, and the vulnerability to all operating systems using Firefox extensions was clearly stated, and while typical Firefox users in 2008 could not be expected to know about extension security, typical users in 2018 need to know about the consequences of using extensions.

3 Extension Security and Privacy Concerns

Many technology-focused news outlets publish stories about extensions, and some of them explore which extensions provide the best security and which extensions should be avoided. Those stories are regularly updated as extensions adopt unacceptable practices and new extensions replace older extensions [14]. It is not the intent of this author to compile lists of good and bad extensions, nor will a list of preferred extension reviewers be recommended. Published papers will be augmented with recent news stories in order to provide both authoritative and current information.

3.1 Extension Security Concerns

- The sale of extensions by their developers, after a large number of users have installed them, has become a common occurrence. Users accept the privacy policy of the extension during installation, but there is nothing to stop future extension owners from changing those policies. A well-known Chrome and Firefox extension (Stylish) became little more than spyware after it began

collecting browser data and Google search results after it was sold to SimilarWeb in January 2017 [15]. Even though Stylish is no longer available for download, the 2 million people who use Stylish have no way of knowing that the extension is collecting their data. In fact, many extensions are purchased solely for the purpose of injecting malware or stealing user data [16].

- Browsers are often written that require high-level privileges and extensions, however, extensions usually can function with much lower privileges and the threat from a poorly written extension can be greatly reduced. An extension that is granted access to all user data on all websites should alarm the user [10, 17].
- Poorly written extensions can expose user data just like those extensions developed with malicious intent. Most extensions are not written by well established companies, and because extension software is easily written, anyone with just a little experience can write an extension. Of course, Firefox and Chrome curate extensions hosted on their stores, but many dangerous extensions are available to the public before their vulnerabilities are fully understood. Extension patches and the removal of extensions from the stores can take months and the user is often never informed [18, 19].

3.2 Extension Privacy Concerns

- Disclosure of extension privacy practices to the user is incomplete. Many user agreement statements are written at higher reading levels than the user can understand, and as a result, most users cannot make an informed decision as to whether privacy safeguards are adequate [20].
- Information leakage can result from extensions collecting HTTP Referer header information. This header informs the current webpage which webpage the user was viewing just before clicking the link to the current webpage [21]. This piece of information is useful to webpage owners so they know how their webpage becomes known to the user, is incidental leakage and not exclusively malicious [22].
- Users generally do not read computer program warnings [23], and those warnings are usually not easily understood. Extension warnings are not comprehensible and the authors may only provide the minimum amount of warning required for the extension to be listed on the Firefox or Chrome stores.

4 Categorizing Extensions

Thousands of extensions have been examined by multiple researchers and many proposals to make them secure have been published. This paper proposes no solution to extension shortcomings, but rather an approach to categorizing extensions in a manner that a typical user can make an informed decision. As discussed earlier, only Chrome and Firefox browser extensions are examined in this paper, but this approach could be extended to extensions integrated into other browsers.

The proposed criteria are listed in Table 1 with explanations below. Other criteria must be explored as this list is not exhaustive. Once criteria are agreed upon, a decision matrix can easily be created in order for a score to be provided to the typical user. Based on this score the typical user can make an informed decision whether an extension is safe to install. This categorization is not intended to assist the user in deciding whether an extension is useful, but it is intended to inform the user about the safety of an extension.

A series of primary decision points should be part of the process, meaning that an extension can be selected for elimination at that point in the process. Just knowing these critical decision points are part of the extension rating process could influence a developer to make better choices. For example, if a developer knows that data sharing prevents a recommendation the developer should know that the extensions design must be designed (or altered) to prevent user data from being transferred to other extensions or outside of the users desired restrictions.

Figure 2 illustrates how extensions could be rated and totaled. Users could determine safety based on higher scores.

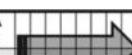
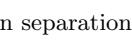
	Privacy (1-9)	Security (1-9)	EULA (1-9)		Total
Ext. 1	5	2	7		72
Ext. 2	8	8	4		85
Ext. 3	1	1	8		40
Ext. 4	3	2	2		34
Ext. 5	2	7	7		55

Fig. 2. Chrome extension separation framework

- Rating how an extension protects against the loss of privacy begins with knowing what is being protected. Not all extensions encounter personally identifiable information (PII) so the grading criteria could be skewed toward extensions with little PII contact. Other privacy protections can be evaluated across all browser, such as browsing history.
- Security grading is more objective than privacy as all extensions must provide a minimum amount of security protection, but will extensions be rated according to data lost due to active attacks or inherent security flaws? Most extensions will not be attacked so should extensions not attacked be rated above extensions that were attacked?
- The user agreement should be easily understood and comprehensive. If there are details that are not understood, the developer should provide explanations either within the agreement or a link to an area where the user can fully understand the risks of installing the extension, and the installation should not proceed until the user acknowledges understanding the agreement.
- Some criteria can be answered yes, or no. Examples are whether the extension comes from a curated source, does it employ JavaScript, and is the extension

reviewed by a community of other developers? These questions are more about declaring aspects of an extension rather than making a judgment on their usage.

- The computational cost associated with an extension will depend on the computing platform and it will not be possible to assign a score for each possibility. A standardized platform should be used, and the computation “delta” score assigned. Because many extensions are multi-platform, this standardized platform should include Windows, Chrome, and MacOS operating systems. To be clear, higher computational costs might be incurred due to additional security procedures and processes. Those higher computational costs should not be used to justify a lower score.
- Update policies should be understood by the user and a choice provided as to whether or not an update requires user authorization and whether the update can be automated when the browser is updated. Users should not be surprised when a browser is updated or when its extensions are updated.
- Most extensions gather user data, but there is little justification for extensions to share user data. Sharing data should be a primary decision point and extensions that share data should be eliminated from further evaluation. Data sharing policies must be fully explained to the evaluators if elimination is to be avoided. If data is shared, those policies must also be explained to the user.
- Browsers are run inside a sandbox and additional security can be gained by running extensions inside a sandbox [24]. This additional layer of sandboxing can protect operating systems and user data by sequestering processes within a defined operational footprint. Privileges are restricted to just what is required for the extension’s process to accomplish its task, and data not needed by that process is safeguarded.
- There are many legitimate reasons why a developer would create and sell a browser extension, but whether the extension is sold for legitimate or illegitimate reasons the user should be aware and provided an option to opt out of the continued use of the extension.
- The “Top 10” criteria should reflect the market share of users who have installed these extensions, but the extension builders are not likely to release this data unless they are at or near the top. In fact, there are multiple categories of Top 10 extensions and none of them can be considered to be authoritative. Finding authoritative sources who will provide a Top 100 is not possible. It should be possible to synthesize the Top 10 or Top 100 list of extensions by surveying trade publications and web stores. As long as the methodology is transparent, the result can be assumed to be without bias.

5 Top Standardization Challenges

If the typical user is going to be able to trust the outcome of this work, there must be a way to assure the user that the outcome is not swayed by market forces or subtle manipulation. No survey can eliminate all uncertainty, but the survey process must reassure the user that the survey results are accurate.

Table 1. Sample extension categorization

	Chrome and Firefox extensions				
	Ext. 1	Ext. 2	Ext. 3	Ext. 4	Ext. Etc.
Chrome					
Firefox					
Privacy					
Security					
User agreement					
Curated					
Java Script					
Community review					
Computation cost					
Update policy					
User data shared					
Sandboxing					
Ownership transfer					
Top 10/Top 100					

5.1 Elimination of Categorization Bias

Many researchers and industry publications provide analysis of extensions, but objectivity is difficult to ascertain. Bias of one browser over another can be difficult to overcome and some publications are quick to publish flashy headlines in order to drive advertising to their website. The work of researchers is peer reviewed and bias can be removed during that process, but the research can take months and the audience is much smaller than publications. An example of a category that is not analyzed is the ranking of the top extensions. Neither the Chrome nor Firefox provide the rankings of extensions. Both of these stores simply parse extensions into categories and the user must sift through them or use a search function. Multiple trade and consumer publications write stories and rank extensions, but most of them are thinly veiled advertisements for their preferred browser or operating system.

5.2 Recognition of User Bias

User bias cannot be underestimated. Many users will not use a different browser than they have become accustomed to using. Chrome is the browser of choice for most desktop internet users, so it is reasonable to assume they will continue to use Chrome instead of switching to Firefox just because a browser extension survey allows the user to select the safest extension.

5.3 Criteria Weighting

Some extensions should be eliminated from consideration if they contain a serious flaw. For example, if a browser shares user data it should not be evaluated any further. But what about more subtle criteria? Is a curated web store preferred over extensions that are community reviewed? Is a readable user agreement as important as the security of an extension?

5.4 Evaluating User Agreements

End-User License Agreements (EULA) are not written so the typical user can understand them and most take a very long time to read [25]. A standardized readability engine should be selected so the writing can be evaluated, and a reading level assigned. A credit card EULA [26] averages out to an 11th grade reading level and is still difficult for most of the population to understand without considerable effort. Preferably, the extension EULA reading level would be lower.

5.5 Extension Update Cycle

A legitimate extension will be updated to stay relevant as the internet changes. The lack of updates can be an indication of a poorly written extension and an unpublicized extension update can indicate an extension contains malicious capabilities. Once a comprehensive survey has been developed, it must be regularly updated if the user is to be able to fully rely on it. Gathering update cycles, policies, and history is labor intensive since there is no known centralized extension update repository and a neutral source of funding the survey will have to be established.

6 Recommendations

After examining extension security and privacy concerns, it is apparent that extension developers, distributors, and hosting browsers are not the only source for solutions. Users must also be part of the solution in current and future platforms.

6.1 Simplified User Decision Matrix

Simply listing required permissions and asking the user to permit access to the extension is not helpful to the uninformed user. Most users do not read the user agreements when presented with the opportunity. As a condition of installing an extension, the user must be required to complete a series of permission authorization steps. Each step must inform the user of the consequences of permitting the extensions activity. This sequential approval process will better inform the user than asking the user to approve the entire EULA with a single click. Before the user ultimately agrees to installing and using the extension, the user must

be presented with a risk analysis that enables the user to make an informed, and collaborative decision. Any user can just agree and then hope they are not compromised, but a collaborative approach to installing and using extensions is helpful for the user, and ultimately protects the extension provider from accusations should the user become a victim after the extension has been installed.

6.2 Expand Scope to Mobile Platforms

This paper focused on desktop browser security, but mobile browsers have had a higher market share than desktop browsers since October 2016 [1]. Mobile browser extensions are not currently permitted on Chrome, but Firefox does allow browser extensions [27]. It is essential that extension privacy and security considerations are surveyed for the typical mobile user as well as the desktop user.

6.3 Supplemental Evaluation

An independent evaluation and categorization of extensions must be viewed as an additional step in the extension release process. The initial and primary evaluations will still be conducted by the extension creators and browser webstores. The process described in this paper should be conducted by representatives of those webstores, non-affiliated professional developers, and academic researchers.

6.4 Standardized Evaluation Criteria

There are standard reporting systems for malware, software vulnerabilities, and zero-day attack possibilities, but standardized browser and extension evaluation criteria does not exist. Standardized criteria should be developed, maintained, and expanded so that browsers and extensions can be evaluated regardless of operating system or hardware platform.

7 Conclusion

Users (typical and experienced) need a decision tool that will help them narrow the gap between what they expect an extension is capable of carrying out and what the extension is actually executing. Most users do not know how powerful extensions are and how much damage they can do if poorly written or if they are maliciously written. If a user has a tool to help identify whether the security tradeoffs are worth the benefits of an extension, then the expectation gap can be narrowed.

Extensions have been examined for security flaws for the last ten years and yet many extensions are still not safe. Steps have been taken to strengthen the development and validation process and the makers of the two main browsers should be commended for their efforts. This paper reviewed extensions of the Chrome and Firefox browsers and developed a high-level survey process that

should enable the typical user to decide whether there are security or privacy concerns before installing the extensions. The survey should not be considered to be all inclusive, but it provides a template for future researchers to use as a more comprehensive survey is developed.

References

1. StatCounter: Desktop browser market share worldwide. StatCounter, 24 July 2018. [Online]. Available <http://gs.statcounter.com/browser-market-share>. Accessed 24 July 2018
2. Chaffin, B.: Apple Releases Safari 12 for High Sierra and Sierra, Combats Ad-Tracking and Increases Security, The Mac Observer, 17 Sept 2018. [Online]. Available <https://www.macobserver.com/news/product-news/safari-12-macos-ad-tracking-security/>. Accessed 27 Sept 2018
3. Golubovic, N.: Attacking browser extensions, 3 May 2016. [Online]. Available <https://golubovic.net/thesis/master.pdf>. Accessed 25 July 2018
4. Hoffman, C.: Beginner Geek: everything you need to know about browser extensions, How-To Geek, 1 Aug 2013. [Online]. Available <https://www.howtogeek.com/169080/beginner-geek-everything-you-need-to-know-about-browser-extensions/>. Accessed 25 July 2018
5. Dornhackl, H., Kadletz, K., Luh, R., Tavolato, P.: Defining malicious behavior. In: 2014 Ninth International Conference on Availability, Reliability and Security, Fribourg, Switzerland (2014)
6. Cisco: Cisco 2017 Midyear Cybersecurity Report, July 2017. [Online]. Available https://www.cisco.com/c/dam/global/es_mx/solutions/security/pdf/cisco-2017-midyear-cybersecurity-report.pdf. Accessed 25 July 2018
7. Bandhakavi, S., Tiku, N., Pittman, W., King, S., Madhusudan, P., Winslett, M.: VEX: Vetting Browser Extensions For Security Vulnerabilities, pp. 91–99. Association for Computing Machinery, Sept 2011
8. Kapravelos, A., Grier, C., Chachra, N., Kruegel, C., Vigna, G., Paxson, V.: Hulk: eliciting malicious behavior in browser extensions. In: Proceedings of the 23rd Usenix Security Symposium, San Diego (2014)
9. Liu, L., Zhang, X., Uan, G., Chen, S.: Chrome extensions: threat analysis and countermeasures. In: 19th Network and Distributed System Security Symposium (NDSS '12). San Diego, California (2012)
10. Barth, A., Felt, A.P., Saxena, P., Boodman, A.: Protecting Browsers from Extension Vulnerabilities, 18 Dec 2009. [Online]. Available <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.378.8542&rep=rep1&type=pdf>. Accessed 26 July 2018
11. Schuh, J.: Saying Goodbye to Our Old Friend NPAPI, Google, 23 Sept 2013. [Online]. Available <https://blog.chromium.org/2013/09/saying-goodbye-to-our-old-friend-npapi.html>. Accessed 1 August 2018
12. Beaucamps, P., Reynaud, D.: Malicious firefox extensions. In: Symposium sur la securit des techniques d'information et de communication. Rennes, France (2008)
13. Ter Louw, M., Lim, J., Venkatakrishnan, V.: Enhancing web browser security against malware extensions. J. Comput. Virol. 4(3), 179–195 (2008)
14. Henry, A.: The best browser extensions that protect your privacy, Lifehacker, 31 Aug 2015. [Online]. Available <https://lifehacker.com/the-best-browser-extensions-that-protect-your-privacy-479408034>. Accessed 26 July 2018

15. Burlacu, A.: Browser Extension Secretly Stole Chrome And Firefox Users' Entire Browsing History, TechTimes, 5 July 2018. [Online]. Available <https://www.techtimes.com/articles/231851/20180706/browser-extension-secretly-stole-chrome-and-firefox-users-entire-browsing-history.htm>. Accessed 26 July 2018
16. Osborne, C.: Firms buy popular Chrome extensions to inject malware, ads, ZDNet, 20 Jan 2014. [Online]. Available <https://www.zdnet.com/article/firms-buy-popular-chrome-extensions-to-inject-malware-ads/>. Accessed 26 July 2018
17. Guha, A., Fredrikson, M., Livshits, B., Swamy, N.: Verified security for browser extensions. In: 32nd IEEE Symposium on Security and Privacy, Berkley, California (2011)
18. Cobb, M.: Web browser extension security: Mitigating browser plug-in threats, SearchSecurity, Nov 2013. [Online]. Available <https://searchsecurity.techtarget.com/tip/Web-browser-extension-security-Mitigating-browser-plug-in-threats>. Accessed 26 July 2018
19. Constantin, L.: Researcher to demonstrate feature-rich malware that works as a browser extension, ComputerWorld, 24 Oct 2012. [Online]. Available <https://www.computerworld.com/article/2492866/desktop-apps/researcher-to-demonstrate-feature-rich-malware-that-works-as-a-browser-extension.html>. Accessed 26 July 2018
20. Martin, D., Smith, R., Brittain, M., Fetch, I., Wu, H.: The privacy practices of Web browser extensions. Commun. ACM **44**(2), 45–50 (2001)
21. Kyrnin, J.: How to Use the HTTP Referer, LifeWire, 4 Apr 2018. [Online]. Available <https://www.lifewire.com/how-to-use-http-referer-3471200>. Accessed 26 July 2018
22. Starov, O., Nikiforakis, N.: Extended tracking powers: measuring the privacy diffusion enabled by browser extensions. In: International World Wide Web Conference Committee, Perth, Australia (2017)
23. Reeder, R., Porter, A., Consolvo, S., Malkin, N., Thompson, C., Egelman, S.: An experience sampling study of user reactions to browser warnings in the field. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems Paper, Montreal, Canada (2018)
24. Hoffman, C.: Sandboxes Explained: How They're Already Protecting You and How to Sandbox Any Program, How-To Geek, 2 Aug 2013. [Online]. Available <https://www.howtogeek.com/169139/sandboxes-explained-how-theyre-already-protecting-you-and-how-to-sandbox-any-program/>. Accessed 1 Aug 2018
25. Madrigal, A.: Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days, The Atlantic, 1 Mar 2012. [Online]. Available <https://www.theatlantic.com/technology/archive/2012/03/reading-the-privacy-policies-you-encounter-in-a-year-would-take-76-work-days/253851/>. Accessed 1 Aug 2018
26. CreditCards: Study: Credit card agreements unreadable to most Americans, CreditCards, 16 Sept 2016. [Online]. Available <https://www.creditcards.com/credit-card-news/unreadable-card-agreements-study.php>. Accessed 1 Aug 2018
27. Knight, J.: Add New Functionality to Your Browser with Extensions, Gadget Hacks, 18 Dec 2017. [Online]. Available <https://android.gadgethacks.com/how-to/firefox-mobile-101-add-new-functionality-your-browser-with-extensions-0181656/>. Accessed 24 July 2018



Exploring Cybersecurity Metrics for Strategic Units: A Generic Framework for Future Work

Mohammad Arafa, Saad Haj Bakry^(✉), Reham Al-Dayel,
and Osama Faheem

Department of Computer Engineering, King Saud University,
Riyadh, Saudi Arabia
shb@ksu.edu.sa

Abstract. Strategic units at the various levels of the cyberspace, from the personal level to the world level, through the enterprise and the country levels, need to protect their information security. In this respect, cybersecurity metrics for such units are important for continuous monitoring, assessment, and response to the various security challenges that they may face. This paper explores the development of such metrics and provides a framework that establishes a common base for the development of such metrics for different strategic units. The paper identifies cybersecurity and strategic units; elaborates on cybersecurity issues; discusses strategic units' cybersecurity problems that require metrics; introduces a generic framework for such metrics; and considers future use of the framework. The paper hopes to highlight the issue of cybersecurity metrics for strategic units, and to help researchers identify related problems for future work.

Keywords: Cybersecurity · Strategic units · Metrics · Standards

1 Introduction

Cyberspace refers to the online world of computer networks and the Internet [1], which covers the whole world. This involves the various worldwide connected enterprises, people and devices. Cybersecurity is concerned with protecting the cyberspace. While the world is facing increasing security challenges, reflected on the cyberspace, cybersecurity is becoming of increasing importance for all connected parties at the national and international levels.

Three reputed organizations concerned with best cybersecurity practices and standards have provided technical definitions to cybersecurity. These organizations are: the International Telecommunication Union (ITU) [2]; the International Standards Organization (ISO) [3]; and the American National Institute for Science and Technology (NIST) [4].

The ITU defines cybersecurity as follows [2]: “Cybersecurity is the collection of tools, policies, security concepts, security safe guards, guidelines, risk management approaches, actions, training, best practices, assurance and technologies that can be used to protect the cyber environment, and organization and user’s assets.

The objectives of cybersecurity are: availability; integrity, which may include authenticity and non-repudiation; and confidentiality”.

The definition offered by ISO is as follows [3]: “Cybersecurity is the preservation of confidentiality, integrity and availability of information in the cyberspace”.

In addition, NIST definition is given in the following [4]: Cybersecurity is the process of protecting information by preventing, detecting, and responding to attacks; a cybersecurity event is a change that may have an impact on organizational operations, including: mission, capabilities, and reputation.

Strategic units in the cyberspace should recognize all the above in protecting themselves, and the other strategic units associated with them.

A strategic unit (SU) is viewed as follows [5]: “An active entity performing functions, providing outcomes, and looking ahead toward better status. It requires a sound strategy to guide its development, and it needs effective management to reach its target”.

Considering this concept, an SU can be associated with different levels, from the level of an individual human being, to the level of the whole world that is considering the cyberspace, including: an enterprise level, a country level, and various other levels with different scopes. In this respect, a higher level strategic unit would involve many lower level strategic units.

The responsibility for cybersecurity is a collective one shared among all SUs at all levels in the cyberspace. Each SU should have its own information security management system, which is defined, by ISO, as follows [6]: “An information security management system (ISMS) is identified as part of the overall management system, based on a business risk approach, to establish, implement, operate, monitor, review, maintain and improve information security”.

All SUs at all levels are controlled by the core SU that is the human being with all his rationality on the one hand, and his recklessness on the other. While SUs should care for protecting, not only themselves, but also others in the cyberspace, some SUs may generate challenges against specific competing SUs, or even against the cyberspace at large.

This paper is concerned with exploring cybersecurity metrics for SUs by developing a framework that incorporates the key issues involved. This would help researchers to identify focus areas for research work that contributes to future development of cybersecurity metrics on the one hand; and it would also help professionals in SUs to cooperate and improve their ISMSs. The framework is developed according to the following four main steps.

- The first step is concerned with identifying the SUs’ cybersecurity issues.
- The second is related to identifying some key cybersecurity metrics’ problems facing SUs.
- The third is associated with developing the targeted cybersecurity metrics framework considering the outcome of the above steps.
- The fourth highlights the use of the framework to identify problems for future work.

The paper is finally concluded with some remarks.

2 Cybersecurity Issues

The key cybersecurity issues involve: the assets of the various SUs in the cyberspace, which require protection; the risks that challenge the assets; the protection controls that can reduce or eliminate risks; and the security objectives according to which protection should be provided. These issues are explored in the following.

2.1 Assets

The central strategic asset of the cyberspace is the SUs' various types of information and their related functions and deliverables, as all other assets would be related to them. These assets include: the related information and communication technology (ICT); people and organizations involved; and the physical premises and support utilities concerned. The challenges facing SUs' assets in the cyberspace may have targets of wide scope, or of limited scope. Within their targets, they would threaten the following:

- The technology value of SUs, including the value of the software and hardware, in addition to networking equipment, concerned with the SUs functions.
- The value of reputation of the SUs and people in charge of the information functions; and the value of satisfaction of the SUs and people making use of these functions.
- The value of the premises and support utilities concerned.

By threatening the above values, these challenges lead to compromising the value of the cybersecurity central strategic asset that is the SUs information and their related functions and deliverables.

2.2 Risks

With the expansion of the cyberspace both geographically and in performed functions, worldwide risks to cybersecurity are on the increase. Risks challenge SUs values, and this challenge is caused by various reasons including: malicious events; accidental events; events resulting from vulnerabilities; events caused by the environment; or any combination of these events. Relative to their targets, the sources of these risks can be related to: technology problems; organizations policies; people behavior; and environment.

Different risks occur with different frequencies causing different levels of loss of value for their various targeted SUs assets. Known risks can be determined, and protection response against them can be provided. For other risks monitoring is needed to enable their detection, provide the needed response, and accumulate experience. Such monitoring should be continuous, to be able to respond to the unfortunate innovation in risk generation. In this respect, one of the important groups concerned is Carnegie-Mellon University (CMU) Computer Emergency Response Team (CERT) whose mission for its home country, the USA, is anticipating and solving the nation's cybersecurity challenges [7].

2.3 Protection Controls

Rational and empirical protection controls have been, and are being, developed for the protection of the cyberspace SUs assets from various risks. Some controls are developed to mitigate or eliminate loss of values resulting from known risks on SUs assets, on the one hand; and to attempt to manage other emerging risks, on the other. Among these controls are those recommended by ISO 27001 [8]. In addition, innovation of new controls is continuously needed to respond to the unfortunate innovative new risks.

Like SUs assets and sources of risks, protection controls can also be associated with: technology, organizations, people, and the environment, as explained in the following [5].

- Considering technology, technical protection controls are needed for access to information, privacy of information, operations, and communications.
- Regarding organization and people, organizational protection controls are needed for protection policies and management considering: human resources, assets, suppliers, incidents, business continuity, and compliance with related requirements.
- On environment, protection controls are for safe premises, and utilities.

2.4 Objectives

Cybersecurity objectives are concerned with: the availability of information; its integrity, including authenticity and non-repudiation; and its confidentiality. The cost of protection controls versus the cost of risk, may be another important factor for some SUs, especially Small to Medium Enterprises (SMEs). In addition, some SUs may emphasize privacy and demand high level of confidentiality, with cost being a minor factor.

It should be noted here that while the objectives above are essential to all SUs in the cyberspace, their required level of protection may differ depending of the business nature of the SU concerned.

3 Cybersecurity Metrics

Metrics for supporting cybersecurity management in SUs would be partly generic for all SUs in the cyberspace, and partly specific depending on the different characteristics and requirements of SUs. In this section, the following five main themes concerned with SUs' cybersecurity metrics are discussed.

- The first considers classification of SUs' cybersecurity issues including: assets; risks; protection controls; and requirements. This classification provides a base upon which metrics can be built.
- The second is associated with transforming recommended SUs security standards into metrics that enable assessing the extent to which the items of the standards are applied.

- The third is concerned with developing cybersecurity metrics for SMEs, which enjoy some specific features.
- The fourth is associated with providing cybersecurity metrics for some SU application; one important application is that of mobile communication.
- The fifth theme considers exploring metrics for other related problems.

3.1 Classification of Cybersecurity Issues (Cyberattacks)

Classification of cybersecurity issues, including: assets, risks and protection controls, is essential for security management and consequently for the development of cybersecurity metrics. An example of such a classification is given in paper [9]. The paper classifies cyberattacks according to the following criterion:

- Purpose, where attacks are viewed according to their target, which may be: access; inspection; denial of services; or other purposes.
- Violation of the legal system, where attacks include: cybercrimes; cyberespionage; cyberterrorism; and cyber-war.
- Severity of involvement, where attacks can be: active attacks that can transmit data to various parties, or block data transmission to such parties; or passive attacks that eavesdrops on the communication between two parties to steal information.
- Scope, where attacks range between: malicious large scale attacks that involve thousands of systems and causes worldwide crash of such systems with loss of huge amount of data; and non-malicious small scale attacks caused by mishandling or operational mistakes done by a poorly trained person which may cause minor loss of data or system crash.
- Type of network, where attacks are classified according to the network types, such as Mobile Ad hoc Networks (MANET) and Wireless Sensor Networks (WSN).

Considering the above, the various cybersecurity issues of different SUs can be considered for some forms of integrated classification, upon which work on metrics can be based.

3.2 Metrics for Recommended Standards

Information security management standards provide SUs with protection controls to mitigate or eliminate the impact of potential risks they face. For example, ISO 27001 document [8], which is the core of ISO 27032 concerned with cybersecurity [3], recommends 112 protection controls for the achievement of 35 objectives. These objectives cover 10 main information security problems. Therefore, ISO 27001 document addresses 10 problems, with an average of 3.5 objectives per solving a problem, and 3.2 protection controls per achieving an objective that is 11.2 controls per problem [8].

The protection controls provided by ISO 27001 are statements stating what should be done to contribute to the achievement of an objective. These statements do not provide any metrics to measure the state of application of the various controls in the SUs concerned. Therefore, refining these control statements into specific metrics would enable assessing their application in various SUs, and would support guiding their full application.

3.3 Metrics for SMEs

SMEs are important SUs in the cyberspace. They represent a large proportion of its participants. These enterprises are considered to have some common features with regards to cyberspace, and these are the following [10].

- The owners of SMEs do not usually allocate enough budget to cybersecurity controls.
- IT staff in SMEs are of limited number and with relatively limited security experience.
- One of the difficult challenges of SMEs is how to balance in-house security resources and building a strong team, while also leveraging third-party security services.

Various efforts on assessing and providing metrics for SMEs in the cyberspace have been reported by both: government organizations, such as that in [11]; and individual researchers, such as that in [12]. Reference [11] addresses cybersecurity risks on SMEs and provides an account on their risk management through planning, implementation and reviewing. The paper in [12] addresses the need for cybersecurity metrics to be understood by all those concerned to promote protection. The literature on the subject has, so far, no commonly available metrics to assess SMEs security in the cyberspace.

3.4 Application Specific Metrics (Smartphones)

The cyberspace is full of various applications associated with SUs at different levels. For example, at the user level, different types of access devices used by individuals are connected to various SUs and networks in the cyberspace. Such devices include: personal computers; notebooks; and smartphones, which are currently in use by 2.53 billion people that is 36% of the world population [13].

Cybersecurity for smartphones is an important problem receiving increasing attention, as mobile devices hold sensitive data, and as companies are adopting smartphones for their business. In addition, BYOD (Bring Your Own Device) is becoming a popular phenomenon. In this case, mobile devices are not only holding personal information, but also business data. The security of BYODs has become a new issue for enterprise administrators and IT professionals [14].

As described in [15], mobile security has special features that should be considered:

- Mobility: the device transfers from place to place; therefore, it can be easily stolen or physically tampered.
- Strong personalization: the owner of device is its unique user and his personal information can be stored in it.
- Strong connectivity: a smartphone allows a user to send e-mails, to check online banking account, to access lot of Internet services; so, malware can infect the device, either through SMS or MMS or by exploiting the Internet connection.
- Technology convergence: a single device can come with different technologies: this may enable an attacker to exploit different routes to perform attacks.
- Reduced capabilities: there are some features that lack on smartphones, e.g. a fully keyboard.

Cybersecurity metrics for smartphones is needed for the various SUs using them, in different fields. This is also the case for various other applications, including applications associated with: cloud computing, the Internet of Things (IoT), big data, and others.

3.5 Metrics for Other Considerations

Care about cybersecurity is important for all SUs, concerned with different applications, at different levels, in the cyberspace. With the continuous advancement and increasing innovation in ICT and its applications, newer and renewed SUs are joining the cyberspace, and facing newer cybersecurity problems. Therefore, care about cybersecurity metrics would be essential for future researcher and for professional concerned with protecting the cyberspace.

4 Metrics Development Framework

This section introduces the targeted framework concerned with the development of cybersecurity metrics for SUs. The principles upon which the framework is based are first laid out. This is followed by addressing its construction considering its scope and management requirements. The use of the framework for metrics development is then highlighted.

4.1 Framework Principles

The framework is developed according to the following main principles.

- The framework should be of comprehensive nature that can accommodate the various types of SUs cybersecurity issues.
- It should also be of modular structure that provides flexible construction, and permits updating of issues whenever needed.
- It should provide useful metrics that enable full assessment of the state of SUs cybersecurity.
- It should allow continuous improvement of performance, as time progresses.

4.2 The Framework

According to the principles above, Fig. 1 illustrates the structure of the targeted framework, which enjoys the following three main features.

- It has a wide structured scope that accommodates comprehensive views of different SUs and provides modularity.
- It considers quality management that supports continuous improvement of cybersecurity for different SUs with response to change.
- It supports the development of metrics for various SUs according to the wide structured scope and to the requirement of quality management.

Elaborations on these three main issues are given in the following.

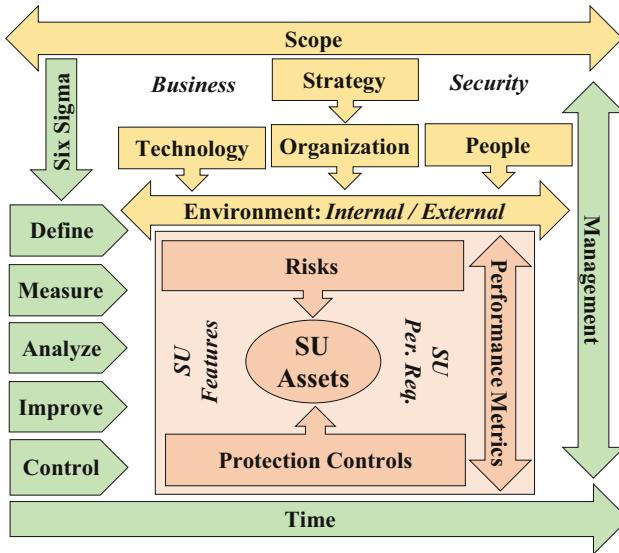


Fig. 1. Metrics development framework for cybersecurity assessment of SUs

4.3 Scope

The chosen wide and modular scope is the STOPE view, which involves looking at the problem concerned considering the five domains of Strategy, Technology, Organization, People, and the Environment. This scope has previously been used in addressing various assessment problems; examples of which are given in the following references [5, 16].

The Strategy of an SU toward cybersecurity would be part of its business strategy. Therefore, SUs may differ about the level of security and privacy they wish to obtain for their business in the cyberspace. In other words, they may differ about the level of residual risk that they can tolerate. In this respect, the cost factor concerned with impact of risk versus protection controls may play a significant role in the strategy. This problem is illustrated in Fig. 2. While some SUs may be more cost sensitive like SMEs, others may be more security and privacy sensitive like banks.

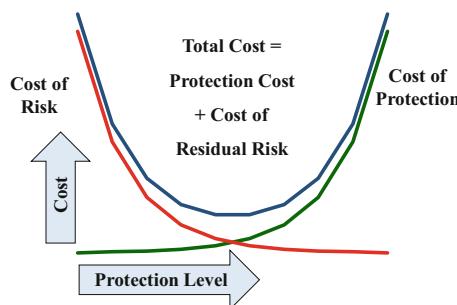


Fig. 2. SUs cybersecurity cost problem

Technology, organizations and people are the main domains according to which SUs assets can be viewed, as mentioned above. Furthermore, according to these domains, and to the environment domain, SUs cybersecurity risks and protection measures can be viewed. With information associated with the state of these domains, metrics concerned with confidentiality, integrity, and availability would also be related to them.

4.4 Management

Management in the framework is based in the six-sigma approach [17]. This approach cares about continuous improvement of performance as time progresses, and this considers continuous monitoring and measurements, which are associated with metrics. In this approach, five main steps are applied periodically to the state of SUs cybersecurity for continuous improvement.

The six-sigma steps, given in Fig. 1, will be applied to the STOPE viewed SU cybersecurity state. They involve: defining the SU cybersecurity state; measuring the state; analyzing the results of the measurement; improving the state according to the outcome of the analysis; and controlling the application of the improvement.

4.5 Metrics Development

Researchers and professionals in the field can use the framework according to the following:

- They can specify various types SUs in the cyberspace including SMEs, hospitals, banks, government departments, and others.
- For every type of SU, they wish to investigate, they can use the STOPE view to define its cyber security state.
- Through STOPE understanding of this state, assessment metrics can be derived considering both strategy of the SU concerned and the CIA performance requirements.
- With the derived metrics, and the STOPE understanding the next steps of six-sigma that is measuring; analyzing; improving; and controlling can be applied.
- Collective experience can be accumulated from applying the above to various types of SUs under different circumstances.
- The collective experience can be used by all for a more comprehensive and integrated view of security in the cyberspace.

It should be finally noted that work in this field will always be active as the generation of new types of SUs, new risks and new protection controls is a continuous phenomenon.

5 Conclusions

This paper has emphasized the need of various SUs in the cyberspace to have cybersecurity metrics that enable them to achieve better security management. To respond to this need, the paper considered elaborating on two main questions. The first is concerned with highlighting this need through addressing the cybersecurity issues, on the one hand; and describing their related metrics development problems, on the other. The second is associated with providing a framework for the development of cybersecurity metrics that enjoys a comprehensive and flexible view, with continuous improvement and updating approach.

The given framework would hopefully provide a base for the development of cybersecurity metrics for various types of SUs, with different requirements.

References

1. Merriam-Webster's Dictionary, <https://www.merriam-webster.com/dictionary/cyberspace>, last accessed 2018/1/1
2. ITU-T X.1205: Overview of Cybersecurity, Series X: Data Networks, Open System Communications and Security, Telecommunication Security (2008)
3. ISO/IEC 27032: Guidelines for Cybersecurity, Information Technology-Security Techniques (2012)
4. NIST: A Framework for Improving Critical Infrastructure Cybersecurity. National Institute of Standards and Technology (2014)
5. Bakry, S.H.: Building the Culture of Responsibility into the Knowledge Society, Global Knowledge Society Forum. ARAMCO, Dhahran, Saudi Arabia (2013)
6. ISO/IEC 27000: Information Technology-Security Techniques-Information Security Management System—Overview and Vocabulary. International Standards Organization, Switzerland (2009)
7. CERT, Computer Emergency Response Team, Carnegie-Mellon University, <https://www.cert.org/>, last accessed 2018/1/1
8. ISO/IEC 27001: Information Security Management: System Requirements, Security Techniques. Information Technology, International Standards Organization, Switzerland (2013)
9. Uma, M., Padmavathi, G.: A survey on various cyber attacks and their classification. Int. J. Netw. Secur. **15**(6), 391–397 (2013)
10. Bresciani, M., Gardner, M., Hickmott, J.: Demonstrating Student Success: A Practical Guide to Outcomes-Based Assessment of Learning and Development in Student Affairs. Stylus Publishing, Sterling, Virginia (2009)
11. HM Government, Small Business: What do you need to know about cybersecurity, UK (2015)
12. Biga, N., Jovanovic, M., Perkovic, M., Mitic, D.: Modern business environment: information technology as a shield against cyber security threats. In: Proceedings of the 8th International Conference on Business Information Security, BISEC 2016, Belgrade, Serbia (2016)
13. Statista (online): Statistics and studies from more than 18000 sources, <https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/>, last accessed 2018/1/1

14. Wang, Y., Vangury, K., Nikolai, J.: MobileGuardian: a security policy enforcement framework for mobile devices. In: 2014 International Conference on Collaboration Technologies and Systems (CTS), pp. 197–202 (2014)
15. La Polla, M., Martinelli, F., Sgandurra, D.: A survey on security for mobile devices. *IEEE Commun. Surv. Tutor.* **15**(1), 446–471 (2013)
16. Bakry, S.H.: Developing security policies for private networks. *Int. J. Netw. Manag.* **13**(3), 203–210 (2003)
17. De Feo, J.A., Barnard, W.W.: *Six-Sigma: Breakthrough and Beyond: Quality Performance Breakthrough Methods*, Juran Institute. Mc-Graw-Hill, New York (2004)



From Access Control Models to Access Control Metamodels: A Survey

Nadine Kashmar^{1(✉)}, Mehdi Adda¹, and Mirna Atieh²

¹ Université du Québec à Rimouski, Rimouski, QC G5L 3A1, Canada

{kasn0002, mehdi_adda}@uqar.ca

² Lebanese University, Hadat, Lebanon

matieh@ul.edu.lb

Abstract. Access control (AC) is a computer security requirement used to control, in a computing environment, what the user can access, when and how. Policy administration is an essential feature of an AC system. As the number of computers are in hundreds of millions, and due to the different organization requirements, applications and needs, various AC models are presented in literature, such as: Discretionary Access Control (DAC), Mandatory Access Control (MAC), Role Based Access Control (RBAC), etc. These models are used to implement organizational policies that prevent the unauthorized disclosure of sensitive data, protecting the data integrity, and enabling secure access and sharing of information. Each AC model has its own methods for making AC decisions and policy enforcement. However, due to the diversity of AC models and the various concerns and restrictions, it's essential to find AC metamodels with higher level of abstraction. Access control metamodels serve as a unifying framework for specifying any AC policy and should ease the migration from an AC model to another. This study reviews existing works on metamodels descriptions and representations. But, are the presented metamodels sufficient to handle the needed target of controlling access especially in the presence of the current information technologies? Do they encompass all features of other AC models? In this paper we are presenting a survey on AC metamodels.

Keywords: Metamodel · Model · Access control · Policy · Security

1 Introduction

As long as there is an increase and development in the use for internet services, there is also an increase and a critical need for computer security solutions [1, 2]. Computer security and the related topics were and still are the main issues in the world of Information Technology (IT).

Over the years until today, IT security and privacy are critical concerns for academic, economic, social, industrial, and governmental organizations. Access Control (AC) is one of the most important and critical aspects of IT security. For this purpose, different AC models and policies are developed. In literature, various AC models are proposed, such as: Discretionary Access Control (DAC) [3, 4], Mandatory Access Control (MAC) [3, 5, 6], Role-Based Access Control (RBAC), Attribute-Based Access

Control (ABAC) [3–5, 7], and Organization Based Access Control (OrBAC) [5]. Each AC model is developed either to overcome limitations found in previous models or as a solution for a specific use case and application.

Finding a unified AC model becomes a significant issue due to the need to include all features offered by the existing AC models, which in some cases, are incompatible and irreconcilable. Also, due to the widespread and heterogeneity of interconnected networks and distributed systems, heterogeneity of platforms and applications, and due to diversity of users, the necessity to design a well coherent AC architecture for enterprises becomes a must. In this context, different AC metamodels, to address these issues, are presented in literature to serve as a unifying framework for specifying and enforcing different AC policies. In the following sections we will summarize existing AC metamodels. Nevertheless, before describing and comparing these metamodels, we have to explore some basic concepts related to AC.

The remaining of this paper is organized as follows: Sect. 2 summarizes the existing AC models, their advantages and limitations. The AC usage in different system levels are presented in Sect. 3. Section 4 describes the concept of AC metamodels, the metamodeling tools, and provides a comprehensive overview about the existing AC metamodels. Section 5 presents the potential research issues. Section 6 concludes this paper.

2 Access Control Models

2.1 A Brief Overview

Access control is defined as an essential security requirement in the field of IT. Each organization has its own information system where a set of policies is defined based on conditions where users can access all or some system resources. Implementing these policies is essential to protect resources. AC methods are carried out at different IT infrastructure levels. They are used in operating systems, databases, networks, and information systems. The goal is to protect files, directories, regulate access to database objects and fields, and protect applications' information (payroll processing, e-health...), etc. However, the primary objective of access controls is the fulfillment of the defined AC policies [1, 2].

Generally speaking [3–5], AC models and mechanisms are defined in terms of subjects, objects and access rights. The subject concept usually refers to a user or program; the object concept refers to an entity a user wants to have access to such as a file, a table or a class. However, a subject may or may not have an access right to an object. Access right means that a subject is able or perform an operation on an object. The operation may be read, write, execute, etc. In other words, a user (subject) can perform an operation (read, write...) on an object (file, class...) if he has a permission to do so. To carry out an operation, access rights are required. To manage who and how operations must be carried out, privileges or AC must be defined. Data resources are protected under different access policies. A model is the projection of the scope of policies and the needed behavior between subject and object entities. Policies are a set

of guidelines, which are generalized, abstracted, formally, or semi-formally described [3]. Several research surveys are presented in literature with detailed descriptions about AC models. In the following sections we summarize the concept of each model.

2.2 Discretionary Access Control (DAC)

In late 1960s, Discretionary Access Control (DAC) model was first introduced by Lampson, a member of a curriculum design team. In DAC, the system protection notion includes three major components: a set of objects, a set of domains, and a matrix. Lampson's work was then extended by Graham and Denning, where the term "subject" was included instead the domain. Then, the extended Lampson's work was developed by Harrison, Ruzzo and Ullman (HRU) to find a formal proof that tracking privilege propagation was undecidable in general [3].

DAC is defined as a user-centric AC model in the sense that a file owner determines the permissions that are assigned to other users requiring access to the file [4]. DAC mechanism allows users control the access rights (read, write...) to their files without the need of a pre-specified set of rules. The access rights are specified by Access Control Matrix (ACM), where AC rights of subject(s) over object(s) are specified. Other variations of implementing AC matrix include Capability Lists (CLs) and Access Control Lists (ACLs). In the concept of CLs the user access rights are stored by rows, whereas in ACLs the access rights for various users on a file are stored by columns. Lampson and Harrison Ruzzo Ullman (HRU) are two DAC models [3].

DAC model is very flexible to assign access rights between subjects and objects. But it also has limitations where the system maintenance and verification of security principles are extremely difficult, since users control access rights to their owned objects. Also, the possible attacks for Trojan horses [3, 5].

2.3 Mandatory Access Model (MAC)

In 1970s, Mandatory Access Control (MAC) protection was presented to include the use of a security kernel. In 1987, a paper was published in IEEE Symposium of Security and Privacy, where crucial differences between commercial and military security requirements were presented by Clark and Wilson [3].

In MAC users cannot define AC rights by themselves. The AC policy is managed in a centralized manner. MAC model is based on the concept of security levels associated with each subject and object, where permissions and actions are derived. Security classes have hierarchical and nonhierarchical components. The hierarchical components include types: unclassified (U), confidential (C), secret (S), and top-secret (TS) where $TS \geq S \geq C \geq U$, to classify subjects and objects into levels of trust and sensitivity. For objects a security level is called the classification level and for subjects it is called clearance level. The nonhierarchical component is represented by a set of categories. Security labels indicate security levels for classification of objects and clearance of subjects, and uses two security properties, simple security property and *-property. Bell and LaPadula (BLP) of multi-level security and BIBA are two MAC variants. In BLP, a subject is allowed to read an object if the subject's clearance is \geq

than the object's classification, and to write if it is \leq . In BIBA, a subject is allowed to read an object if the subject's clearance is \leq than the object's classification, and to write if it is \geq [3, 5].

This model is presented to overcome the limitations of DAC model, which is the Trojan Horse attacks, by centralizing access control. In [6], it is mentioned that MAC model is relatively straightforward and is considered to be a good model for commercial systems that operate in hostile environments such as web servers and financial institutions where the risk of attack is very high. Also, it has limitations, since it is difficult to implement due to the dependence on trusted components, and the necessity for applications to be rewritten to adhere to MAC labels and properties. Similarly, the assignment of security levels by the system places limits on user actions which prevents dynamic modification of the original policies.

2.4 Role-Based Access Control (RBAC)

Based on historical practices, Role-Based Access Control (RBAC) was defined as a job a user performs, and he/she can be assigned one or more roles to indirectly associate permissions with users [3].

RBAC model is considered as an alternative approach to MAC and DAC. In RBAC, users can be assigned several roles and a role can be associated with several users [4]. A role means a collection of permissions to use objects to perform a job function that combines the authority and responsibility assigned to a subject who plays this role, e.g. accountant, director, engineer, etc. each role is associated with privileges or permissions [3]. The aim of RBAC is to facilitate the administration of the AC policy. It governs the access of a user to information through roles for which the user is authorized to perform. RBAC model is based on several entities, which are, users, roles, permissions, actions, operations, and objects. Each role can have many permissions, and permissions may be assigned to many roles. A subject can operate or play many roles and a role can be performed by different subjects [5]. Several RBAC models are proposed in the literature, Flat RBAC (RBAC0), Hierarchical RBAC (RBAC1), Constrained RBAC (RBAC2), and Symmetric RBAC (RBAC3) [5, 7].

This model has many benefits, it has central administration of role memberships and ACs. It may also be applied in distributed areas because it is based on the concept of constraints and inheritance [5, 6]. In RBAC, role hierarchy specifies which roles and permissions are available to subjects based on different inheritance mechanisms. Role hierarchies and permission inheritance in RBAC models are explained in [8, 9]. Also, in distributed areas where different resources are shared among users, RBAC has powerful means of specifying AC decisions [10]. Conversely, it also has some drawbacks. It is frequently criticized for the difficulty of setting up an initial role structure and for inflexibility in rapidly changing IT technologies. For example, RBAC provide poor support for dynamic attributes such as time of day, which might be needed when determining user permission [11]. Another drawback is reflected in large systems, where role inheritance and the need for customized privileges make administration potentially heavy [6].

2.5 Organization Based Access Control (OrBAC)

Organization Based Access Control (OrBAC) is first presented in 2003. The aim of this model is to solve some problems in the previous AC models (DAC, MAC and RBAC), to find a more abstract control policy. It is designed to address the subject, object and action, in such a way that the policy determines what subject(s) has some action(s) to access some object(s). Each organization (clinic, banks, hospitals...) is comprised of a structured group of subjects having certain roles, or entities. This model exceeds the concept of only granting permissions to subjects, it also addresses the concept of prohibitions, obligations and recommendations [5]. A role may have a permission, prohibition or obligation to do some activity on some view given an associated context. In this model seven entities are defined, (a) the abstract level or organizational (1-Role, 2-Activity, 3-View) and (b) the concrete level (4-Subject, 5-Action, 6-Object). The seventh entity is Context lies between the two levels to have a correspondence between the elements of each level. The context is presented in OrBAC to express dynamic rules for relations between entities, for example, Permission, Prohibition, Isprohibited, Recommendation, Ispermitted, Isobligatory, Isrecommended, Obligation [5, 12].

The OrBAC has an advantage in eliminating conflicts between security rules. It also has some vulnerabilities to some kinds of attacks, e.g. covert channels [5].

2.6 Attribute-Based Access Control (ABAC)

Attribute Based Access Control (ABAC) concepts have paralleled that of RBAC. It has some advantages over RBAC, because of its benefits in authorization management and its ability to support dynamic attributes. The main idea behind ABAC is to grant or deny user requests based on arbitrary attributes of the user and selected attributes of the object that may be globally recognized [3, 11]. It enables precise AC which allows a higher number of discrete inputs into an AC decision. This provides a larger set of possible combinations of variables to reflect a larger set of possible rules to express policies [8]. Recently, this model gained attention from businesses, academia and organizations due to the limitations in RBAC model. Two standards that widely address the ABAC framework are: The Extensible AC Markup Language (XACML) and Next Generation AC (NGAC) with AC facility for applications and other important features [3].

In ABAC, subjects are enabled to access a wider range of objects without specifying individual relationships between each subject and each object. As well as, there are three types of attributes: subject, object and environmental attributes. The first two are common to all ABAC models. The third type used in some models that depend on the availability of system sensors that can detect and report values, they may include the current time or day of the week. Despite the benefits of ABAC model over the other models and its flexibility to assign policies and security features, it has a number of limitations such as [2]:

- The problem of determining the current set of permissions available to all users.
- Its implementations require significant time to run.
- It is often not possible to compute the set of users that may have access to a given resource.

- It is difficult to efficiently calculate the resulting set of permissions for a given user as all objects would need to be checked against all relevant policies.

The historical AC models are summarized to provide an overview of their concepts, benefits and limitations. This overview gives a conception about creating new AC models, combining some of their features, or finding models with higher level of abstraction (metamodels).

3 Access Control Usage

This section explains how AC models are generally integrated in any Information System (IS). In general, the IS is composed of six components: hardware, software, data, procedures, people and communication. IT security concepts are related to each component of any IS. For a secure environment, security must be carefully managed. Moreover, any connected computer to the internet is vulnerable to attacks. Thus, the IS components are also under threat. Subsequently, various AC models and principles are deployed to protect the IS environment. Their function is to control which object have access to which subject in the system (files to read, programs to execute, data to share, etc.). Figure 1 shows AC at different levels in a system. Applications may be written on top of middleware, such as a database management system. The middleware use services provided by the underlying operating system. Also, the operating system of ACs usually relies on hardware features provided by the processor or by associated memory management hardware [13].

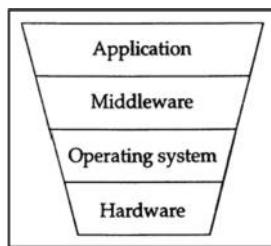


Fig. 1. Access control levels in a system [13]

The IT security requirements are widely explained in literature. Some examples of these requirements are: protection from improper access, user authentication, data integrity, etc. In the following sections we will present AC usage in operating systems, databases, networks, cloud computing, and Internet of Things (IoT).

3.1 Access Control in Operating Systems

Access control mechanisms are provided with Operating Systems (OSs) to authenticate system administrators and users using some procedures, for example, passwords. After authentication, AC play an important role in allowing users to access files,

communications ports, and other system resources. User permissions are modeled as a matrix of access permissions, where columns represent files/folders and rows represent users. In this context, a file owner determines the permissions that are assigned to other users requiring access to the file. However, only these users may have some permissions (privileges), on the file, to read, write, execute, etc. An ACM is used to show the users' access rights on a system and the files in it [4]. Although ACMs can be used to implement protection mechanisms, they don't scale well with many users. For example, in an institution with a large number of users and applications, a plenty of entries will occur and this will cause a performance problem and administrative mistakes. Two practical ways to overcome those issues are presented in [13]. Firstly, by using groups or roles to manage the privileges of large sets of users simultaneously. Secondly, by storing the ACM either by columns (ACL) or rows (CLs).

3.2 Access Control in Databases

Today, it is common for institutions to have databases (DBs) with critical data, such as: bank accounts, employment records, hospital reports, etc. Thus, the security needs become very critical, especially for online users. DB management systems (Oracle, SQL...) have their control mechanisms and users should have different types of permissions based on their job functions. In this context, OS play an important role to separate the DB from other applications running on the same computer by identifying users. Besides, ACLs and CLs are mixed to provide the needed mechanisms for DBs [13]. Database users should have different types of permissions based on their job functions. Each user should only have the permissions which are granted for him after his login to the system, which is known as authorization. These permissions are assigned to determine the user actions within the DB.

Different DB tasks and maintenance procedures must be occurred, these tasks must be implemented by DB administrators. These tasks include creating DBs, removing unneeded DBs, managing disk space allocation, monitoring performance, and performing backup and recovery operations. DB platforms allow default system administrators to perform such tasks and delegate permissions to other users [14].

3.3 Access Control in Networks

Network (NW) security is also an essential part of IS. In this domain, different issues must be taken into consideration, which are: the NW design, NW device security, firewalls, virtual private NWs, Intrusion Detection and Prevention Systems (IDPS), etc. The section below describes and summarizes these considerations [15].

The requirements of NW security and budgets are specified based on NW design. For this, different aspects must be taken into consideration in designing NWs, such as: availability, cost, performance, number of users, etc. In security, it is important to enable effective and secure links to other NWs, provide a platform that is helpful for securing sensitive NW assets, and identify critical security controls and understand the consequences of a failure of these controls. Also, NW device security concern is how to use routers and switches to increase the security of the NW. The internetworking protocol in use today is known as Transmission Control Protocol/Internet Protocol

(TCP/IP). It is a suite of protocols and applications that have discrete functions that map to the Open Systems Interconnection (OSI) model. Each connected device on a NW has two NW addresses: The Media Access Control (MAC) address, and the IP address. However, there are several configuration steps to configure the device (router, switch...) for increased security. The various steps are: switch security practices, ACLs, administrative practices, Internet Control Message Protocol (ICMP), logging to routers, etc. Furthermore, firewalls are defined as the first line of defense between the internal NW and untrusted NWs like the Internet. They play an important role in controlling application communication and other functions, such as: Network Address Translation (NAT), antivirus, e-mail (spam) filtering, IDPS, etc. Virtual Private Networks (VPNs) are created by establishing a virtual connection using dedicated connections is to provide a secured communication channel over public NWs. To secure VPNs, different issues must be addressed, e.g. the authentication process, client configuration, etc. A security administrator always checks the system and security log files looking for something abnormal. Audit tools are used by administrators to detect a wide range of rogue events, such as: unauthorized registry changes, protocol attacks, Denial of service (DoS) attacks, etc. Also, for website security different methods are handled, e.g. prevention of SQL injection, using complex passwords, management of cookies, and others.

The AC models are developed to match the security needs in all aspects of IT. In the presence of new technologies, such as: Cloud Computing and IoT, it is worth presenting some AC mechanisms in such fields.

3.4 Access Control in Cloud Computing

Cloud Computing (CC) is an emerging technology. Due to the huge amount of data which are generated from different end users' applications and information systems, CC is considered as an efficient solution for easier and faster storage retrieval of data. In CC, users can access computer services via the internet and this makes their data vulnerable to attacks. For this reason, different AC methods for CC are presented in literature. This section presents some of AC in CC methods.

Onankunju, in [16], introduces a method for providing secure AC in CC after presenting the possible CC attacks, e.g. DoS and authentication attacks. Hence, a hierarchical structure using a clock is presented to upload, download, and delete files to and from the cloud. The root of the hierarchical structure is the trusted authority which authorizes the top-level domain authorities. Also, domain authorities authorize cloud users. The system is composed of 4 parts: cloud owner, untrusted cloud, clock and cloud users. The user, with a key, encrypts his data before uploading it to the untrusted cloud. For the user, to access his data, he should send a request to the cloud owner which in turn sends him back a key. The key remains available for a certain period, then it becomes invalid after the clock stops counting. The user should access his data within the time limit. Another method is proposed in [17]. The method avoids using static passwords, it uses a one-time password and one day password. Whereas, the first password expires in two minutes, and the second one after twenty-four hours. The user receives passwords with encryption via e-mail for each login session.

3.5 Access Control in IoT

In [18], IoT is defined as “a world-wide network of interconnected objects uniquely addressable, based on standard communication protocols”. Which means a huge number of heterogeneous objects, with different technologies and platforms, are communicating together via the internet. Hence, all devices connected to the internet are vulnerable to attacks, and this is the case for IoT devices. In this context, various authentication and AC methods in IoT are also presented in literature, to integrate security issues with this technology.

Liu et al. in [19] propose a model to find a secure communication between things by a certain procedure. The main idea of this procedure is based on verifying identities between two IoT devices. The method is based on implementing authentication protocol in the presentation layer, where identification key establishment occurs. They adopt the concept of authorization in RBAC model, and for secure key establishment they implement Elliptic Curve Cryptosystem (ECC). Moreover, authors in [20] propose a “smart contract-based framework to implement distributed and trustworthy access control”. Their aim is “to apply the smart contract-enabled blockchain technology to achieve distributed and trustworthy AC for the IoT”. This framework contains multiple Access Control Contracts (ACCs), one Judge Contract (JC), and one Register Contract (RC). ACCs are implemented between subjects and objects for AC. JC is used for judging the unpleasant behavior of the subject during AC. RC is used to manage the ACCs and JC. To demonstrate the framework feasibility, case studies are addressed.

Different other AC proposals are presented in this field, which reflects the evolutionary stage of CC, IoT, and security concerns due to the presence of attacks.

4 Access Control Metamodels

The need to use AC methods in different system levels and technologies, imposes the necessity of finding AC models with combined features from two or more models. Due to the continuous increase and upgrade of information technology features, the presence of security threats also increases. The technological environment which is open to all types of users is a crucial concern, because it is also open to various types of attacks. This makes security enforcement, through AC models, an urgent need. So, many AC models are presented in literature with combined features from two or more AC models based on research motivations and needs. For example, we can find many models with combined features from both RBAC and ABAC. Some hybrid RBAC and ABAC models and others are presented for this purpose.

In [11] and due to RBAC’s difficulty to set up an initial role structure in rapidly changing environments, and because RBAC does not support dynamic attributes, the idea of adding attributes to RBAC is addressed. The aim is to find a model that supports dynamic attributes specially in organizations. These features are presented to handle relationship between roles and attributes to provide better AC features in dynamic environments. Also, authors in [21] target the idea of enhancing features from both RBAC and ABAC, because both have complimentary features to each other. Hence, Attribute Enhanced RBAC model (AERBAC) is presented. The model “retains the

flexibility offered by ABAC, yet it maintains RBAC's advantages of easier administration, policy analysis and review of user permissions [21]".

However, AC models must consider the continuous developments and changes. The new technologies (CC, IoT...), the variety of platforms and applications, users' types, etc. comprise a complex fact in controlling secure and private accesses to the needed resources in different areas. All this makes AC models and even combining some features of them are insufficient to handle the needed target. This fact forces the need to find models with higher level of abstraction, which is called AC metamodels. The aim of AC metamodels is to serve as a unifying framework for specifying and enforcing any AC policy. For this purpose, different research works are present and still conducting for finding an AC metamodels. The following sections present some existing AC metamodels.

4.1 Metamodel Definition

Before exploring the existing AC metamodels, it is worth mentioning some essential definitions and ideas about metamodeling. Metamodel in [22] is defined as a textual, graphical/visual, or formal representation of concepts and how they are linked together. In other words, it is a structure of a collection of concepts in a certain domain. These concepts might be terms, rules, guidelines, etc. for an institution or organization. In [23], metamodeling is defined as modeling of a model, where they should describe the permitted structure to which models must adhere. Furthermore, models and metamodels need adaptable supporting tools due to changing requirements and policies. As mentioned earlier, metamodels can be illustrated using textual, graphical or formal representations. Though, different metamodeling tools and languages are presented in [22, 23] such as: Unified Modeling Language (UML), Meta-Object Facility (MOF), Eclipse Modeling Framework (EMF), MetaEdit, Conceptbase, and other tools. Figure 2 shows the metamodel abstraction levels. M_3 is an instance of a model, it defines a specific information domain. M_2 is an instance of metamodel, it defines a language to describe an information domain. M_1 is an instance of a Meta-metamodel, it defines the language for specifying a model. M_0 is the infrastructure for a metamodeling architecture, it defines the language for specifying metamodels [22].

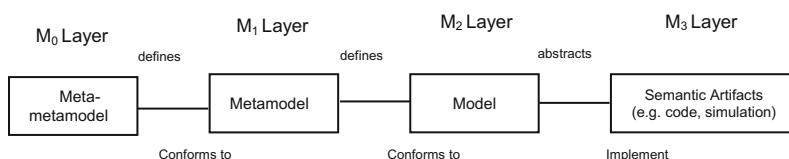


Fig. 2. The four-layer metamodeling architecture [23]

4.2 State of the Art

Korman et al. in [24] propose a unified metamodel designed using Enterprise Architecture (EA) modeling language, the ArchiMate, and explain its use on many scenarios

and two business cases. Their aim is to combine different AC models in a single EA model, and to propose an extension to an established EA modeling language. Their model targets the challenge of flexibly modeling policies of authorization according to the most well-known AC models: DAC, MAC (BPL, BIBA, and Chinese Wall), RBAC, and ABAC, in terms of EA. The authors summarize some of the existing AC models, then define the vocabulary of these models, such as: subjects, objects, sessions, etc. to use them later in their presented metamodel. Then, they represent the conceptual models for the most basic common terms of AC, which are: subject, object, and access mode for the same purpose of later use. Also, the configurations of DAC, BLP, BIBA, Chinese Wall (CW), RBAC_{0,1,2,3}, and ABAC are represented as metamodels, e.g. in Fig. 3 they represent the metamodel for ABAC. Finally, these predefined steps are used as an introduction before presenting their unified metamodel in Fig. 4. The model mostly builds on the conceptual model of ABAC for its ability to emulate most functions of the other AC models.

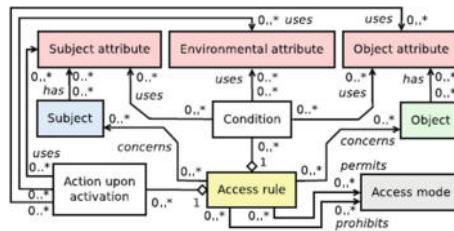


Fig. 3. Metamodel for expressing configurations of ABAC [24]

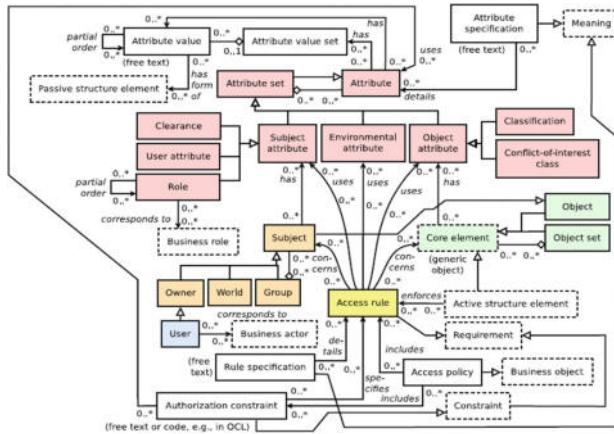


Fig. 4. Unified metamodel for modeling authorization [24]

Authors in [25] propose an AC metamodel to concurrently handle multiple AC models. Their metamodel consider four models, which are: CW, BLP, BIBA, and RBAC models. They start from the general concepts of metamodels; the object, subject, access mode, role, and the instances of associations between metamodeling elements (subject and role). They present the decision concept which relies behind applying logical rules that are expressed in terms of AC metamodel elements. As proposed, each AC metamodel has a special element called Decision Handler. Figure 5 illustrates the initial concept of their metamodel. Then, kernel AC elements are presented to later illustrate the associations between metamodel elements. These elements are: Object, Objects Group, Subject, Subjects Group, Access Mode, Additional Attribute, Query, Environmental Attribute, and Decision. Objects Group and Subject Group respectively represent sets of objects and subjects sharing some properties. The AccessMode represents, for example, some actions like: read, write, execute, etc. Query is the access request on an object by a subject. Environmental Attribute is used to hold information related to access request events such as time, place, temperature, etc. Additional Attribute represents a construct associated to a metamodel element to support the specification of some property of that element. Decision (e.g. permit, deny, indeterminate, NotApplicable...) is where the response is issued by the AC request. The metamodel is implemented with the four AC models (CW, BLP, BIBA, and RBAC) as shown in Fig. 5. ACmetaModelElement is a generalization of any element of the AC metamodels other than DecisionHandler. The proposed AC metamodels relies on a subset of UML but without determining how an instance of the metamodel returns an AC decision. To do so, First Order Logic (FOL) mapping is used for relating entities to their types, specifying relationships between entities, and for expressing a decision logic based on relations. To specify the integration of several AC metamodels of a hybrid AC policy, they presented an example of Integration Metamodel (IM) based on Ascending Decisions Tree (ADT). Hybrid AC policies mean applying multiple AC specifications and policies. Figure 6 shows the ADT metamodel example which concludes with only one AC decision as output, in response to a set of multiple AC decisions as input. ADT nodes are, DecisionHandler instances and nodes applying Combining Algorithms (ComAl). DecisionHandler instance issues its AC decision based on its metamodel decision logic (explained in [25]). ComAlNode issues its

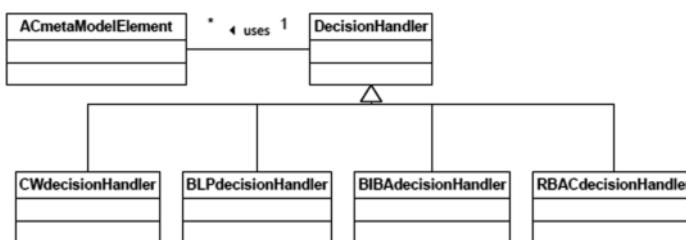


Fig. 5. DecisionHandler specializations in access control metamodels [25]

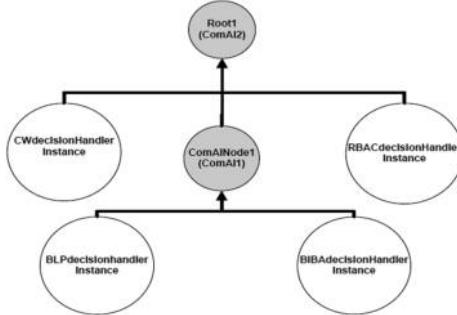


Fig. 6. An ADT integrating multiple AC metamodel instances [25]

decision by applying its ComAI on the decisions of its direct children nodes. It has a unique root which returns the decision of the whole tree with the hybrid AC policy. Finally, the proposed IM metamodel is illustrated in Fig. 7, which encompasses the whole above concepts.

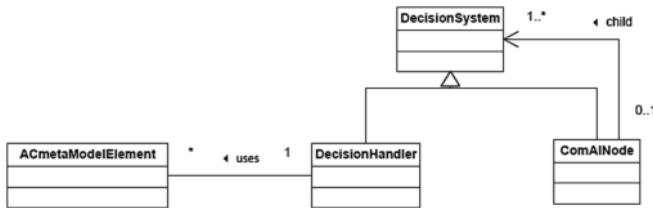


Fig. 7. The main view of IM metamodel [25]

Besides, designing AC metamodels for distributed environments (consisting of several sites) are also taken into account. This is due to the importance of finding dynamic or collaborative policies, to consider system changes or collaborate with other policies of several sites. This concept is covered in [26], where authors present the importance of finding a formal specification language to define AC models and policies in such environments. For this purpose, the concept of rewriting techniques (for security policies and protocols) is suggested to provide semantics for distributed AC mechanisms. Also, they mention some several available rewrite-based programming languages for fast prototyping, such as, Muade. Their proposed rewriting techniques are defined as an instance of a metamodel based on Distributed Event Based AC model (DEBAC). However, the authors first explain the advantages of rewriting systems, then explore the existing AC models to introduce the concept of a unifying metamodel for AC. Also, they describe the main features of the AC metamodel and define the extension of the metamodel for distributed environments. Second, they propose a formal specification of the distributed metamodel in a rewrite-based language, constructed on core concepts of AC models, and focus on the modular properties of the system. In this context, the federation notion is considered, like database systems where

several systems are integrated by a federated system and each system preserves its autonomy. Third, general policy combining operators are well-defined to define combinations of policies. Algebraic terms are used in all the steps to define and rewrite policies, properties, operational semantics of the distributed metamodel, and integrating combination operators in the distributed metamodel. In [26], detailed explanations about the used expressions are also provided. Similarly, a metamodel extension mechanism is proposed as a solution in the context of MoNoGe French collaborative project [27]. This project is based on a textual Domain Specific Language (DSL). The aim is to face up current limitations and the lack of standard solutions in the existing project. In addition to building a generic lightweight metamodel extension approach for the industrial environment where rapid and efficient adaptations of the used modeling tools are needed. Authors begin by defining the concept of modeling in real industrial projects, which deal with different models and metamodels, in addition to supporting tools. Hence, they present the main industrial use case of MoNoGe, which comes from DCNS (a world-leading company in naval defense and energy that especially develops Combat Management Systems, CMS, for ships). DCNS use two separate modeling tools: DoDAF (U.S. Department of Defense Architecture Framework) standard, and Modelio supporting software design and development. Then, a metamodel extension operators and a DSL are introduced to easily use them. Also, two different implementations of their proposed extension mechanism, based on Eclipse/EMF and the Modelio modeling environment, are also described. Figure 8 shows a sample of the grammar of their proposed metamodel extension textual DSL.

```

Model:'define' extensionName=ID 'extending' metamodel+=Metamodel ':'
prefix+=Prefix (" " metamodel+=Metamodel ':' prefix+=Prefix)*
'{ extensions += Extension* '}';
Extension: Create | Refine | Generalize | ModifyClass | FilterClass;
Metamodel: name=ID;
Prefix: name=ID;
Create: 'add class' class=ID;
Refine: 'add class' classNew=ID 'specializing' prefix=[Prefix] '.'
    classOriginal=ID;
Generalize: 'add class' classNew=ID 'supertyping' prefix=[Prefix]
    '.' class+=ID ("," prefix=[Prefix] '.' class+=ID);
ModifyClass:
    'modify class' prefix=[Prefix] '.' class=ID '{'
        modifyOperators += ModifyOperator*
    '}';
ModifyOperator: AddProperty | ModifyProperty | FilterProperty |
    AddConstraint | FilterConstraint;
AddProperty: 'add property' property=ID 'type' type=ID;
ModifyProperty: 'modify property' property=ID value+=ValueAssignment
    ("," value+=ValueAssignment)*;
ValueAssignment: attribute=ID '=' value=EString;
FilterProperty: 'filter property' property=ID;
FilterClass: 'filter class' prefix=[Prefix] '.' class=ID;
AddConstraint: 'add constraint' constraint=ID value=EString;
FilterConstraint: 'filter constraint' constraint=EString;

```

Fig. 8. The grammar metamodel extension textual DSL [27]

Furthermore, AC metamodels also consider the Web Content Management Systems (WCMSs). WCMSs are frameworks that are widely used for web applications development for enterprises, e.g. Drupal, Wordpress, and Joomla. Users with little technical knowledge can fully develop technical systems due to their integrated environment. This environment provides design definition, layout, content management and

organization of the application. In this context, Martínez et al. in [28] highlight the importance of security requirements, since WCMSs may contain sensitive information. Authors propose a metamodel to the representation of WCMS AC policies, to ease the analysis and manipulation of security requirements by abstracting them from vendor-specific details. They focus on the idea of facilitating WCMS configuration, to minimize the possibility of vulnerabilities because users often lack depth technical and security knowledge. Besides, authors enumerate some law-level security aspects, for example, management of cookies, and prevention of SQL injection vulnerabilities. Although AC techniques are integrated in most WCMS systems, some limitations still exist in such systems, e.g. knowing the level of protection of the implemented access policy in a WCMSs is complex and error-prone task. For this purpose, authors propose to raise the level of abstraction of the AC implementation, to be represented according to a vendor-independent metamodel. Figure 9 represents the proposed WCMS metamodel sample, which is inspired by RBAC.

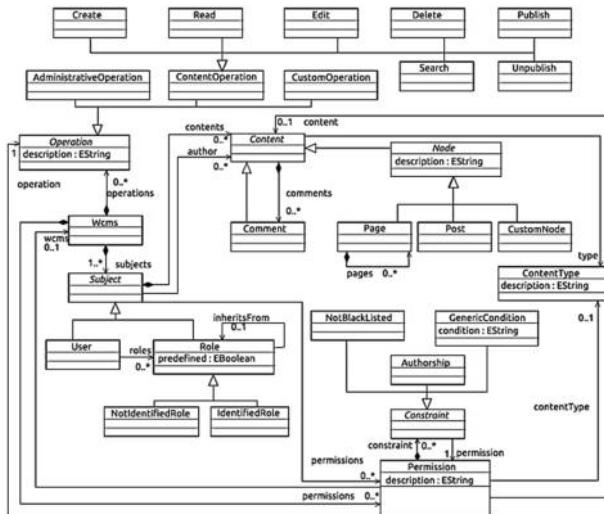


Fig. 9. WCMS metamodel sample [28]

The four metamodel basic elements functionalities explained in [28] are: content, actions, permissions and subjects. The idea of their process is to automatically extract the AC information in the domain of WCMSs. Moreover, they mention that even their metamodel could be manually filled by investigating AC information using WCMS administration tools, it should also be filled by an automatic reverse engineering approach. Thus, they present an automatic process for Drupal in Fig. 10, where Drupal contents with AC information are stored in the backend database. SQL queries over the database are injected to obtain a model that conform their proposed WCMS metamodel. This process allows the possibility of defining extra AC rules or modifying them programmatically. The abstract representation of the WCMS AC model is

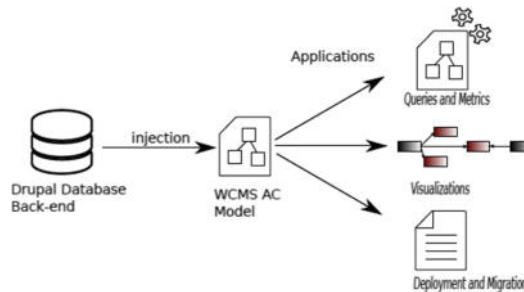


Fig. 10. Drupal access control extraction [28]

developed using Model Driven Engineering (MDE) where the relation between metamodel elements can be easily realized. Concerning WCMS migration, they present using their metamodel as a pivot representation. This illustration is to represent the AC information of the old WCMS in a way corresponding to their metamodel, to facilitate its analysis. Another web service metamodel is proposed in [29] to handle the verification of authorization in Web Service Oriented Architecture (WSOA). The aim of this metamodel is to improve the existing AC models for better features to match the requirements of WSOA. The proposed metamodel is depicted in Fig. 11. The metamodel is an enhancement for hierarchical RBAC and ABAC. Conceptual UML modeling is used to present the metamodel and define the sets and relations. In this metamodel the commonly used type of operations, e.g. read, write, which are carried between permission and object elements are removed. Instead, the focus on placing the input parameters of web service operation. As shown in Fig. 11, there are two relations from permission to an object: (1) indirect relation via policy, (2) direct relation to the input parameter, where “a parameter does not need to know if its value is evaluated for access control [29]”. Web services Composition “consists of multiple invocations of other Web Service Operations in a specific order [29]”. This element plays an important role to stop execution in case of missing authorization at an early stage. The proposed

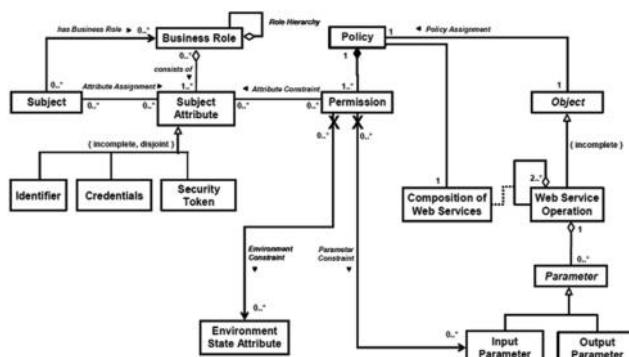


Fig. 11. Metamodel for AC in WSOA [29]

metamodel is mapped to an authorization verification service, which is part of Identity Management (IdM) architecture. Then it is linked with the core concern of WSOA, and the feasibility of this approach is illustrated in a case study. Also, Web Services Description Language (WSDL) is used for service interface definition. Likewise, addressing network security is also a critical concern. Martínez et al. in [30] propose a model driven approach to extract network AC policies enforced by firewalls within a network system. Their concept tackles the problem of filtering the traffic of a network with the presence of a number of filtering rules. based on their analysis, “the network topology, that may include several firewalls, may impose the necessity of splitting the enforcement of the global security policy among several elements. [30]”. In this context, their aim also is “to raise the level of abstraction of the information contained in the firewall configurations files so that the AC policy they implement is easier to understand, analyze, and manipulate [30]”. However, Eclipse tool (Xtext) is used to extract AC information out of the net-filter iptables language. The features of RBAC and OrBAC AC models are implemented in this approach. Figure 12 shows the network connection metamodel. The metamodel consists of two entities, host and connection. The former represents a network host, e.g. IP address. The latter represents connections between hosts, where the port and the protocol are specified to establish connections and specify if the connection is allowed or denied.

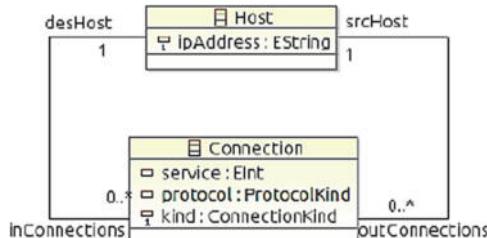


Fig. 12. Network connection metamodel [30]

5 Potential Research Issues

Developing AC metamodel, that covers the features of all other AC models, is a challenging issue especially if the aim is to find a metamodel that encompasses all existing AC models. The dynamic requirement for enforcing security issues and the rapid propagation of technology makes it an urgent need.

The presented metamodels come with some advantages, and many case studies are addressed to handle each metamodel design. But we notice that each metamodel is itself a case and does not encompass a general base concept. Table 1 summarizes the presented AC metamodels, which depicts different metamodels in IT systems. Also, it indicates that metamodeling is a recent research field, especially in the last few years. As we notice, all the presented metamodels are designed for dedicated case scenarios or

Table 1. Summary of presented access control metamodels

Metamodel features					
Ref.	Publication	Designed for	Type	Used Models	Modeling tools
[24]	2016	Enterprise architecture	Unified	DAC, MAC RBAC, ABAC	ArchiMate
[25]	2015	Enterprise	Hybrid	CW, BLP, BIBA, RBAC	UML, FOL
[26]	2014	Distributed environment	Metamodel extension	DEBAC	Muade
[27]	2015	Industrial project	Metamodel extension	MoNoGe project	DSL, EMF, Modelio
[28]	2013	WCMSs	Metamodel extraction	RBAC	Drupal, MDE
[29]	2007	Web service	Metamodel integration	RBAC ₁ , ABAC	UML, WSDL
[30]	2012	Network firewalls	Metamodel extraction	RBAC, OrBAC	Xtext

projects based on some features of AC models. Furthermore, in addition to the achieved progressions, there are still issues to be addressed. In fact, despite the advantages of the presented metamodel in [24], authors present some of its limitations. For example, it misses the concept of logging, in addition to the difficulty for a potential implementation of automated analytical capabilities of the unified metamodel. So is the case for the other metamodels, they are not generic enough to include all AC models features. As we can see some combined features from some models to cure some existing deficiencies in some projects or enhancing some service features. In addition to the concept of applying the same complex process in assigning relationships between model elements in some metamodels.

In this paper, we try to spot on the idea of metamodels, the existing metamodels in literature, and to look into future with some raised questions concerning this matter. Subsequently, in addition to the existing concerns and metamodels, many questions are raised, such as: is it possible to find a more general concept of metamodels? Is it possible to implement easier and general unified structures of metamodels, and visualize their elements more readily? Are the existing metamodels handle the feature of flexibility for any new extensions or transformations? Or, are the current metamodeling frameworks flexible and dynamic enough for any changes? If so, what are the possible ways, steps or strategies to merge metamodel elements? Although there are many metamodels are built, based on many AC models, do they overcome the existing limitations of these AC models? Last but not least, is the existing metamodeling tools and languages enough to answer all the above questions or some of them?

Additionally, as presented in this survey we can see that metamodels are implemented for different scenarios: AC models, WCMSs, and distributed environments. Thus, is there any opportunity to find a metamodel design or plan that encompasses the

different scenarios? Or is it more efficient to find a unified metamodel for each scenario? As a result, currently we may not have answers to the above inquiries, but at least we know that metamodels introduce a new era of enforcing policies and controlling access in IT world.

Another interesting feature, that is missing in current AC metamodels, is the ease of the migration from an AC model to another. In fact, having a metamodel, should make it possible to translate an existing AC policy between the different AC models covered by the metamodel.

6 Conclusion

We covered in this survey existing AC metamodels, and the AC models they generally cover (DAC, MAC, RBAC...). We also presented a brief explanation about how these AC models are used in different fields e.g. Databases, operating systems, IoT, etc. Moreover, Metamodels are proposed in literature to concurrently handle multiple AC models, also in distributed environments and web systems. The main goal is to develop a metamodel that is general enough to instantiate all existing AC models and that may also help organizations to easily migrate from an AC model to another.

Acknowledgements. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number 06351].

References

1. Matt, B.: *Introduction to Computer Security*. Pearson Education India (2006)
2. De Capitani di Vimercati, S., Paraboschi, S., Samarati, P.: Access control: principles and solutions. *Softw. Pract. Exp.* **33**(5), 397–421 (2003)
3. Hu, V.C., Kuhn, D.R., Ferraiolo, D.F.: *Attribute-Based Access Control*. Norwood, Artech House (2018)
4. Kayem, A.V., Akl, S.G., Martin, P.: A presentation of access control methods. In: *Adaptive Cryptographic Access Control*, pp. 11–40. Springer, Berlin (2010)
5. Ennahbaoui, M., Elhajji, S.: Study of access control models. In: *Proceedings of the World Congress on Engineering* (2013)
6. Ausanka-Crues, R.: Methods for access control: advances and limitations. *Harvey Mudd Coll.* **301**, 20 (2001)
7. Sandhu, R., Ferraiolo, D., Kuhn, R.: The NIST model for role-based access control: towards a unified standard. In: *ACM workshop on Role-Based Access Control* (2000)
8. Crampton, J.: On permissions, inheritance and role hierarchies. In: *Proceedings of the 10th ACM Conference on Computer and Communications Security*. ACM (2003)
9. Belokosztolszki, A.: *Role-based access control policy administration*. University of Cambridge, Computer Laboratory (2004)
10. Zhang, C.N., Yang, C.: Designing a complete model of role-based access control system for distributed networks. *J. Inf. Sci. Eng.* **18**(6), 871–889 (2002)
11. Kuhn, D.R., Coyne, E.J., Weil, T.R.: Adding attributes to role-based access control. *Computer* **43**(6), 79–81 (2010)

12. OrBAC: Organization Based Access Control. 2010; Available from: http://orbac.org/?page_id=21
13. Anderson, R.: Security Engineering. Wiley, New York (2008)
14. Rhodes-Ousley, M.: Information Security: The Complete Reference. McGraw Hill Education (2013)
15. Rajpoot, Q.M., Jensen, C.D., Krishnan, R.: Attributes enhanced role-based access control model. In: International Conference on Trust and Privacy in Digital Business. Springer, Berlin (2015)
16. Onankunju, B.K.: Access control in cloud computing. *Int. J. Sci. Res. Publ.* **3**(9), 1 (2013)
17. Hussain, S.: Access control in cloud computing environment. *Int. J. Adv. Netw. Appl.* **5**(4), 2011 (2014)
18. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. *Comput. Netw.* **54**(15), 2787–2805 (2010)
19. Liu, J., Xiao, Y., Chen, C.P.: Authentication and access control in the internet of things. In: 2012 32nd International Conference on Distributed Computing Systems Workshops (ICDCSW). IEEE, New York (2012)
20. Zhang, Y., Kasahara, S., Shen, Y., Jiang, X., Wan, J.: Smart Contract-Based Access Control for the Internet of Things (2018). arXiv preprint [arXiv:1802.04410](https://arxiv.org/abs/1802.04410)
21. Rajpoot, Q.M., Jensen, C.D., Krishnan, R.: Integrating attributes into role-based access control. In: IFIP Annual Conference on Data and Applications Security and Privacy. Springer, Berlin (2015)
22. Assar, S.: Meta-modeling: concepts, tools and applications. In: IEEE 9th International Conference on Research Challenges in Information Science, IEEE RCIS 2015, Athens, Greece; Available from: <https://www.computer.org/cms/ComputingNow/education/said-assar-metamodeling-tutorial.pdf>
23. Sprinkle, J., Rumpe, B., Vangheluwe, H., Karsai, G.: 3 Metamodelling. In: Model-Based Engineering of Embedded Real-Time Systems, pp. 57–76. Springer, Berlin (2010)
24. Korman, M., Lagerström, R., Ekstedt, M.: Modeling enterprise authorization: a unified metamodel and initial validation. *Complex Syst. Inf. Model. Q.* **7**, 1–24 (2016)
25. Abd-Ali, J., El Guemhioui, K., Logrippo, L.: A metamodel for hybrid access control policies. *JSW* **10**(7), 784–797 (2015)
26. Bertolissi, C., Fernández, M.: A metamodel of access control for distributed environments: applications and properties. *Inf. Comput.* **238**, 187–207 (2014)
27. Bruneliere, H., Garcia, J., Desfray, P., Khelladi, D.E., Hebig, R., Bendraou, R., Cabot, J.: On lightweight metamodel extension to support modeling tools agility. In: European Conference on Modelling Foundations and Applications. Springer, Berlin (2015)
28. Martínez, S., García-Alfaro, J., Cuppens, F., Cuppens-Boulahia, N., Cabot, J.: Towards an access-control metamodel for web content management systems. In: International Conference on Web Engineering. Springer, Berlin (2013)
29. Emig, C., Brandt, F., Abeck, S., Biermann, J., Klarl, H.: An access control metamodel for web service-oriented architecture (2007)
30. Martínez, S., Cabot, J., García-Alfaro, J., Cuppens, F., Cuppens-Boulahia, N.: A model-driven approach for the extraction of network access-control policies. In: Proceedings of the Workshop on Model-Driven Security. ACM (2012)



A Potential Cascading Succession of Cyber Electromagnetic Achilles' Heels in the Power Grid

The Challenge of Time Synchronization for Power System Disturbance Monitoring Equipment in a Smart Grid Amidst Cyber Electromagnetic Vulnerabilities

S. Chan^(✉)

Vit Tall Cyber Unit & IE2SPOMTF, San Diego, CA 92192, USA
schan@vittall.org

Abstract. The instantiation of various Phasor Measurement units (PMUs) at pertinent, disparate points across a smart grid facilitates certain insights. Time synchronization among the involved PMUs and other involved disturbance monitoring equipment (DME) is vital for situational awareness within the smart grid so as to avoid catastrophic failures, such as large-scale blackouts. Current PMUs often utilize Global Positioning System (GPS) substation clocks for time synchronization. From a cybersecurity vantage point, this subjects the PMUs, and in turn, the entire involved smart grid, to various cyber electromagnetic vulnerabilities, such as GPS blocking/jamming, and spoofing. Current mitigation strategies are not yet robust enough and need buttressing. In this paper, an architectural schema is proposed, wherein a modified Best Master Clock Algorithm (BMCA) (equipped with a modified “Compare Unit”) is executed along different pathways and then harmonized, via a modification of an N-Input Voting Algorithm (NIVA). A modified Fault Tolerant Average Algorithm (FTAA) is then applied against the results of the various NIVAs so as to determine a Master Clock Group (MCG). Variants of sync integrity protection mechanisms (SIPMs) were utilized prior to the final confirmation of the Grand Master Clock (GMC) election and prior to any time information being utilized and/or syndicated for data synchronization and/or event correlation purposes.

Keywords: Phasor measurement unit (PMU) · Smart grid · Time synchronization · Disturbance monitoring equipment (DME) · Situational awareness · Large-scale blackouts · Global positioning system (GPS) · Cyber electromagnetic vulnerabilities · Timestamping · Best master clock algorithm (BMCA) · Data synchronization · Event correlation

1 Introduction

Power system monitors, specifically power system disturbance monitoring equipment (DME), include devices, such as the Dynamic Disturbance Recorder (DDR), Digital Fault Recorder (DFR), Phasor Measurement Unit, Digital Protective Relay, Power

Quality (PQ) Analyzers, and Traveling Waves Fault Locators (TWs FL), among others. These devices, particularly the PMU, are an essential part of the modern-day smart grid, and they all rely upon the correct correlation of data to time (i.e. accurate timestamping) to perform their respective functions.

As an exemplar, taking the PMU, a closer look is taken at the source of its timestamping. Electric utilities tend to utilize Global Positioning System (GPS)-based substation clocks (with oscillators)—equipped with GPS receiver modules/chips—to ensure accurate timestamping. The current prototypical interplay among the described components (i.e. GPS satellite, GPS antenna, GPS substation clock, oscillator, GPS receiver chip, DFR, PMU, Digital Protective Relay, PA Analyzers, TWs FL, etc.) is portrayed in Fig. 1. As can be seen, there is a high dependency of the DMEs upon the GPS substation clock, particularly its oscillator and its GPS receiver chip, to be able to perform accurate timestamping.

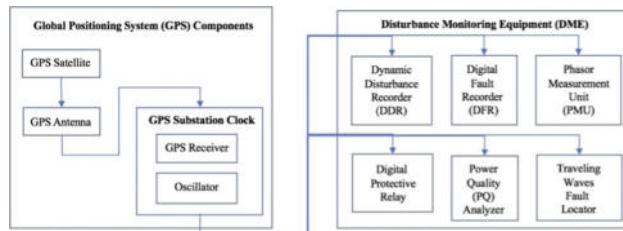


Fig. 1. Prototypical interplay among global positioning system (GPS) components and disturbance monitoring equipment (DME)

Indeed, GPS substation clocks have become foundational to PMUs and the ecosystem of DMEs. Yet, despite this criticality, GPS substation clocks are susceptible to a variety of issues and represent a potential cyber “Achilles heel” for the modern-day smart grid. Along this vein, the term of art “cyber,” particularly within the context of our discussed case of the GPS substation clock, should be more clearly delineated. Among a variety of sources, the U.S. Army Cyber Warfare Field Manual (FM) 3-38 [1], “Cyber Electromagnetic Activities” (supplanted by FM 3-12 “Cyberspace and Electronic Warfare”) contends that “Cyber Electromagnetic Activities” encompass not only conventional cyber activities (e.g. distributed denial of service or DDoS attacks¹),

¹ A Distributed Denial-of-Service (DDoS) attack is an attack by which multiple compromised computer systems attack a targeted resource, such as a GPS substation clock. The torrent of incoming messages, connection requests force the targeted resource to slow down or shut down, thereby denying service for legitimate use.

but also activities involving electronic warfare (e.g. GPS jamming, GPS spoofing, etc.) and spectrum management operations.² This is delineated in Fig. 2.

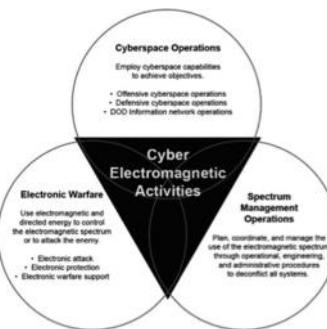


Fig. 2. Cyber electromagnetic vulnerabilities

As can be seen by way of vulnerability databases (e.g. National Vulnerability Database or NVD), such as that produced by the National Cybersecurity and Communications Integration Center (NCCIC) Industrial Control Systems Cyber Emergency Response Team (ICS-CERT), there exists several GPS substation clock vulnerabilities that can affect the accuracy of the clock. For example, an actual ICS-CERT vulnerability (ICSA-14-345-01) characterization is as follows: “an attacker with specialized radio equipment and knowledge could transmit signals that can disrupt the (GPS substation) clock” [2]. In addition to GPS spoofing, other phenomenon, such as GPS blocking/jamming, loss-of-lock (LOL), multipath interference, and equipment failure, can all profoundly disrupt accurate timestamping.

Presuming that these issues are adequately addressed prior to deployment and the involved GPS substation clock is put into actual service, the GPS substation clock’s timestamping paradigm is typically used to synchronize power system DMEs of the modern-day smart grid.³ However, the actual implementation can lead to yet another potential cyber “Achilles heel.” As one example, some electric utility implementations

² Spectrum management refers to the management of the spectrum. By way of example, the U.S. spectrum is managed by the Federal Communications Commission (FCC) for non-governmental applications as well as by the National Telecommunications and Information Administration (NTIA) for governmental applications. Spectrum management is a burgeoning problem due to the growing number of spectrum uses, such as over-the-air broadcasting, government and research uses (e.g. defense, public safety), commercial services to the public (e.g. wireless broadband), and industrial, scientific, as well as medical services.

³ Since the modern-day smart grid is a complex and interconnected system [3], the value of time synchronization should be clear; what happens in one part of the power grid affects operations elsewhere in the involved power grid. When these complex interactions, usually in response to an event, lead to cascading faults and large-scale power outages (i.e. blackouts), it is necessary to make sense of the timestamped data generated by the PMU and other DMEs. The data will have a spatial frame of reference (i.e. what happened where) and a temporal frame of reference (i.e. what happened when).

at substations utilize a single GPS antenna, in conjunction with a GPS antenna cable splitter, to drive multiple devices. Other implementations use a single GPS substation clock (with a single internal GPS module/receiver chip) to provide the multiple outputs to drive multiple devices [4]. In either case, the single GPS antenna/single GPS receiver module/chip paradigm constitutes a potential cascading succession of single points of failure or “Achilles heels” for the various end-devices (e.g. the list of DMEs provided earlier), and the single GPS antenna and/or single GPS receiver module/chip is subject to, among other issues, the experiencing of GPS satellite technical anomalies (e.g. such as the anomaly that affected almost all the GPS receivers used in several substations in Japan [5]), natural interference (e.g. such as the solar flare that disrupted GPS satellites on 6 September 2017 [6]), and equipment problems (e.g. such as the replacements that needed to be effectuated by a particular GPS substation clock manufacturer for 4 of 7 substation installations). The net effect of this one-to-many failure paradigm is that a successful attack at the identified single points of failure or “Achilles heels” (the single GPS antenna and/or single GPS receiver module/chip) could potentially segue to a paradigm of cascading timestamping failure consequences across a key substation of a power grid, potentially disrupt normal operations, compound the complexity of event analysis, and negate any type of robust regression analysis or predictive analytics.

This paper posits an architectural schema that endeavors to mitigate the identified single points of failures (more fully described in Sects. 2 and 3), via a hybridization of several modified algorithms (that are of value-added proposition beyond the current schemas described in Sects. 4 and 5). Collectively, these modified algorithms (within the context of the architectural schema described in Sect. 6) determine the timestamps that will be utilized given the real-time, ongoing circumstances at-hand.

2 Societal Impact of Cyber Breaches and Power Outage Incidents

Cyber breaches have become fairly prevalent in modern society. According to the Federal Trade Commission (FTC), the 2017 Equifax data breach affected 143 million people [7]. The U.S. Office of Personnel Management (OPM) states that its 2015 OPM data breach affected 22.1 million people [8]. Health insurance provider Anthem Inc. states that its 2015 data breach affected 79 million people [9]. Department store retailer Target Corp. states that its 2013 data breach affected 41 million people [10]. Clearly, the number of people affected by data breaches is quite significant.

Of course, there is the well-known first-of-its-kind cyber incident of 17–18 December 2017 that resulted in a power outage affecting 225,000 people in the Ukraine capital city of Kiev; the cyber attackers had also sabotaged some of the involved power distribution equipment, thereby complicating the ensuing attempts to restore power. Preliminary findings indicate that workstations and Supervisory Control and Data Acquisition (SCADA) systems linked to power supplier Ukrenergo’s substation called “North” were compromised and that the cyber incident was “a premeditated and multi-level invasion” [11]. Marina Krotofil, the lead cybersecurity researcher at Honeywell International, Inc., who assisted in the investigation, stated, “It was an intentional cyber incident ... they actually attacked more, but couldn’t achieve all their goals” [12].

In a comparable fashion to that of data breaches, power outages have profoundly impacted modern society. Each year, thousands of airline travelers are affected by power outages affecting airlines and airports.⁴ Historically, cascading power failures, such as the Northeast Blackout of 2003, have affected millions of people. Other notable power outages include that of Super Bowl XLVII, which affected 111 million viewers and 70,000 spectators in attendance [16].⁵ More recently, on 15 August 2017, a massive power blackout affected “half of all households on the island” of Taiwan, which has 7.54 million households [17].⁶ Suffice it to say, the societal impact, via the number of people affected, of data breaches and that of power outage incidents are on par with each other.

3 Vulnerabilities of Power Grids

The GPS issue is clearly of significance, so it should be of no surprise that, among other directives issued by various countries around the world, on 11 May 2017, a U.S. Presidential Executive Order on “Strengthening the Cybersecurity of Federal Networks and Critical Infrastructure” (which includes the power grid) was issued. On 24 May 2018, the U.S. House of Representatives incorporated what is known as the “National Timing Resilience and Security Act of 2018” into the National Defense Authorization Act for Fiscal Year (FY) 2019 (House of Representatives or H.R. 5515), where it appears as Section 4514, “Backup Global Positioning System.”

⁴ A 9-h power outage affected San Diego International Airport (a.k.a. Lindbergh Field) on 25 May 2013. On 10 October 2013, a power outage caused major delays at London-Stansted Airport. On 27 March 2015, a power outage struck the Amsterdam Airport Schiphol. On 8 August 2016, Delta Air Lines suffered an information technology (IT) systems failure due to a power outage and stated that it had lost \$150 million in revenue as a result [13]. On 27 May 2017, a power outage at the British Airways data center caused an IT systems failure that resulted in more than a \$100 million in lost revenue and “the expense of accommodating, re-booking, and compensating thousands of passengers” [14]. More than 50 airports across Japan were affected when the national airline carrier, All Nippon Airways (ANA), experienced an outage of its IT systems on 21 March 2016. On 17 December 2017, Atlanta’s Hartsfield-Jackson International Airport, the world’s busiest airport, experienced a power outage that led to more than a thousand flight cancellations and stranded thousands of passengers in the dark [15].

⁵ On 3 February 2013, one minute and thirty-eight seconds into the third quarter of the second half of Super Bowl XLVII, there was a power outage at the Mercedes-Benz Superdome that interrupted the Super Bowl game for 34 min. With 111 million viewers and another 70,000 in attendance, it may be the most infamous power outage in history, as it earned the game the nickname of “Blackout Bowl.”

⁶ On 15 August 2017, President Tsai Ing-wen of the Republic of China (a.k.a. Taiwan) stated, “Electricity is not just a problem about people’s livelihoods [...] but also a national security issue. A comprehensive review must be carried out to find out how the electric power system can be so easily paralyzed” [18]. She also noted that the 15 August 2017 power failure should not be seen as an “isolated case.” In particular, the incident caused a ripple effect that resulted in a blackout affecting all of the Hsinchu Science Park and the fabrication facilities of semiconductor wafers located there. According to Annabelle Hsu, a senior research manager in Taiwan at International Data Corp., Hsinchu Science Park is the “heart” of Taiwan’s semiconductor industry. Hsu also stated, “Factories there are running 24 hour a day, so sudden power cut without warning could cause a big damage” [19].

Interestingly, these backup and resiliency-related actions occurred 16 years after the first comprehensive assessment—of the vulnerabilities impacting the civilian GPS infrastructure—was conducted. This referenced assessment was published in 2001 by the Volpe National Transportation Systems Center [20]. The assessment concluded that, among the various types of attacks, GPS spoofing is the most malicious and difficult to detect [21]; since PMUs utilize the GPS signals received by a GPS substation clock to derive a timestamp, they are indeed vulnerable to spoofing [22]. In particular, a spoofing attack can cause the GPS receiver of a GPS substation clock to compute an erroneous clock offset, thereby resulting in erroneous timestamp calculations, which in turn, introduces error into the various PMUs' phase angle measurements [23]. This described phenomenon is devastating, as time synchronization across PMUs is crucial to its function of wide-area situational awareness of a power grid [24]; indeed, a substantive portion of applications in powers system rely upon PMU measurements, such as power grid health monitoring algorithms, automatic control systems, remedial action schemes [25], and other applications, as shown in Fig. 3 on the following page.

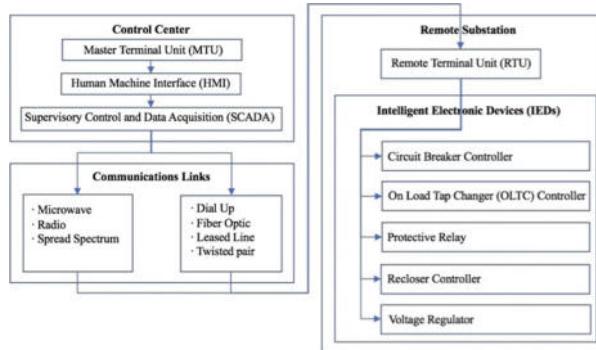


Fig. 3. Time synchronization as a common denominator of the control center, communications links, and remote substation (including the intelligent electronic devices or IEDs) of a power grid

At the Control Center, there is a Master Terminal Unit (MTU) and computer networks. Time synchronization is critical, as debugging the involved computer network involves determining when events happen. Indeed, time synchronization is necessary to accurately correlate log files between these devices so as to determine network usage and examine information technology (IT) security breaches. The same concept also applies to the delineated Communications Links and the Intelligent Electronic Devices (IEDs) at the delineated Remote Substation of Fig. 3.

At the Remote Substation, the RTU transmits telemetry data from the IEDs back to the MTU at the Control Center and, in turn, utilizes messages from the MTU to control the IEDs. The IEDs, which receive telemetry data from electrical power equipment (e.g. circuit breakers), can issue control commands, such as tripping circuit breakers, if anomalies are sensed in the voltage, current, or frequency. This is generally controlled

by a setting file, which contains thresholds. The testing of setting files (and their contained thresholds) is typically one of the most time-consuming roles of a power grid protection engineer.

Indeed, various power grid vulnerabilities have been illuminated at key (remote) substations, at the key substation clocks of key substations, and at Supervisory Control and Data Acquisition (SCADA)/Industrial Control Systems (ICS) systems (that rely upon the GPS substation clocks), among other vulnerabilities. This is detailed, via the following Subsects. 3.1 through 3.4.

3.1 Key Substations

In March 2014, a study released by the Federal Energy Regulatory Commission (FERC) indicated that 30 substations in the U.S. play a significant role in power grid operations; according to this same FERC study, a major blackout could occur, if simply 9 of these substations were taken out of commission [26]. Interestingly, the 16 April 2013 sniper shooting on the Pacific Gas and Electric Company (PG&E) transmission substation in Coyote, California (near the border of San Jose, California) had clearly underscored the physical vulnerability of substations.⁷

This is particularly troubling, for of significant note, in many cases, a single key substation (often a terminal substation) supplies 100% of the load for the involved critical infrastructure (e.g. airport). In other cases, the key substation supplies a substantive portion of the load (rather than 100% of the load); in either case, the criticality of the key substation is not negated.

3.2 Key GPS Substation Clock at the Key Substation

Besides the physical vulnerabilities of substations themselves, there are key components at the substations, such as the GPS receiver within the GPS substation clock, that are particularly vulnerable. Despite the knowledge that a successful attack at the identified potential single point of failure (e.g. a single GPS substation clock driving multiple devices at a key substation) can potentially have cascading timestamping failure consequences across the DMEs serving the key substation and, thereby, profoundly impact the involved power grid supplying airlines and airports, among other venues, the alarming significance of this particular potential critical point vulnerability has neither been extensively articulated nor robustly addressed. Indeed, its cyber electromagnetic effect has not been accentuated as diligently as other cyber effects, such as data breaches, although the societal impacts are on par.

⁷ The sniper incident was referred to as the “Metcalf Sniper Attack,” wherein gunmen fired on 17 electrical transformers and caused \$15 million worth of damage. To avoid a blackout, power was rerouted from nearby Silicon Valley-based power plants. Former Chairman of the FERC Commission, Jon Weillinghoff, described the attack as “the most significant incident of domestic terrorism involving the grid that has ever occurred” [27]. A ranking member of the U.S. House Committee on Energy and Commerce, Henry Waxman, stated that the attack was “an unprecedented and sophisticated attack on an electric grid substation with military-style weapons. The attack inflicted substantial damage, and it took weeks to replace the damaged parts. Under slightly different conditions, there could have been serious power outages or worse” [28].

In addition, the commonalities with regards to the various “means to a breach” (i.e. attack vector) are clear: distributed denial of service (DDoS) attacks, phishing, malware, ransomware, and various other vulnerability exploits. By way of example, the National Institute of Standards and Technology (NIST) National Vulnerability Database (NVD) lists a particular GPS substation clock that “allows remote attackers to cause a denial of service (i.e. disruption)” [29]. The NVD utilizes a Common Vulnerability Scoring System (CVSS), which provides an open framework for communicating the characteristics and impacts of Information Technology (IT) vulnerabilities. The CVSS score for the referenced GPS substation clock exemplar is shown in Table 1.

Table 1. Characteristics of a national vulnerability database (NVD) vulnerability

Characteristics of an NVD vulnerability	Description
CVSS 2.0 base score	<i>High</i> 7.8
Vulnerability type(s)	Denial of service
Availability impact	<i>Complete</i> (there is a total shutdown of the affected resource. The attacker can render the resource completely unavailable)
Access complexity	<i>Low</i> (specialized access conditions or extenuating circumstances do not exist. Very little knowledge or skill is required to exploit)
Authentication	<i>Not required</i> (authentication is not required to exploit the vulnerability)

The NVD’s CVSS Specification Documents provide the severity explanations. In any case, a CVSS Base Score of 7.8 is high, whether it is for the CVSS v2.0 Specification Document or for the CVSS v3.0 Specification Document, as is articulated (bolded and italicized) in Tables 2 and 3.

Table 2. CVSS V2.0 ratings

Severity	Base score range
Low	0.0–3.9
Medium	4.0–6.9
<i>High</i>	<i>7.0–10.0</i>

Table 3. CVSS V3.0 ratings

Severity	Base score range
None	0.0
Low	0.1–3.9
Medium	4.0–6.9
<i>High</i>	<i>7.0–8.9</i>
Critical	9.0–10.0

To exacerbate this issue, attackers can utilize various search engines for the Internet of Things (IOT) or Internet-connected devices, such as SHODAN (<https://www.shodan.io/>), to find various technologies including SCADA/ICS, which rely upon GPS substation clocks. Attackers can perform bulk searching and processing of SHODAN queries, via software called SHODAN Diggity, which provides a list of 167 search queries in a dictionary file, known as the SHODAN Hacking Database (SHDB). The described process, as shown in Fig. 4, is streamlined and enhanced by Search Diggity, which is a graphical user interface (GUI) application developed for the Google Hacking Diggity Project. In essence, the attack surface area of the GPS substation clock may be at greater exposure due to the combinatorial of elements (e.g. Search Diggity, SHODAN, SHODAN Diggity, SHDB, etc.) that may be utilized maliciously by an attacker as accelerators.

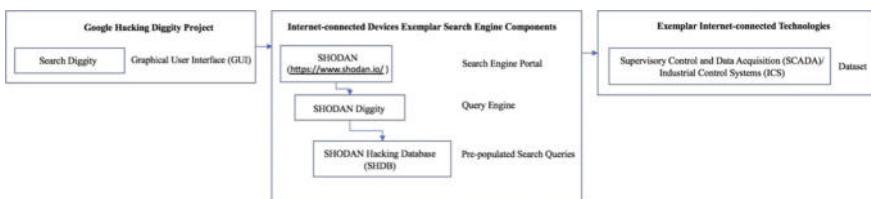


Fig. 4. Interplay among the components of an internet-connected devices search engine and exemplar internet-connected technologies

3.3 Protocol Limitations

The modern power system utilizes several methods for time synchronization: (1) obtaining accurate time synchronization by way of an IRIG-B signal synchronized to GPS time (i.e. GPS satellites), (2) effectuating synchronization by Network Time Protocol (NTP) or Simple Network Time Protocol (SNTP), and (3) improving the accuracy of time synchronization by the IEEE 1588 Precision Time Protocol (PTP). However, each described method has its own inherent weaknesses. The GPS substation clock may experience Loss-of-Lock (LOL), thereby affecting the IRIG-B signal. The uncertainties of Ethernet transmission delay may affect the accuracy of both NTP and SNTP protocols [30]. IEEE 1588 PTP is heavily dependent upon the stability of the oscillator [31] as well as the GPS signal.

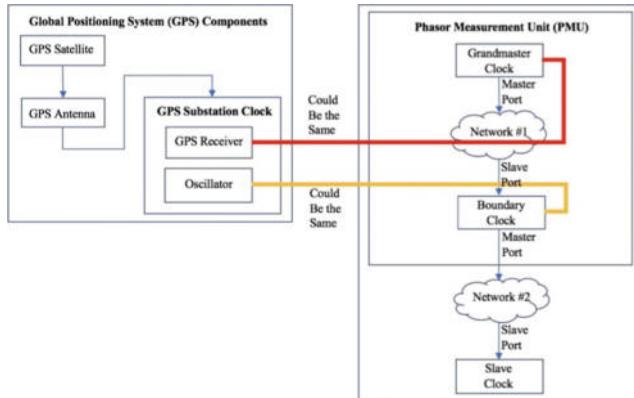
Collectively, these time synchronizations protocols are presented below in Table 4 [32]. In essence, each protocol has an associated potential “Achilles heel.”

Yet, the IEEE 1588 PTP standard as of 2008 (IEEE 1588-2008) remains the most current paradigm. It was envisioned as mitigating the deficiencies of NTP and GPS. A PTP network can be utilized in combination with GPS receivers to serve as a Grand Master Clock (GMC)⁸ (i.e. the time source for that network). As can be seen in Fig. 5,

⁸ A Grand Master Clock is a clock that synchronizes to a GPS receiver. It always runs in PTP Master Mode and distributes its time throughout the Network (e.g. Network #1), via its Master Port.

Table 4. Time synchronization protocols

Protocol	Media	Synchronization accuracy
NTP/SNTP	Ethernet	50–100 ms
IRIG-B	Coaxial	1–10 μ s
PTP	Ethernet	20–100 ns

**Fig. 5.** Exemplar precision time protocol (PTP) network involving global positioning system (GPS) components and a phasor measurement unit (PMU)

in a PTP network: (1) a GMC acts as a time source for the network; (2) a Boundary Clock⁹ synchronizes itself to this GMC; and (3) the Boundary Clock synchronizes another part of the network, via a Slave Clock.¹⁰

Among others, a problematic issue that arises is the fact that IEEE 1588-2008 PTP is predicated upon a master/slave principle,¹¹ which has a potential critical point failure [33]. The failure of a Master Clock necessitates the re-election of a new Master Clock (the PTP standard utilizes an algorithm called the ‘Best Master Clock Algorithm’ or BMCA to select, from among the various clocks, the one to serve as the Master Clock). The process of electing and selecting a new Master Clock, which is then designated as the GMC, requires a certain amount of time during which the various “Slave Clocks” are not synchronized; rather, they are running disparately and singularly. Unfortunately, the paradigm of various “Slave Clocks” running disparately and singularly is

⁹ A Boundary Clock is a clock that synchronizes to a PTP Master, via its Slave Port, and distributes the time to another part of the Network (e.g. Network #2), via its Master Port.

¹⁰ A Boundary Clock is a clock that synchronizes to a PTP Master, via its Slave Port.

¹¹ The master/slave principle, in the context of time synchronization, refers to a paradigm, wherein a master clock is an electronic device that provides time synchronization signals to a number of slave clocks on a network; typically, the master clock derives its time from a time source, such as a GPS signal from a satellite.

unacceptable for anomaly detection, as correct correlation of data to time (i.e. accurate timestamping) is needed to establish the baseline against which anomalies can be detected.

3.4 Data Graph of Phasor Measurement Unit Time (Based upon GPS Substation Clock) at Key Substation

The temporal scale of Phasor Measurement Unit (PMU) time is in the millisecond (ms) range. In theory, when graphed against actual time (International Atomic Time or TAI¹²), the graph will yield a diagonal line, as shown in Fig. 6.

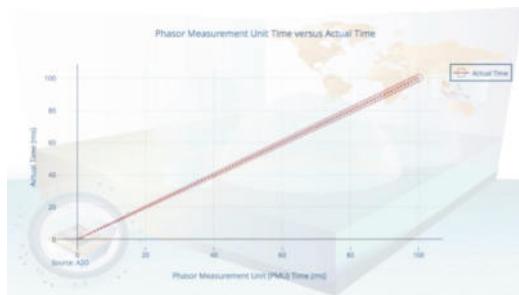


Fig. 6. Phasor measurement unit (PMU) time versus actual time

In actuality, PMU time is subject to a variety of factors: (1) the electromagnetic interference experienced by the GPS signal, (2) GPS signals in the atmosphere experiencing a delay or latency, and (3) jitter within the 1PPS output generated by GPS receivers, and other issues. Accordingly, when PMU time is graphed against actual time, the graph will yield an approximate diagonal line, as shown in Fig. 7.

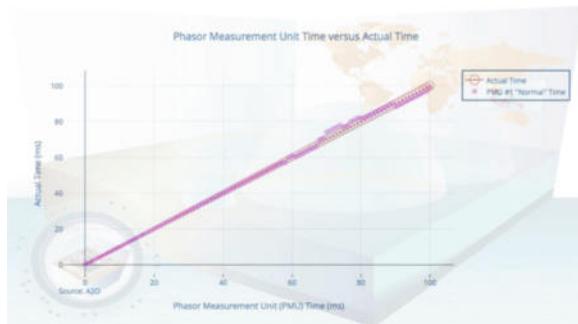


Fig. 7. Phasor measurement unit (PMU) #1 “Normal Time”

¹² International Atomic Time (TAI) is one of the mainstays of Coordinated Universal Time (UTC), which is the time scale used to determine local times around the world.

Given the “baseline” diagonal line exhibited in Figs. 7 and 8, “anomalous” time (the exhibited non-diagonal line) can be graphed as a layer atop PMU time graphed against actual time (an approximate diagonal line), as shown in Fig. 8.

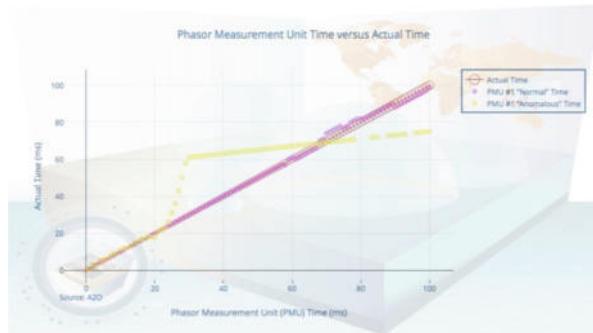


Fig. 8. Phasor measurement unit (PMU) #1 “Anomalous Time”

4 Current Mitigation Strategies to Address the Critical Point Failure

The issue of “accurate time-stamping” becomes more significant as critical infrastructure increasingly relies upon services currently provided by Global Positioning System (GPS). Despite the increasing cycles of adaptation by cyber attackers and the evolution of the cyber threat landscape, the defenses to the GPS critical point failure remains limited. Several mitigation strategies have been employed to address the known critical point failure of the GPS substation clock. First, for some time, it was hoped that enhanced LOnG RAnge Navigation (LORAN) (a.k.a. eLORAN or E-LORAN, formerly Loomis Radio Navigation or LRN) could serve as an alternative. Second, when it was found that multiple receivers were useful for the detection of spoofing [34], the paradigm of multiple GPS substation clocks was utilized as a mitigation strategy. Third, an IRIG-B decoder has proved to be a useful mitigating tool, as it can scrutinize the input signal for data integrity¹³ and data validity¹⁴ as well as assist in determining whether a problem resides with the GPS substation clock, GPS receiver, or end-device (e.g. DMEs). This is detailed, via the following Subsects. 4.1 through 4.3, below.

¹³ Data Integrity refers to the assurance that data is unchanged from its source and has not been accidentally (e.g. through programming errors) or maliciously (e.g. through IT security breaches) altered or destroyed.

¹⁴ Data Validation refers to the evaluations utilized to determine compliance with specifications and requirements so as to ensure correctness and reasonableness of the data.

4.1 Enhanced Long Range Navigation (eLoran)

For scenarios where GPS is unavailable or degraded, enhanced LORAN (eLoran) seeks not only to build upon LORAN-C¹⁵—the most recent version of LORAN, which operates in the 3 kHz–300 GHz radio frequency (rf) portion of the electromagnetic spectrum—but also to mitigate against the known vulnerabilities of GPS. As such, eLORAN includes additional pulses, which can transmit auxiliary data such as Differential Global Positioning System (DGPS)¹⁶ corrections. The eLORAN receivers also use “all in view” reception, incorporating signals from all stations in range, not solely those from a single Group Repetition Interval (GRI),¹⁷ thereby incorporating time signals and data from up to 40 stations [37].

4.2 Multiple Global Positioning System (GPS) Antennas and Multiple GPS Receivers

As previously discussed, GPS receivers indicate Lock/Sync status; of course, GPS receivers also indicate Loss-of-Lock (LOL). In addition to Lock/Sync status, most GPS receivers indicate the signal level that has been acquired from each of the satellites. For optimal performance, this should be checked against the manufacturer’s recommended signal strength level for the GPS receiver.

Oftentimes, multiple GPS receivers come in the form of multiple GPS substation clocks. In Sect. 2, the massive power outage that occurred in Taiwan is referenced. It seems apropos to cite an ongoing effort in Taiwan to enhance the resiliency of GPS-based timestamping; one such effort involves a multiple time source compare system (MTCS) to enhance the reliability of the time sources for time dissemination services [38]. In terms of system architecture, the MTCS utilizes a 3-source system: (1) Coordinated Universal Time (UTC) IRIG-B time code generated from a cesium-beam

¹⁵ LORAN-C was a ground-based navigation system operated by the U.S. Coast Guard. In May 2009, President Obama declared the system obsolete. That same year, Congress debated whether to retain and upgrade the LORAN-C infrastructure to become eLORAN, which was envisioned as a national backup to GPS. This utilization as a national backup has not yet occurred. In October 2009, Congress decided to terminate LORAN-C term. The Coast Guard began shutting it down in February 2010. In accordance with the 2010 U.S. Department of Homeland Security (DHS) Appropriations Act, the U.S. Coast Guard terminated the transmission of all U.S. LORAN-C signals on 8 Feb 2010. In February 2014, the U.S. House of Representatives Transportation and Infrastructure Committee reopened the topic [35]. However, the utilization of eLoran as a national backup has not yet occurred.

¹⁶ A Differential Global Positioning System (DGPS) refers to enhancements to GPS, via improved location accuracy, from the 15-m nominal GPS accuracy to better than 10 cm, for the optimal implementation cases [36].

¹⁷ Group Repetition Interval (GRI) refers to the time interval between the reoccurrence of a Master pulse, which is the pulse from the first station (a.k.a. “master”), which then triggers the second station (a.k.a. “slave”) into transmitting a similar pulse after a pre-determined time delay.

atomic clock¹⁸ and time-code generator, (2) Taipei, Taiwan local time (UTC + 8 h) IRIG-B time generated from another cesium-beam clock and time-code generator, and (3) GPS time generated from a True Time GPS receiver. All 3 codes are decoded and fed into a “Compare Unit.” The compared results are then sent into a “Control Unit” to determine whether the two switches, as shown in Fig. 10, should be on or off. Unfortunately, the posited “Compare Units” and “Control Units” all use single number thresholds, such as “10 μ s” for the presented source. Hence, the various posited solutions presume certain thresholds, but the variances are, in actuality, far greater by a factor of 3, such as “30 μ s” [40]. Other sources cite a range of PMU measurement error variances [41]. Collectively, in summation, a single number threshold is considered to be potentially brittle and is of dubious value. The aforementioned described architecture is articulated in Fig. 9.

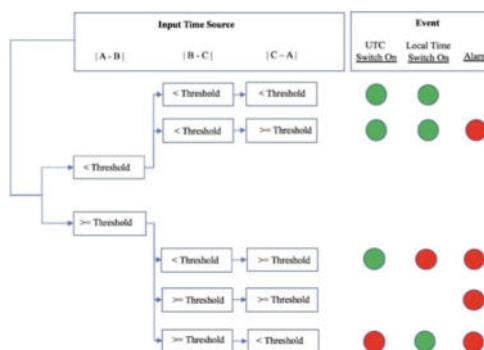


Fig. 9. Multiple global positioning system (GPS) substation clock “Compare Unit” and “Control Unit” interplay

4.3 Inter-Range Instrumentation Group (IRIG)-B Decoders

Engineers from various utilities assert that the monitoring and logging of IRIG-B timing output from the GPS substation clock can be quite insightful, as the substantive portion of the problematic issues experienced have been traced to the GPS substation clock firmware, not the GPS receiver firmware or end-device (e.g. PMU, DFR, etc.)

¹⁸ Atomic clocks are responsible for synchronizing time for much of the technology in modern society, including electric power grids. On 3 April 2014, NIST officially launched their new standard for time using the NIST-F2 atomic clock. According to NIST’s time and frequency division, NIST-F2 is accurate to one second in 300 million years [39]. The NIST-F2 was certified by the International Bureau of Weights and Measures as the world’s most accurate time standard. Both NIST-F2 and the NIST-F1 (the standard that NIST F-2 replaced) are known as cesium-based atomic clocks or atomic fountain clocks (an atomic fountain refers to a cloud of atoms that is tossed upwards in the Earth’s gravitational field by lasers, and if the cloud of atoms were visible, it would resemble the water in a fountain; albeit weightless in the toss, the atoms are measured to set the frequency of an atomic clock), which means that they determine the length of a second by measuring the natural vibration inside a cesium atom.

firmware; they further assert that without monitoring the IRIG-B output of clock, it cannot be ascertained whether the problem resides with the clock, receiver, or the end-device [42].

Utilizing an IRIG-B decoder seems prudent. The core function of an IRIG-B decoder is to read and check time codes, which are in the IRIG-B format, for integrity and validity and to decode and display the Binary Coded Decimal (BCD) information contained within. The IRIG-B decoder verifies the integrity and validity of the input signal and, if the result is positive, it derives from the IRIG-B code a pulse per second (PPS or 1PPS) synchronized with the markers of the time code. Furthermore, the unmodulated input IRIG-B code is replicated at the output. The time information can then also be made available over a serial digital interface [43]. These signals allow synchronization with the time signal generated by the GPS satellite(s). There are a variety of commercial IRIG-B analyzers available (e.g. Tektron, Verilog, etc.). There are also various do-it-yourself (DIY) versions.

To conclude Subsects. 4.1 through 4.3, even with the various lines of effort towards mitigating the GPS critical point failure, the current mitigation strategies—even as an amalgam—do not suffice to robustly address the remaining potential critical point failures that exist with regards to the plethora of timestamping vulnerabilities.

5 Architectural Requirements for a Robust Power Grid

Incorrect timestamping segues into an inability to robustly perform the basic power grid functions of: (1) Monitoring and Diagnostics, (2) Operation and Control, and (3) Planning. According to research, such as that funded by the U.S. Department of Energy DE-AR0000340 [44], the function of Monitoring and Diagnostics resides within the 10^{-2} – 10^5 Hz range. According to the studies conducted as part of the referenced research [45], System Identification resides within the 10^{-2} – 10^2 Hz range as does State Estimation. Event Detection and Identification resides within the 10^{-1} – 10^5 range. The function of Operation and Control resides within the 10^{-4} – 10^2 range. The function of Planning resides within the 10^{-9} – 10^{-4} range. In turn, System Health Monitoring resides within the 10^{-9} – 10^{-4} range. These are presented in Fig. 10.

As pertains to Monitoring and Diagnostics, lower sampling rates might reside between the 10^2 – 10^3 range, and higher sampling rates might reside in the 10^4 – 10^5 range. Taking the example of a cycle frequency of 50 Hz (50 cycles/s), 1 cycle will have the time period of 1/50 s. Hence, 50 cycles/s at 512 samples/cycle means that there are 25,600 samples/s (i.e. between the 10^4 and 10^5 range). At 60 Hz (60 cycles/s), 1 cycle will have the time period of 1/60 s. Hence 60 cycles/s at 512 samples/cycle means that there are 30,720 samples/s (i.e. between the 10^4 and 10^5 range).

Clearly, at the higher sampling rates, timestamping becomes even more important as the granularity increases. This requisite timestamping paradigm is examined against the current mitigation strategies to address the discussed GPS critical point failure: (1) eLoran, which is not yet operational as a backup; (2) IRIG-B decoders, which unfortunately introduce latency into the equation, as they need time to read and check time codes for integrity and validity; and (3) multiple GPS antennas/multiple GPS



Fig. 10. Power grid functions and sampling rates against frequency *Source* Based upon [45]

receivers, which further necessitates diversification (i.e. varied manufacturers) to avoid the negation of the logic of multiple units, particularly if they are all subject to the same vulnerabilities. Accordingly, a paradigm of utilizing varied GPS substation clock manufacturers (with varied oscillators) along with varied PMU manufacturers (equipped with varied GPS receivers, but with varied external oscillators), whose output is verified by low-latency IRIG-B decoders, was explored.

6 New Proposed Architecture and Mitigation Strategy to Address the Critical Point Failures

In the operationalization of the multiple GPS substation clocks paradigm, a specific methodology was utilized. First, clocks from various manufacturers (anonymized, pursuant to the FICC review process guidelines) were incorporated into the experiment to honor the principle of species diversification. This is shown in Fig. 11. Of significance, various comparisons (utilizing the acronyms/designators of Fig. 11) of Ant 1 to Ant 2 to Ant 3, Rec 1 to Rec 2 to Rec 3, and Osc 1 to Osc 2 to Osc 3 were performed to ascertain true species diversification among Mfr 1, Mfr 2, and Mfr 3.

Second, Phasor Measurement Units (PMUs) that have their own built-in GPS receivers (e.g. Energoservice), of varied manufacturers, were incorporated into the experiment. Third, regarding the oscillation service for these PMUs, external oscillators (e.g. Microsemi) Osc 4, Osc 5, and Osc 6 were incorporated into the experiment, as shown in Fig. 12.

When viewed from the vantage point of applying the Best Master Clock Algorithm (BMCA), Figs. 11 and 12 can be re-labeled as shown in Fig. 13.

The BMCA is then applied to the various groupings. For this particular example, the groupings are: (1) Rec 1, Osc 1, Rec 4, Osc 4; (2) Rec 2, Osc 2, Rec 5, Osc 5; and (3) Rec 3, Osc 3, Rec 6, Osc 6. This is shown in Fig. 14.

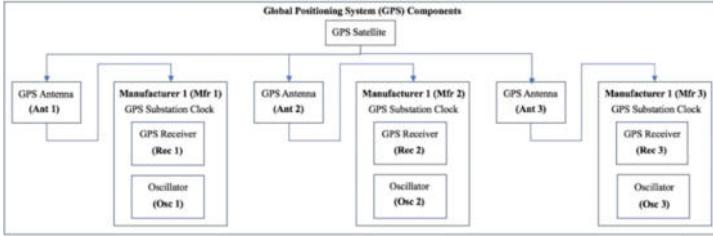


Fig. 11. Determination of species diversification for varied GPS substation clock selection

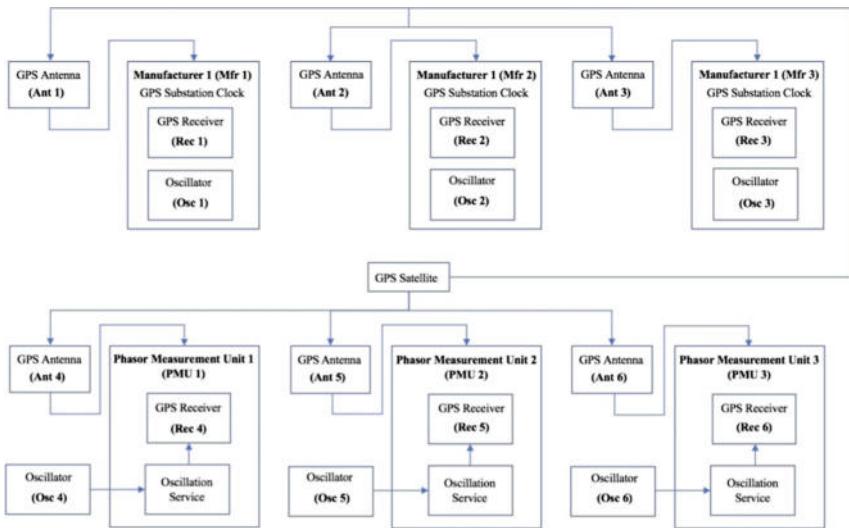


Fig. 12. Determination of species diversification for varied PMU selection

In essence, the classical BMCA operates [46] as shown in Fig. 15. All clocks which are not the “better master clock” and not elected as the “best clock” will go passive and not send messages known as *Announce Messages*. The elections are based upon various heuristical priorities. For example, if MC 1 has a better oscillator than MC 2, which is otherwise identical, MC 1 will be the elected as the “better master clock” and be designated the Grand Master Clock (GMC). However, if MC 1 has a Loss-of-Lock, MC 2 will be elected as GMC. When MC 1 re-obtains a Lock to the GPS signal, MC1 will be re-elected as the GMC. Clocks discover the properties other clocks by the fields in each clock’s *Announce Message*. Clocks send an *Announce Message*, if they do not receive an *Announce Message* from a clock with better properties. IEEE 1588-2008 calls for a clock to populate the parent data set with their own attributes on initialization, and then replace them later with the MC that it synchronizes to.

For our proposed architecture, atop the classical BMCA schema [47], we utilize a modified N-Input Voting Algorithm (NIVA) to determine reasonableness, via a comparator logic range (as contrasted to the single number threshold of Fig. 9) against a

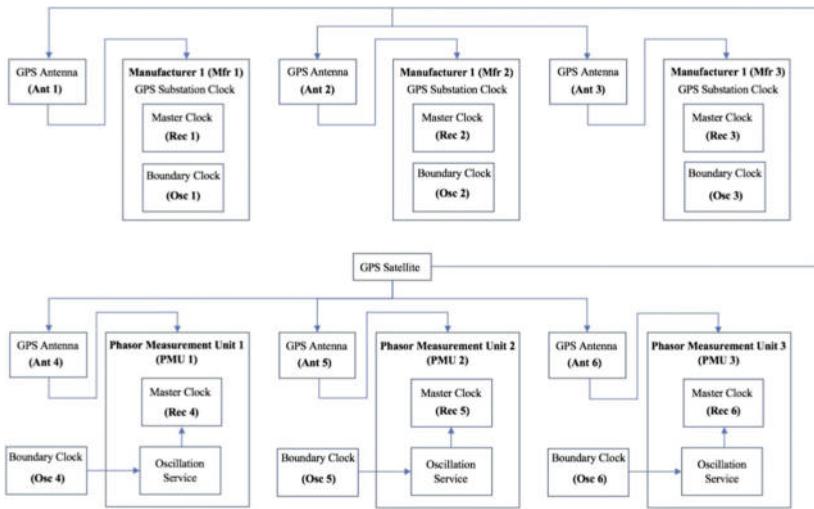


Fig. 13. Interplay of master clocks (Rec 1 through Rec 6) and boundary clocks (Osc 1 through Osc 6)

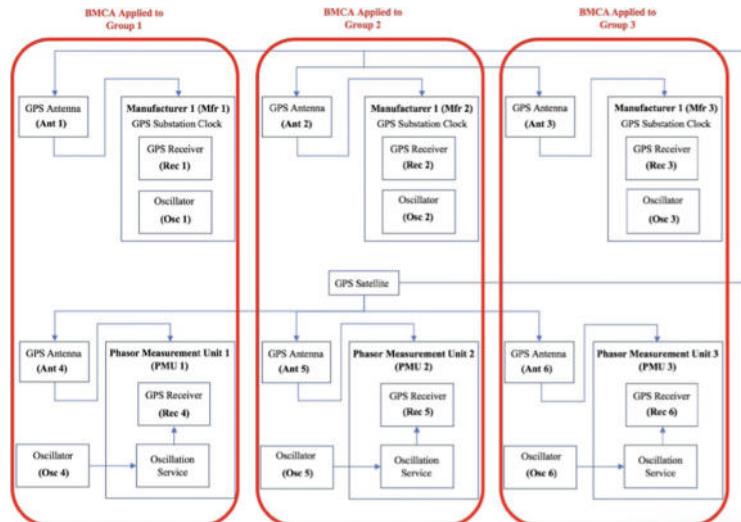


Fig. 14. Best master clock algorithm (BMCA) applied to groups 1 through 3

baseline-determined accuracy threshold range. It is widely accepted that “voting is an important operation in [a] multichannel computational paradigm ... [for the] ... realization of ultrareliable and real-time control systems that arbitrate{s} among the results of N redundant variants” [48]. If all seems reasonable, then the designation is formalized, and the “Best Master Clock” becomes designated as the “Newly Selected GMC.” A modified Fault Tolerant Average Algorithm (FTAA) is then applied, as shown in

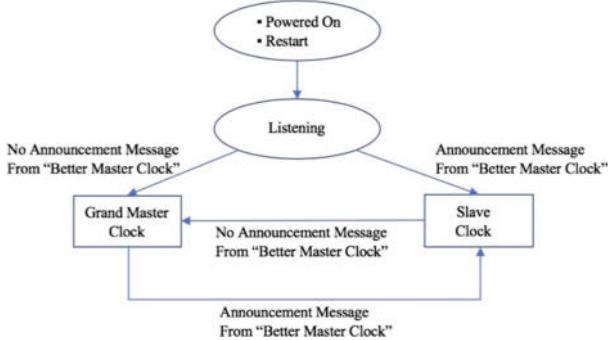


Fig. 15. The logic of the best master clock algorithm (BMCA)

Fig. 16, against the results of the various NIVAs, which were applied to the Best Master Clock Algorithms (BMCA) that were applied to their respective groups. In essence, the FTAA will determine the mean of the results (which encompass the three “best master clocks” selected by the BMCA, as applied to Group 1, Group 2, and Group 3) “excluding t fastest and t slowest” [49]. Once the “New Clock Value” is determined from the FTAA, as applied to the NIVA-generated values, the base reference point is established from which the nearest NIVA-generated value can be determined. The Group associated with that NIVA-generated value becomes the Master Clock Group (MCG) and the MC from within is elected and designated the GMC.

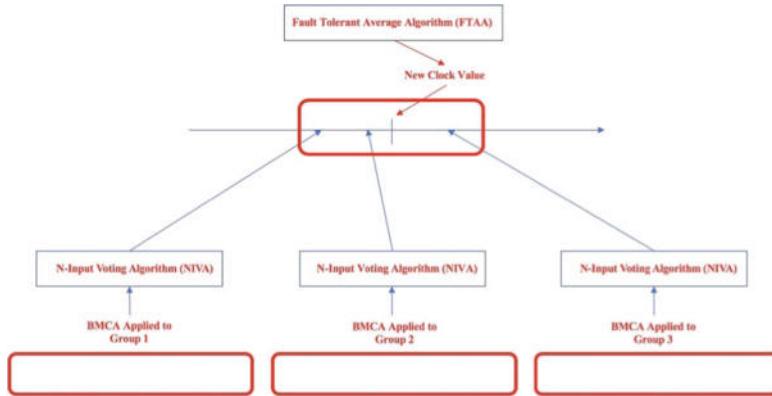


Fig. 16. The logic of the fault tolerant average algorithm (FTAA)

After several security related problems were discovered with PTP upon release of the standard as IEEE 1588-2002 [50], Annex K of IEEE 1588-2008 (a.k.a. Version 2) called for an integrity protection mechanism [51]. An integrity protection mechanism utilizes a Message Authentication Code (MAC) to verify that the message has not suffered any unauthorized modification in transit. In a comparable fashion, under our

proposed architecture, a sync integrity protection mechanism (SIPM) (i.e. SIPM 1) is implemented prior to the final confirmation of the GMC election. This is shown in Fig. 17.

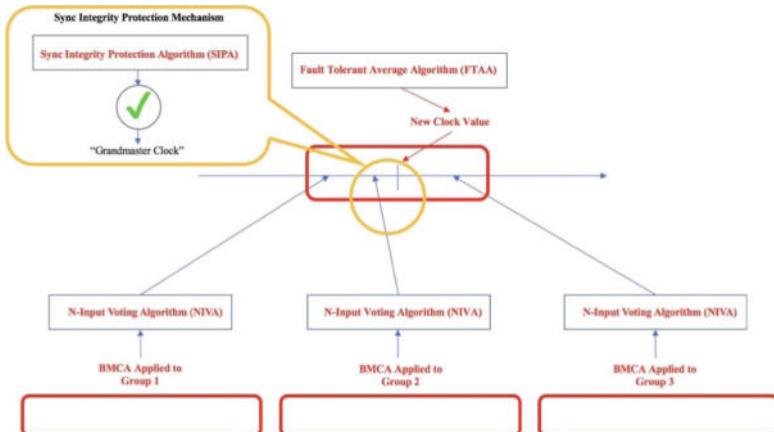


Fig. 17. Utilization of a sync integrity protection algorithm (SIPA) prior to formal "Grand Master Clock" designation

Another SIPM (i.e. SIPM 2) (equipped with an IRIG-B decoder) is utilized before any time information is made available or syndicated (e.g. over a serial digital interface) to the rest of the network. Success was gauged by whether the FTAA values were consistently in closer proximity to the SIPA validated GMC NIVA values than the other NIVA values.

7 Conclusion

In review, timing is used to correlate events recorded at different locations [4]. However, timing or time synchronization is complex. Utilities have noted that "substation timing has been taken for granted—it hasn't been as carefully scrutinized and tested as other equipment" [4]. A principal issue is that "manufacturers of GPS receivers are adding more configuration and connectivity 'features' that normally have security as an afterthought" [40]. Yet, in many cases, manufacturers have designed PMUs, such that the GPS Receiver and Grand Master Clock (GMC) are the same (the Oscillator and Boundary Clock could also be the same); in other cases, the GMC synchronizes to the GPS receiver. In either case, the GPS receiver becomes a potential critical point of failure.

Among other issues, one particular issue is that IEEE 1588-2008 PTP is predicated on a master/slave principle, which has an inherent potential critical point failure [31]: the master clock that is serving as the GMC. The failure of an elected master clock (that has been designated as the GMC) necessitates the re-election of a new master clock to

serve as the GMC; however, the re-election and ensuing designation process requires a certain amount of time during which the various “Slave Clocks” are not synchronized; rather, they are running disparately and singularly. This period of uncertainty is sub-optimal for Disturbance Monitoring Equipment (DME). For example, the fact that PMUs are either equipped with a GPS receiver or rely upon a GPS receiver means that PMUs also inherit “all the insecurity and problems related to GPS, which in turn makes the implementation of a secure time synchronization algorithm more difficult” [36]. Furthermore, the re-election of a new master clock to serve as the GMC still results in the reliance upon only one master clock, which is not validated. In this paper, an architectural schema is proposed, wherein a modified Best Master Clock Algorithm (BMCA) (equipped with a modified “Compare Unit”) is executed along different pathways and then harmonized, via a modification of an N-Input Voting Algorithm (NIVA). A modified Fault Tolerant Average Algorithm (FTAA) is then applied against the results of the various NIVAs so as to determine a Master Clock Group (MCG). Variants of sync integrity protection mechanisms (SIPMs) were utilized prior to the final confirmation of the Grand Master Clock (GMC) election and prior to any time information being utilized and/or syndicated (e.g. via serial digital interface) for data synchronization and/or event correlation purposes.

The described work has been benchmarked against various permutations involving BMCA, NIVA, and FTAA. The MCG and SIPMs are confirmation and validation steps, respectively, at the tail-end of the work flow process and were not a party to the permutations. The preliminary results of the modified BMCA-NIVA-FTAA sequence seem promising. An extensive review of the prior work related to fast recovery mechanisms for BMCA, NIVAs for fault-tolerant systems, and efficient FTAAAs based upon an assortment of techniques had been conducted. Future work will involve a review of updated techniques for benchmarking purposes as well as the potential involvement of other useful algorithmic modifications, via an orchestration layer to be incorporated into the posited architectural schema.

References

1. U.S. Army Cyber Warfare Field Manual (FM) 3-38. Department of the Army, Feb 2014
2. Advisory (ICSA-14-345-01). U.S. Department of Homeland Security’s Industrial Control Systems Cyber Emergency Response Team (ICS-CERT), Jan 2015
3. Nasiruzzaman, A., Pota, H.: Transient stability assessment of smart power system using complex networks framework. In: IEEE Power and Energy Society General Meeting, July 2011
4. Hawks, P., Orndorff, R., Thomas, K.: GPS timing in substations at dominion energy. In: Proceeding of CIGRE USNC Grid of the Future, Oct 2017
5. Itagaki, D., Ohashi, K., Shuto, I., Ito, H.: Field experience and assessment of GPS signal receiving and distribution system for synchronizing power system protection, control and monitoring. In: Proceedings 2006 IEEE Power India Conference, 10–12 Apr 2006
6. Fazekas, A.: How the Strongest Solar Flare in a Decade Is Affecting Earth. National Geographic, Sept 2017
7. McCrank, J., Finkle, J.: Equifax Breach Could Be Most Costly in Corporate History. Reuters, Mar 2018

8. Nakashima, E.: Hacks of OPM Data Compromised 22.1 Million People, Federal Authorities Say. *The Washington Post*, July 2015
9. Pierson, B.: Anthem to Pay Record \$115 Million To Settle U.S. Lawsuits Over Data Breach. *Reuters*, June 2017
10. Hurtado, P.: Target Agrees to pay \$18.5 Million to End Data-Breach Probes. *Bloomberg*, May 2017
11. Minchev, Z., Bogdanoski, M.: Countering Terrorist Activities in Cyberspace. IOS Press and NATO Emerging Security Challenges Division (2018)
12. Poliyuk, P., Vukmanovic, O., Jewkes, S.: Ukraine's Power Outage was a Cyber Attack: Ukrenergo. *Reuters*, Jan 2017
13. Carey, S.: Delta Airlines CEO Takes Responsibility for Outage. *Wall Street Journal*, Aug 2016
14. Tadeo, M., Jasper, C.: British Airways Owner Says Power Outage Cost 80 Million Pounds. *Bloomberg*, June 2017
15. Hedgpeth, D.: Atlanta Crews Restore Power to World's Busiest Airport; Travelers Still Stranded, Flights Canceled. *The Washington Post*, Dec 2017
16. Super Bowl XLVII Draws 108.7 Million Viewers, 26.1 Million Tweets. *Nielsen*, Feb 2013
17. Yu, J., Kao, J.: Taiwan Probes Massive Power Cut That Affects Millions of Households. *Reuters*, Aug 2017
18. Chow, J.: Taiwan To Review Electrical Grid After Flawed Power Supply Replacement Process Caused Massive Blackout. *The Straits Times*, Aug 2017
19. Mishap Triggers Taiwan Blackout as Power Policies Draw Scrutiny. *Bloomberg*, Aug 2017
20. Vulnerability Assessment of the Transportation Infrastructure Relying on the Global Position System. John A. Volpe National Transportation Systems Center, Aug 2001
21. Jiang, X., Zhang, J., Harding, B., Makela, J., Dominguez-Garcia, A.: Spoofing GPS receiver clock offset of phasor measurement units. *IEEE Trans. Power Syst.* **28**(3), 3253–3262 (2013)
22. Fan, X., Du, L., Duan, D.: Synchrophasor data correction under GPS spoofing attack: a state estimation based approach. *IEEE Trans. Smart Grid* (Feb 2017) ISSN 1949-3053 and <https://doi.org/10.1109/TSG.2017.2662688>
23. Silverstein, A.: Electric Power Systems and GPS. Civil GPS Service Interface Committee, North American Synchrophasor Initiative, Sept 2016
24. Gharavi, H., Hu, B.: Synchrophasor sensor networks for grid communication and protection. *Proc. IEEE* **105**(7), 1408–1428 (2017)
25. Ferrer, H., Schweitzer III, E.: Modern Solutions for Protection, Control, and Monitoring of Electric Power Systems. Schweitzer Engineering Labs, Pullman, WA, USA (2010)
26. Smith, R.: U.S. Risks National Blackout From Small-Scale Attack. *The Wall Street Journal*, Mar 2014
27. Smith, R.: Assault on California Power Station Raises Alarm on Potential for Terrorism. *The Wall Street Journal*, Feb 2014
28. Harris, S.: 'Military-Style' Raid on California Power Station Spooks U.S. Foreign Policy, Dec 2013
29. CVE-2014-9194 Detail. U.S. Department of Commerce National Institute of Standards and Technology's (NIST) National Vulnerability Database (NVD), Jan 2015
30. Smith, T., Crites, C.: Case studies in facility-wide time synchronization. In: *Proceedings of the 71st Annual Conference for Protective Relay Engineers (CPRE)*, pp. 1–4, 26–29 Mar 2018
31. Wang, H., Zhu, G., Hou, M., Wang, S.: Time synchronization based on multiplexing RPR channel and IRIG-B time code. In: *5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies*, pp. 869–872, 26–29 Nov 2015
32. Matson, J.: Choosing the Correct Time Synchronization Protocol and Incorporating the 1756-TIME Module Into Your Application. Rockwell Automation, May 2013

33. Gaderer, G., Rinaldi, S., Kero, N.: Master failures in the precision time protocol. In: 2008 IEEE International Symposium on Precision Clock Synchronization for Measurement, Control and Communication, Sept 2008
34. Borio, D., Gioia, C.: A dual-antenna spoofing detection system using GNSS commercial receivers. In: Proceedings of the 28th International Technical Meeting of the Satellite Division of the Institute of Navigation, pp. 325–330, Sept 2015
35. LORAN-C Infrastructure & E-LORAN. National Coordination Office for Space-Based Positioning, Navigation, and Timing's GPS.gov, June 2018
36. High Accuracy-Nationwide Differential Global Positioning System Program Fact Sheet. U.S. Department of Transportation's Federal Highway Administration (2003)
37. Safar, J., Williams, P., Grant, A., Vejrazka, F.: Analysis, Modeling, and Mitigation of Cross-Rate Interference in eLoran. *J. Inst. Navig.* **63**(3), 295 (2016)
38. Lin, C., Lin, S., Liao, C.: Development of a multiple time source comparison system for disseminative services in Taiwan. In: Proceedings of the 31st Annual Precise Time and Time Interval (PTTI) Meeting, pp. 317–322, Dec 1999
39. NIST Launches a New U.S. Time Standard: NIST-F2 Atomic Clock. U.S. Department of Commerce's National Institute of Standards and Technology (NIST), Apr 2014
40. Onica, R., Neves, N., Casimiro, A.: Fault-Tolerant Precision Time Protocol Smart Grids. Laboratório de Sistemas de Grande Escala (LaSIGE)/Departamento de Informática Faculdade de Ciencias, Universidade de Lisboa (2015)
41. Du, J., Ma, S., Wu, Y., Poor, H.: Distributed hybrid power state estimation under PMU sampling phase errors. *IEEE Trans. Signal Process.* **62**(16), 4052–4063 (2014)
42. Hawks, P., Orndorff, R., Thomas, K.: GPS Timing in Substations at Dominion Energy. [International Council on Large Electric Systems] CIGRE [US National Committee] USNC Grid of the Future, Oct 2017
43. Gonzalez, D., Carrillo, J., Espitia, J.: Using an SEL Satellite-Synchronized Clock to Synchronize Devices Via Serial Port Time Broadcasting. SEL, Mar 2014
44. Meier, A., Culler, D., Poola, K., Andersen, M., Brady, K., Arnold, D., Stewart, E., McEachern, A., Moffat, K.: Micro-Phasors for Distribution Systems. Berkeley Electrical Engineering and Computer Sciences (2014)
45. Ardakanian, O., Yuan, Y., Dobbe, R., Meier, A., Low, S., Tomlin, C.: Event detection and localization in distribution grids with phasor measurement units. In: IEEE Power & Energy Society General Meeting, July 2017
46. Arnold, D.: What Makes a Master the Best. Meinberg, Nov 2013
47. Data Center Fabric with Nanosecond Accuracy—Use IEEE1588 PTP on Nexus 3000 Switches. Cisco, June 2012
48. Karimi, A., Zarafshan, F., Ramli, A.: A Novel N-Input Voting Algorithm for X-by-Wire Fault-Tolerant Systems. *The Scientific World Journal* (2014)
49. Lonn, H.: Clock synchronization. In: Parallel and Distributed Systems: Clock Synchronization, p. 11 (2015)
50. Tsang, J., Beznosov, K.: A Security Analysis of the Precise Time Protocol. *Information and Communications Security* (2006)
51. IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems. IEEE Std 1588-2008 (Revision of IEEE Std 1588-2002), pp. c1-269, July 2008



On the Relation Between Security Models for HB-like Symmetric Key Authentication Protocols

Miaomiao Zhang^(✉)

Manhattan College, Riverdale, New York, NY 10471, USA
mzhang01@manhattan.edu

Abstract. The purpose of this paper is to provide a basis and comprehensive view for evaluating security models in the context of HB-like symmetric key authentication protocol development. We consider the man-in-the-middle security in the concurrent setting and proposed a new security notion, c2MIM, which we believe is simple and nature. A number of existing security models including our c2MIM are summarized and compared. Some general considerations for designing and using security models are presented.

Keywords: Man-in-the-Middle · Security models · Authentication · HB

1 Introduction

Embedded or handheld devices are getting more and more prevalent in network communications. These platforms often have strict constraints in terms of computational resources. As a result, standard public-key cryptographic algorithms, which are critical for ensuring the authenticity and integrity of communications, can be too slow or too energy-consuming to be suitable for these types of devices. Designing lightweight cryptographic solutions for resource-constrained devices is therefore an important practical and major theoretical research challenge.

Secret-key authentication is a two party protocol, in which one party (the prover) proves its identity to another party (the verifier) by demonstrating knowledge of the shared key. Theoretically speaking, constructions of such protocols (with strong security, to be discussed below) exist from any one-way function. Practical two-round (challenge-response) protocols can be built from any message-authentication code (MAC), and efficiently instantiated from secure cryptographic primitives, e.g. AES or SHA-256. However, constraints on power consumption and circuit size for low-cost IoT devices make it problematic to deploy conventional cryptographic algorithms like AES or modular exponentiation.

The Learning Parity with Noise (LPN) problem involves a binary secret vector \mathbf{x} of length n , and an adversary who is given a number of LPN samples. Each

sample has the form $(\mathbf{a}, \langle \mathbf{a}, \mathbf{x} \rangle \oplus e)$, where \mathbf{a} is a uniformly random vector, e is a bit that is 1 with probability ε . The adversary's goal is to compute \mathbf{x} or distinguish the samples from completely random (in the decision version of the problem).

The LPN problem was introduced by Angluin and Laird [1]. It soon became notorious for having no efficient noise-tolerant algorithm. It was proven by Kearns [26] that the class of noisy parity concepts is not learnable in the statistical query model; and all known efficient learning algorithms for noisy concepts can be cast in this model. The initial cryptographic applications of LPN was shown by Blum et al. [7].

Because of the computational simplicity of binary arithmetic and the strong security guarantee due to the failure to find polynomial-time algorithms for it, LPN has been intensively studied in building efficient cryptosystem.

The work on LPN-based authentication protocols began with Hopper and Blum [22], whose HB protocol was later proven to be secure against Passive attacks assuming the hardness of LPN. The original motivation for the HB protocol was to enable unaided human authentication: the goal was for the protocol to be simple enough to be carried out without the help of a computational device. Subsequent work has found that the key sizes and error rates required to ensure security may be too large for humans to employ with ease comparable to, say, password-based authentication. Nevertheless, as noted by Juels and Weis [23], HB-like protocols are lightweight enough to be potentially applicable in the RFID setting.

There is a long line of research [8, 10, 11, 15–21, 25, 27–31, 34] devoted to devising efficient secret-key authentication protocols based on the hardness of LPN problem and its variants. However, these solutions are based on different security notions and there is no unified security model among these LPN-based authentication protocols, in particular the HB-like protocols (*cf.* Sect. 1.2 for the definition). The relation among some of those security notions is unclear and mentioned as a gap in the literature. In this work, we study the models of attacks on HB-like authentication protocols, including concurrent attack models.

1.1 Our Contribution

In this paper our contributions can be summarized as follows:

- We give the first formal definition of the concurrent Man-in-the-Middle security for the HB-like symmetric key authentication protocols. We name this new model as concurrent two-phase MIM (c2MIM). c2MIM model is a natural extension (concurrent version) of the two-phase MIM security that has been used in the literature for the HB-like protocols. In the first phase (training phase), the attacker is allowed to interact with arbitrarily many instances of the prover and the verifier concurrently. The attacker can eavesdrop on and modify any message, and interleave executions, hoping to learn enough to be able to impersonate the prover in the second phase (attack phase).

- We study and summarize the relation among all security notions for the HB-like authentication protocols (Fig. 1). We show that c2MIM security does not imply sequential MIM (sMIM) security and cMIM security is slightly stronger than c2MIM security. We also investigate the split phase security models in the cryptography literature.

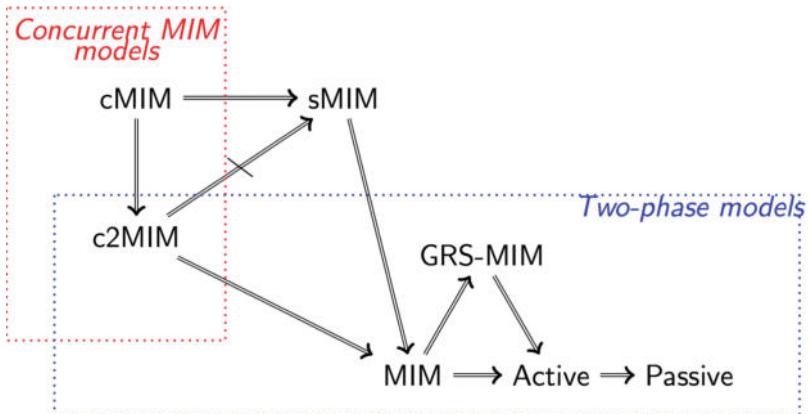


Fig. 1. The relationship among the security notions

1.2 Related Work

The HB-like Protocols The original motivation for the HB protocol [22] was to enable unaided human authentication: the goal was for the protocol to be simple enough to be carried out without the help of a computational device.

The HB protocol has very simple circuit representations. For example, the interaction between the prover, \mathcal{P} and the verifier, or \mathcal{V} , in the HB protocol consists of two messages: first, \mathcal{V} sends a random challenge $\mathbf{a} \in \mathbb{F}_2^n$. Next, \mathcal{P} samples $e \in \mathbb{F}_2$ according to the Bernoulli distribution Ber_ε (i.e. $\Pr[e = 1] = \varepsilon$). \mathcal{P} sends $z = \mathbf{a}^\top \mathbf{x} + e$ to \mathcal{V} , where $\mathbf{x} \in \mathbb{F}_2^n$ is a key shared between \mathcal{P} and \mathcal{V} . \mathcal{V} accepts if $z = \mathbf{a}^\top \mathbf{x}$. This basic authentication step has soundness $\frac{1}{2}$ and completeness $1 - \varepsilon$, but this can be improved via sequential or parallel composition (cf. Sect. 2.2).

In [23], Juels and Weis introduced HB^+ , which was shown to be secure in the stronger Active security model. Gilbert, Robshaw, and Seurin ([20]) showed that HB^+ is vulnerable to a mimic attack. A number of variants of HB^+ were proposed to remedy this defect, including HB^{++} [11], HB^* [15], HB-MP [30], $\text{HB-MP}'$ [28], and Trusted-HB [10]. However, all of these were proven insecure. Gilbert, Robshaw, and Seurin ([18]) extended their attack on HB^+ to break HB^{++} , HB^* , HB-MP , $\text{HB-MP}'$, and Frumkin and Shamir [17] showed that Trusted-HB is insecure. Gilbert, Robshaw, and Seurin [19] introduced $\text{HB}^\#$, which was secure against the

same attack that succeeded against HB^+ . However, Oaifi et al [31] presented an Man-in-the-Middle attack on $\text{HB}^\#$.

Katz et al. [25] provided the first proof of security for HB (Passive secure) and HB^+ (Active secure) for any error rate $\varepsilon < 1/2$, via black box reductions. However, for HB^+ the reduction used rewinding, so that it achieved Active security $\sqrt{\varepsilon}$ assuming LPN is hard for noise rate ε .

Pietrzak then introduced Subspace LWE [32], a more flexible formulation of LPN that is nevertheless equivalent to LPN. In a major advance, Kiltz et al. [27] built on Subspace LWE [32] to construct a two-round Active-secure protocol Auth, as well as two secure MACs, which imply two-round Man-in-the-Middle-secure protocols. As for Auth, Kiltz et al. suggested a modified version where the communication complexity is reduced at the expense of higher storage complexity and claimed that the Auth variant is at least as secure as Auth.

Bosley et al. proposed a simple authentication protocol whose security is based solely on the LPN problem that is secure against Man-in-the-Middle attacks [8].

Rizomiliotis and Gritzalis [34] revisited the security of the two-round Auth protocol [27] and proved that the Auth variant is secure against the MIM attacks under the assumption that it is Active secure (claimed in [27]). Endo and Kunihiro [16] presented a rigorous Active security proof and therefore completed the MIM security proof of Auth variant. However, Endo and Kunihiro [16] also pointed out that the Auth variant is not secure in the sequential Man-in-the-Middle (sMIM) secure model proposed by Lyubashevsky and Masny [29].

Rizomiliotis and Gritzalis [33] proposed $\text{GHB}^\#$, an improvement on $\text{HB}^\#$ based on the Gold Boolean functions. It essentially replaced the linear functions in $\text{HB}^\#$ by non-linear functions. The resulting three-round protocol is proven secure in the MIM model.

Heyse et al. proposed Laptin [21], a lightweight authentication system based on the Ring-LPN problem, a variant of LPN. It is claimed to be provably secure against Active attacks. Sooner after, Bernstein and Lange [5] presents an attack against Ring-LPN-512 and Lapin-512. The attack is based on [6], hence not practical, but nevertheless violates specific security claims in [21].

We classify the above mentioned related works as HB -like protocols. Such symmetric authentication protocols consist of n (parallelable) iterations of the basic authentication step. They can either run sequentially (verify after polynomially many basic steps) or in a parallel manner with two or three rounds, just like the original HB protocol.

Authentication Based on MAC or Weak PRF Under LPN. In the breakthrough work of Kiltz et al. [27], besides Auth, two secure MACs (Message Authentication Code) based on LPN were proposed. As secure MAC implies two-round cMIM-secure authentication protocol, this work gives the first two solutions of building Man-in-the-Middle secure authentications from LPN. However, both Man-in-the-Middle-secure constructions based on MAC require the use of an (almost) Pairwise Independent Permutation on approximately $O(n^2)$ bits.

Furthermore, the first MAC's security reduction is loose, achieving security $\sqrt{\varepsilon}$, while the second construction is much more complicated and requires a longer key.

Dodis et al. [14] explored the direction of authentication from weak pseudorandom primitives and constructed a three-round authentication protocol secure against active attacks from any weak PRF. They proposed another LPN-base MAC, which also leads to a two-round concurrent Man-in-the-Middle-secure (cMIM) authentication scheme. However, this LPN-MAC based protocol has the drawback of large key size.

Lyubashevsky and Masy [29], subsequently, proposed a three-round protocol based on any weak PRF, which is secure against sMIM attacks. The construction is generic and can be instantiated with any version of the LPN function, including classic LPN and its more efficient variants Toeplitz-LPN and Ring-LPN. However, these protocols are not concurrent Man-in-the-Middle (cMIM) secure and all require three rounds.

Recently, Cash et al. [12] proposed two-round secret-key authentication protocol that is secure against sMIM attacks (the same model as [29]). The construction follows from a generic transformation from Active-secure protocol of certain forms, and can be instantiated with concrete problems such as LPN, DDH, and weak PRFs.

Damgård and Park [13] proposed an approach to construct two-round secret-key authentication protocols making black-box use of any pseudorandom generator (PRG) that follows from LPN. Different from other LPN-based work, the proposed protocol has perfect completeness from a variant MAC. It is also claimed to achieve concurrent Man-in-the-Middle security. However, the concurrent model they used is a prover-stateful model, i.e. this protocol requires the prove side to keep a small amount of state.

1.3 Organization

The rest of the paper is organized as follows: In Sect. 2 we first give a summary of the basic notations and terminology. Then we briefly introduce the LPN problem, and the HB, HB⁺ protocols. Next, in Sect. 3, we review the security models for symmetric authentication protocols that has been proposed in the literature, and introduce the new c2MIM model. Section 4 examines the relation between the Man-in-the-Middle security models: c2MIM, sMIM and cMIM. Finally, we conclude the paper in Sect. 5.

2 Preliminaries and Notations

We write $x \xleftarrow{\$} X$ to denote the process of assigning a value sampled from the distribution X to the variable x . If S is a finite set, we write $s \xleftarrow{\$} S$ to denote assignment to s of a value sampled from the uniform distribution on S . We use $[n]$ to denote the set $\{1, 2, \dots, n\}$. Vice versa, we will abuse set-notation to identify a distribution X with its support; for example, we write $x \in X$ to

denote that x is in the support of X . If \mathcal{A} is a probabilistic algorithm, we let $\mathcal{A}(x)$ denote the output distribution of \mathcal{A} on input x , and write $y \xleftarrow{\$} \mathcal{A}(x)$ to denote the process of running algorithm \mathcal{A} on input x and assigning its output to y . We write:

$$\Pr[x_1 \xleftarrow{\$} X_1, x_2 \xleftarrow{\$} X_2, \dots, x_n \xleftarrow{\$} X_n : \phi(x_1, \dots, x_n)]$$

to denote the probability that the predicate $\phi(x_1, \dots, x_n)$ is true, when for all $i \in [n]$, x_i is drawn from distribution X_i , possibly depending on the values drawn for x_1, \dots, x_{i-1} . When $n = 1$, $\hat{x} \in X_1$, and $\phi(x_1)$ is of the form “ $x_1 = \hat{x}_1$ ”, we use the shorthand $\Pr[\hat{x}_1 \xleftarrow{\$} X]$ to denote $\Pr[x_1 \xleftarrow{\$} X_1 : x_1 = \hat{x}_1]$. For two probability distributions X_1, X_2 , we write $X_1 \equiv X_2$ if and only if $\forall \hat{x} \in X_1 \cup X_2$, $\Pr[\hat{x} \xleftarrow{\$} X_1] = \Pr[\hat{x} \xleftarrow{\$} X_2]$.

Let \mathbb{F}_q represent the finite field with q elements. We denote the uniform distribution over \mathbb{F}_2^n by U_n , and the Bernoulli distribution with bias ε by Ber_ε . So $\Pr[1 \xleftarrow{\$} \mathsf{Ber}_\varepsilon] = \varepsilon$ and $\Pr[0 \xleftarrow{\$} \mathsf{Ber}_\varepsilon] = 1 - \varepsilon$.) We use the binary operator $\oplus: \mathbb{F}_2 \times \mathbb{F}_2 \rightarrow \mathbb{F}_2$ to represent finite field addition, and for $b \in \mathbb{F}_2$, we let $\bar{b} = 1 \oplus b$ be the complement of b . For an event S , \overline{S} represents its complement, the event that S does not occur.

We denote column vectors by lower-case bold letters such as \mathbf{x} , and matrices by upper-case bold letters such as \mathbf{X} . We denote the transpose of \mathbf{X} by \mathbf{X}^\top . $\mathbf{0}$ denotes the all-zero column vector. We will often consider column vectors $\mathbf{x}, \mathbf{y} \in \mathbb{F}_2^\ell$ as matrices in $\mathbb{F}_2^{\ell \times 1}$. Considering \mathbf{x}, \mathbf{y} as matrices allows us to extend operations on matrices to vectors. For example, we can form the outer product $\mathbf{x}\mathbf{y}^\top \in \mathbb{F}_2^{\ell \times \ell}$, and form the kernel $\ker(\mathbf{x})$. The dot product of two column vectors \mathbf{x}, \mathbf{y} can be written as the matrix multiplication $\mathbf{x}^\top \mathbf{y}$. For a vector \mathbf{x} , we denote the scalar i -th element of \mathbf{x} by $\mathbf{x}[i]$. For a vector \mathbf{x} , let $|\mathbf{x}|$ denote the number of nonzero entries of \mathbf{x} .

We denote an arbitrary polynomial function of n by $\text{poly}(n)$. We write $f = \text{negl}$ to mean that f is negligible as a function of n , that is, $f = o(n^{-c})$ for any constant $c > 0$.

2.1 Learning Parity with Noise (LPN)

Roughly speaking, the problem of Learning Parity with Noise amounts to distinguishing two distributions over $\mathbb{F}_2^n \times \mathbb{F}_2$: the uniform distribution and the LPN distribution. For a fixed secret vector $\mathbf{x} \xleftarrow{\$} \mathbb{F}_2^n$, the LPN distribution is in turn defined in terms of its sampling algorithm $\mathsf{LPN}_\varepsilon^\mathbf{x}$, shown in Algorithm 1. Algorithm $\mathsf{LPN}_\varepsilon^\mathbf{x}$ is initialized with a uniform secret vector $\mathbf{x} \xleftarrow{\$} \mathbb{F}_2^n$. Thereafter, whenever an LPN sample is requested, the algorithm chooses random $\mathbf{a} \xleftarrow{\$} \mathbb{F}_2^n$ and $e \xleftarrow{\$} \mathsf{Ber}_\varepsilon$ and outputs (\mathbf{a}, b) , where $b = \mathbf{a}^\top \mathbf{x} \oplus e$. For $\varepsilon = \frac{1}{2}$, LPN becomes the uniform distribution.

```

1: function LPN
    $\overset{\mathbf{x}}{\underset{\varepsilon}{\leftarrow}}$ 
2:    $\mathbf{a} \overset{\$}{\leftarrow} \mathbb{F}_2^n$ 
3:    $e \overset{\$}{\leftarrow} \text{Ber}_\varepsilon$ 
4:    $b = \mathbf{a}^\top \mathbf{x} + e$ 
5:   return ( $\mathbf{a}, b$ )

```

Algorithm 1. LPN

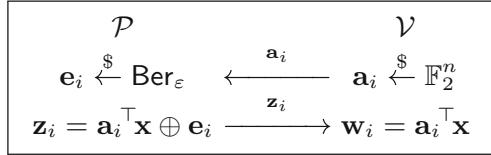
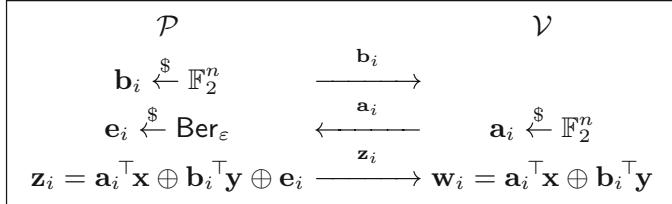
We will use the decisional version of the LPN hardness assumption, which is defined using an indistinguishability game. It has been shown [25] that hardness of the decisional version is equivalent (up to polynomial factors) to hardness of recovering the entire key. The decisional variant of LPN is hard if it is difficult to distinguish between an oracle with distribution $\text{LPN}_\varepsilon^{\mathbf{x}}$ for random \mathbf{x} versus an oracle with a random distribution $U_n \times U_1$ can be represented as $\text{LPN}_{1/2}^{\mathbf{x}}$. More formally, the advantage of an algorithm \mathcal{A} against LPN for a given (ε, n) is defined using a game in which the adversary attempts to guess which oracle was selected:

Definition 1. The decisional LPN assumption states that for all efficient adversaries \mathcal{A} , $\text{Adv}_{\mathcal{A}}^{\text{LPN}}(\varepsilon, n) \leq \delta_{\text{LPN}} = \text{negl}$, where $\text{Adv}_{\mathcal{A}}^{\text{LPN}}(\varepsilon, n)$ is defined as

$$\text{Adv}_{\mathcal{A}}^{\text{LPN}}(\varepsilon, n) = \left| \Pr \left[\begin{array}{l} \mathbf{x} \overset{\$}{\leftarrow} \mathbb{F}_2^n, \beta \overset{\$}{\leftarrow} \mathbb{F}_2, \\ \mathcal{O}_\beta = \begin{cases} \text{LPN}_{1/2}^{\mathbf{x}} & \text{if } \beta = 0 \\ \text{LPN}_\varepsilon^{\mathbf{x}}, & \text{if } \beta = 1 \end{cases}, : \hat{\beta} = \beta \\ \hat{\beta} \overset{\$}{\leftarrow} \mathcal{A}^{\mathcal{O}_\beta}(1^n) \end{array} \right] - \frac{1}{2} \right|$$

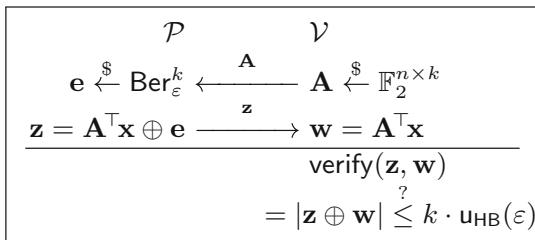
2.2 HB and HB⁺ protocols

The HB, HB⁺ protocols consist of $k = \text{poly}(n)$ iterations of what is known as a “basic authentication step”. The protocols are executed by two parties: the prover \mathcal{P} , and the verifier \mathcal{V} . The key for HB is a vector \mathbf{x} of length n , where n is the security parameter. For HB⁺, the key consists of two vectors \mathbf{x}, \mathbf{y} of length n . For $i \in [k]$, $\mathbf{a}_i, \mathbf{b}_i \in \mathbb{F}_2^n$ are column vectors used in the execution. In HB, as shown in Fig. 2, a prover \mathcal{P} and a verifier \mathcal{V} share a random secret key $\mathbf{x} \in \mathbb{F}_2^n$. In the i -th authentication step, the verifier sends a random challenge $\mathbf{a}_i \in \mathbb{F}_2^n$ to the verifier, and the prover replies with $\mathbf{z}_i = \mathbf{a}_i^\top \mathbf{x} \oplus \mathbf{e}_i$, where $\mathbf{e}_i \overset{\$}{\leftarrow} \text{Ber}_\varepsilon$. HB⁺ adds a second secret $\mathbf{y} \in \mathbb{F}_2^n$ and a third round, as shown in Fig. 3. In both HB and HB⁺, at the end of k steps, \mathcal{V} checks to see what fraction of answers \mathbf{z}_i were correct. If more than $k \cdot u(\varepsilon)$ are correct, for $u(\varepsilon)$ some function of ε , then the verifier accepts. Otherwise, the verifier rejects. k and $u(\varepsilon)$ should be set high enough to allow the honest prover to authenticate w.h.p., but low enough that a malicious third party should not be able to authenticate by randomly guessing. In particular, as noted in [25], for both HB and HB⁺, $u(\varepsilon) = (1 + \tau)\varepsilon$ suffices to

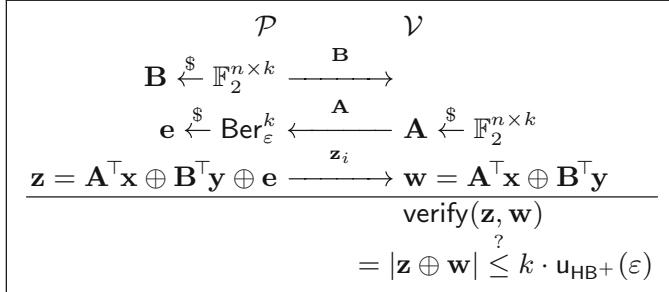
**Fig. 2.** HB (the i -th authentication step)**Fig. 3.** HB⁺ (the i -th authentication step)

achieve completeness error negligible in the security parameter, for any positive constant τ .

We can use matrix notation to simplify working with the HB and HB⁺ protocols in parallel, as shown in Figs. 4 and 5. We adapt the parallel notation in the rest of the paper. Let $\mathbf{A}, \mathbf{B} \in \mathbb{F}_2^{n \times k}$ be matrices for which $\forall i \in [k], \mathbf{Ad}_i = \mathbf{a}_i, \mathbf{Bd}_i = \mathbf{b}_i$. That is, the i -th columns of \mathbf{A}, \mathbf{B} respectively are the vectors $\mathbf{a}_i, \mathbf{b}_i$ respectively. Then in the HB protocol, for example, \mathcal{V} sends the challenge $\mathbf{A} \xleftarrow{\$} \mathbb{F}_2^{n \times k}$. \mathcal{P} replies with $\mathbf{z} = \mathbf{A}^\top \mathbf{x} \oplus \mathbf{e}$, where $\mathbf{e} \xleftarrow{\$} \text{Ber}_{\varepsilon}^k$. \mathcal{V} computes $\mathbf{w} = \mathbf{A}^\top \mathbf{x}$, and accepts iff $|\mathbf{z} \oplus \mathbf{w}| \leq u_{\text{HB}}(\varepsilon)$.

**Fig. 4.** HB (Parallel Notation)

Similar to HB and HB⁺, HB-like variants such as HB⁺⁺ [9], HB[#] [24], HB^N [8], etc. all consist of n parallelable iterations of the basic authentication step. They can either run in a parallel or sequential manner with two (like HB) or three rounds (like HB⁺).

**Fig. 5.** HB^+ (Parallel Notation)

3 Security Models for Symmetric Key Authentication Protocols

In this section, we first present several natural security models that have been used for authentication and for two round HB-like protocols in particular. The more general models are **Passive**, **Active**, and **MIM**, which are presented in order of increasing strength, and the stronger attacks imply the weaker ones. All the three attacks run in two phases and only differ in the first phase.

In the weakest model, **Passive** attack model, the adversary in the first phase may observe several sessions between an honest prover and an honest verifier. Later in the second phase, the adversary will try to impersonate the prover, trying to authenticate as the prover to the verifier.

A stronger type of model, **Active** attack model, is the adversary who in the first phase may run many authentication protocols with the prover. But the adversary has no power to reset the prover. Then afterward, without access to the prover, the adversary tries to fool the verifier in the second phase.

An even stronger one, **MIM** (**Man-in-the-Middle**) (two-phase) attack model, allows the adversary first sequentially run many authentication protocols in both the verifier's role with the real prover and the prover's role with the honest verifier, hoping to learn enough to be able to impersonate the prover in a future time. Note that the adversary learns the decision made by the verifier, i.e. whether she is accepted or not. But an accept by an honest verifier in this stage is not considered as a winning. After the first stage, the adversary loses access to the prover and tries to authenticate to the honest verifier.

Additionally, several works have used an intermediate attack model, **GRS-MIM**, which is stronger than **Active** model yet weaker than the **MIM** model. The **GRS-MIM** model of Gilbert et al. [19] is a variant of the **Man-in-the-Middle** model, in which the adversary is only allowed to modify the messages from the verifier to the prover, but not in other direction.

A more realistic setting, especially in the era of the Internet, is one that allows the concurrent execution of protocols. In a concurrent attack [4], an adversary firstly acts as a verifier. It interacts with many prover instances that have inde-

pendent inner states and random tapes, but the same secret key. It may use some cheating trick in messages to collect information of the secret key from the responses from provers. After the completion of acting as a verifier, in the second phase the adversary tries to impersonate the prover against a victim verifier using that collected information. Then concurrent attack model can be view as an concurrent extension of the Active model.

“Man-in-the-Middle attack” is a broad term for any attack in which communication between honest parties may be corrupted by an adversary. Man-in-the-Middle attacks are well-known to the security protocols community and then brought into the Cryptography community. A security protocol will typically be proved secure against a Yao [35] attacker, a powerful definition of Man-in-the-Middle from 1983. However, defining security against Man-in-the-Middle attacks for cryptographic protocols is not simple, as many cryptographic primitives still have no universally agreed-upon definitions for security against man-in-the-middle attacks.

Most of the earlier work of secret-key authentication did not consider the concurrent execution of protocols and there was a lack of a clear and agreed definition on the concurrent Man-in-the-Middle security model. Now considering the adversary can get an advantage by participating in several concurrent executions of the protocols. We give a natural extension of the (two-phase) MIM model—concurrent two-phase Man-in-the-Middle model (c2MIM, for short) which allows the attacker to interact with arbitrarily many instances of the prover and the verifier concurrently in the first phase. The attacker can eavesdrop on and modify any message, and interleave executions, hoping to learn enough to be able to impersonate the prover in a future (the second phase). Additionally, the attacker learns the decisions made by the verifier \mathcal{V} . But, the success of the adversary getting accepted by any verifier instance in phase I, is not considered as a winning. In the attack stage (phase two), without access to the prover, the adversary interacts once with the honest verifier. The formal definition for c2MIM is given as follows.

Definition 2. (*Concurrent two-phase MIM (c2MIM) security*). A secret-key authentication protocol $(\mathcal{P}, \mathcal{V})$ with shared secret key \mathbf{x} is secure against c2MIM attacks if for any probabilistic poly(n)-time (PPT) adversary \mathcal{A} which in the first phase can interact arbitrarily polynomially many times with an honest prover \mathcal{P} and/or an honest verifier \mathcal{V} (the interactions may be concurrent), and learns the accept/reject decisions of the verifier. Then afterward, in the second phase, without access to \mathcal{P} and \mathcal{V} , the adversary interacts once with an honest verifier \mathcal{V}^* , it holds that¹

$$\begin{aligned} \text{Adv}_{\mathcal{A}}^{\mathcal{P}, \mathcal{V}, \mathcal{V}^*} &= \Pr \left[\Sigma_{\mathcal{A}} \xleftarrow{\$} \mathcal{A}^{\mathcal{P}, \mathcal{V}}(1^n) : (\mathcal{A}(\Sigma_{\mathcal{A}}), \mathcal{V}^*) = \text{accept} \right] \\ &\leq \text{negl} \end{aligned}$$

The security models discussed so far can be summarized in the following figures Figs. 6, 7, 8 and 9. In an HB-like interactive protocol between two parties,

¹ $\Sigma_{\mathcal{A}}$ is the state information from \mathcal{A} in Phase I.

we use \mathbf{R} to denote a random challenge generated by the party on the right, \mathbf{L} the randomness generate by the party on the left, and \mathbf{z} the response computed based on both parties random values. Another concurrent definition cMIM (concurrent Man-in-the-Middle attack) can be adapted from the CR2 attack in public-key identification setting [2,3]. In a cMIM attack, there is no line drawn there to split it into two phases. At the attack stage, the adversary may simultaneously communicate with multiple instances of prover (in the role of a prover) and/or verifier (in the role of a verifier). The adversary wins whenever it manages to let one of the verifier instances accept. The definitional issue in the cMIM model is complex and of course an real attack should excludes simply passing messages back and forth between prover and verifier. One method to address this can be introducing session ids in a similar way as [3]. Clearly, a c2MIM attack is a special case of a cMIM attack. An analogy with CCA (chosen-ciphertext attacks) in encryption might be useful to understand the distinction between the (non) two-phase setting, and helps explain why we consider the line we have drawn to split the two phases in between to be nature. We will discuss this more in next section.

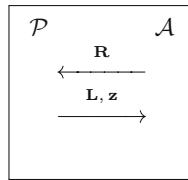


Fig. 6. Phase I (Active)

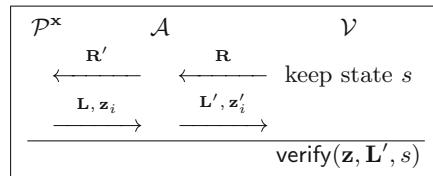
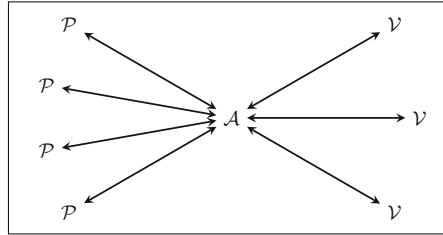
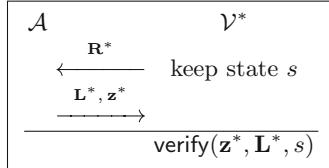


Fig. 7. Phase I (MIM)

The security notion of sequential security against Man-in-the-Middle attacks (or sMIM security, for short) [29] weakens cMIM to only allow the adversary to interfere with independent (non-overlapping) sequential sessions between the same pair of prover and verifier. The adversary wins whenever it manages to let the verifier accept in some session and has changed at least one of the messages sent by the prover or the verifier. Note that in this model, similar to the cMIM, the attack was not formed in two phases as above. If view the last session in

**Fig. 8.** Phase I (c2MIM)**Fig. 9.** Phase II (All Models)

which the adversary finally wins as the second phase in the view of a two-phase model, the adversary was actually given access to the prover while trying to impersonate the prover to the verifier in this attack model. So **sMIM** is stronger than **MIM**.

4 Relation among the Man-in-the-Middle Security Models

From the definitions and discussions in the previous section, it is clear that the series of two phases security models for HB-like symmetric key authentication protocols can be ordered with increasing strength as follows: **Passive** < **Active** < **GRS-MIM** < **MIM** < **c2MIM**; And for the non-splitting models, **sMIM** < **cMIM** as **sMIM** is weakened from **cMIM** [29]. In this section, we present more discussions on the Man-in-the-Middle security models and their relationship as shown in Fig. 1.

The **cMIM** and **sMIM** model as we introduced earlier have no split phase for the attack stage, i.e. the adversary may still access the prover instance (concurrently) while trying to cheat the verifier. An analogy with chosen-ciphertext attack (CCA) model in encryption is useful to understand the difference of such authentication models from the two-phase models. Recall that in the formulation of CCA security, the decryption oracle is provided to an adversary and the adversary must provide response to the challenge ciphertext. If view such setting as the analogue of the adversary (when attacking an authentication scheme) is given access to prover instances and must have interaction with the verifier instance, the IND-CCA1 setting, allowing access to the decryption oracle only prior to the appearance of the challenge, is the analogue of the two-phase **Man-in-the-Middle** setting; the IND-CCA2 setting, allowing access to the decryption oracle even after the appearance of the challenge is analogue of the non-phase-split model.

In an IND-CCA2 setting, the adversary can trivially win by querying its decryption oracle with the challenge ciphertext; the analogue is that the adversary in the authentication protocol could make the verifier accept by simply passing messages back and forth between the prover and verifier. The definitional “fix” of the IND-CCA2 setting is to simply disallow this one query. However, the fix for authentication setting is less trivial. The sMIM model fixed this by defining the winning of the adversary as “whenever it manages to let the verifier accept in some session and has changed at least one of the messages sent by the prover or the verifier”. Cash et al. [12] further enhanced the definition by introducing session ids.

From the above discussion, we can see the c2MIM can be viewed as a special case of cMIM, which yields $c2\text{MIM} < c\text{MIM}$. The c2MIM security does not imply sMIM security, and we don’t know the other direction yet.

However the sMIM definition is not clear enough for the HB-like protocols. As we stated before, the HB-like protocols consist of n (parallelable) iterations of the basic authentication step. They can either run sequentially (verify after polynomially many basic steps) or in a parallel manner with two or three rounds. The parallel nature of HB-like protocols seems difficult to achieve with factoring or discrete-log type assumption. While runs in parallel, all the HB-like protocol including the MIM secure authentication protocols such as HB^N [8], Auth variant [27] and $\text{GHB}^\#$ [33] suffers a same attack as shown in [16] under the current definition of sMIM. The attack is simple. The adversary just simply flip one bit of the response (vector) sent from the prover (\oplus the response with any unit vector) and pass the slightly perturbed response to the verifier. The verifier accept such modified response still with high probability. However, when the HB-like protocols run in serial, the above attack is obvious not valid. As the adversary is still honestly forwarding back and forth the messages of the majority of basic authentication steps.

Similarly, HB^N and other HB-like protocols suffer a “malleable” attack in the cMIM model, in which, the adversary can first decompose the challenge sent by the verifier into two or more pieces; concatenate the “challenger piece” with some random information to form new challenges; then make queries to multiple prover instances on each new challenger (which can be viewed as a superset of the real challenge piece); and finally retrieve the correct response to the verifier from the multiple responses of the prover instances.

Such attacks are essentially caused by the parallel nature of HB-like protocol and are not real attacks. Even the adversary successfully passed the verification of the authentication protocol, she essentially has no information about the shared secret key. So the definition should exclude it. For example, in the c2MIM model, such “malleable” attack is impossible. We leave this as open discussions whether more strict definitions are needed for sMIM and cMIM, and how this issue can be fixed, e.g. define that during the attack stage, the challenge an adversary send to prover has empty intersection to the challenge she received from the verifier.

5 Conclusion and Future Work

HB-like protocols are based on the hardness of the Learning Parity with Noise (LPN) problem. Such protocols based on bit XOR operations may suitable for low-cost “Internet of Things” devices, whose limited circuit size and power constraints rule out the use of more heavyweight operations such as modular exponentiation. There are various security models proposed for such HB-like symmetric authentication protocols in the literature. However, a clear relation between these security notions are missing. In this work, we aimed at filling this gap. We first proposed a new concurrent Man-in-the-Middle security model (c2MIM) and then summarized the relation between all the security models in Fig. 1. During our work on this topic, we identified some delicate definitional issues of the cMIM and sMIM security notions for authentication protocols. Possible extensions of this work that can be undertaken as future research work include formalizing the cMIM security model definition, improving and understanding the relationship between c2MIM and sMIM model, and design authentication schemes secure under cMIM and/or c2MIM model.

References

1. Angluin, D., Laird, P.D.: Learning from noisy examples. *Mach. Learn.* **2**(4), 343–370 (1987)
2. Bellare, M., Rompel, J.: Randomness-efficient oblivious sampling. In: 35th Annual Symposium on Foundations of Computer Science, pp. 276–287. IEEE (1994)
3. Bellare, M., Fischlin, M., Goldwasser, S., Micali, S.: Identification protocols secure against reset attacks. In: Pfitzmann, B. (ed.) *Advances in Cryptology, EUROCRYPT 2001: International Conference on the Theory and Application of Cryptographic Techniques*, pp. 495–511. Springer, Berlin, Heidelberg (2001)
4. Bellare, M., Palacio, A.: GQ and Schnorr identification schemes: Proofs of security against impersonation under active and concurrent attacks. In: Yung, M. (ed.) *Advances in Cryptology—CRYPTO 2002*, pp. 162–177. Springer, Berlin, Heidelberg (2002)
5. Bernstein, D.J., Lange, T.: Never trust a bunny. IACR Cryptology ePrint Archive, vol. 2012, 355 (2012)
6. Blum, A., Kalai, A., Wasserman, H.: Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM (JACM)* **50**(4), 519 (2003)
7. Blum, A., Furst, M.L., Kearns, M.J., Lipton, R.J.: Cryptographic primitives based on hard learning problems. In: Proceedings of the 13th Annual International Cryptology Conference on Advances in Cryptology, CRYPTO ’93, pp. 278–291. Springer, London, UK (1993)
8. Bosley, C., Haralambiev, K., Nicolosi, A.: HB^N : a variant of HB secure against man-in-the-middle attacks (2011)
9. Bringier, J., Chabanne, H.: Trusted-HB: HB against man-in-the-middle attacks. EPrint
10. Bringier, J., Chabanne, H.: Trusted-HB: a low-cost version of HB secure against man-in-the-middle attacks. arXiv (2008)

11. Bringer, J., Chabanne, H., Dottax, E.: HB⁺⁺: a lightweight authentication protocol secure against some attacks. In: Second International Workshop on Security, Privacy and Trust in Pervasive and Ubiquitous Computing (SecPerU 2006), pp. 28–33. IEEE Computer Society (2006)
12. Cash, D., Kiltz, E., Tessaro, S.: Two-round man-in-the-middle security from LPN. In: Theory of Cryptography: 13th International Conference. TCC 2016-A, Tel Aviv, Israel, 10–13 Jan 2016, Proceedings, Part I, pp. 225–248. Springer, Berlin, Heidelberg (2016)
13. Damgård, I., Park, S.: Towards optimally efficient secret-key authentication from PRG. Cryptology ePrint Archive, Report 2014/426 (2014)
14. Dodis, Y., Kiltz, E., Pietrzak, K., Wichs, D.: Message authentication, revisited. In: Advances in Cryptology—EUROCRYPT 2012—31st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Proceedings, pp. 355–374, Cambridge, UK, 15–19 Apr 2012
15. Duc, D., Kim, K.: Securing HB against GRS man-in-the-middle attack. cais-lab.icu.ac.kr (2008)
16. Endo, K., Kunihiro, N.: On the security proof of an authentication protocol from eurocrypt 2011. In: Yoshida, M., Mouri, K. (eds.) Advances in Information and Computer Security: 9th International Workshop on Security, IWSEC 2014, Proceedings, pp. 187–203, Hirosaki, Japan, 27–29 Aug 2014. Springer International Publishing (2014)
17. Frumkin, D., Shamir, A.: Un-Trusted-HB: Security Vulnerabilities of Trusted-HB. EPrint (2009)
18. Gilbert, H., Robshaw, M., Seurin, Y.: Good variants of HB⁺ are hard to find. In: Proceedings of Financial Cryptography and Data Security, pp. 156–170 (2008)
19. Gilbert, H., Robshaw, M., Seurin, Y.: HB[#]: increasing the security and efficiency of HB. In: Proceedings of EUROCRYPT, vol. 4965, pp. 361–378 (2008)
20. Gilbert, H., Robshaw, M., Sibert, H.: Active attack against HB⁺: a provably secure lightweight authentication protocol. Electron. Lett. 4(21), 1169–1170 (2005)
21. Heyse, S., Kiltz, E., Lyubashevsky, V., Paar, C., Pietrzak, K.: Lapin: an efficient authentication protocol based on ring-LPN. In: FSE, pp. 346–365 (2012)
22. Hopper, N., Blum, M.: Secure human identification protocols. In: Proceedings of ASIACRYPT (2001)
23. Juels, A., Weis, S.: Authenticating pervasive devices with human protocols. In: Proceedings of CRYPTO, pp. 293–308 (2005)
24. Katz, J., Shin, J.S.: Parallel and concurrent security of the HB and HB⁺ protocols. Eurocrypt (2006)
25. Katz, J., Shin, J.S., Smith, A.: Parallel and concurrent security of the HB and HB⁺ protocols. J. Cryptology **23**(3), 402–421 (2010)
26. Kearns, M.: Efficient noise-tolerant learning from statistical queries. In: Proceedings of the 25th ACM Symposium on Theory of Computing, pp. 392–401. ACM (1993)
27. Kiltz, E., Pietrzak, K., Cash, D., Jain, A., Venturi, D.: Efficient authentication from hard learning problems. In: Proceedings of Eurocrypt, pp. 7–26 (2011)
28. Leng, X., Mayes, K., Markantonakis, K.: HB-MP+ protocol: an improvement on the HB-MP protocol. In: 2008 IEEE International Conference on RFID (2008)
29. Lyubashevsky, V., Masny, D.: Man-in-the-middle secure authentication schemes from LPN and weak PRFs. In: Advances in Cryptology—CRYPTO 2013—33rd Annual Cryptology Conference, Santa Barbara, CA, USA, 18–22 Aug 2013. Proceedings, Part II, pp. 308–325 (2013)

30. Munilla, J., Peinado, A.: HB-MP: a further step in the HB-family of lightweight authentication protocols. *Comput. Netw.* **51**(9), 2262–2267 (2007)
31. Ouafi, K., Overbeck, R., Vaudenay, S.: On the security of HB $^{\#}$ against a man-in-the-middle attack. In: Proceedings of ASIACRYPT (2008)
32. Pietrzak, K.: Subspace LWE (2010), manuscript available at <http://homepages.cwi.nl/~pietrzak/publications/SLWE.pdf>
33. Rizomiliotis, P., Gritzalis, S.: GHB $^{\#}$: A Provably Secure HB-Like Lightweight Authentication Protocol, pp. 489–506. Springer, Berlin, Heidelberg (2012)
34. Rizomiliotis, P., Gritzalis, S.: Revisiting lightweight authentication protocols based on hard learning problems. In: Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks, WiSec'13, pp. 125–130. ACM, New York, NY, USA (2013)
35. Yao, D.D.A.: On the security of public key protocols. *IEEE Trans. Inf. Theor.* **29**(2), 198–208 (1983)



Development of Students' Security and Privacy Habits Scale

Naurin Farooq Khan^(✉) and Naveed Ikram

Riphah International University, Islamabad, Pakistan
{naurin.zamir, naveed.ikram}@riphah.edu.pk

Abstract. The cyber space offers many opportunities to general public but it has its dark side in terms of cyber crimes. Apart from technical security, the human factor plays an important role in safe guarding the security and privacy of systems. The end users especially students need to be aware of protective measures they can adopt to safe guard themselves through security awareness and trainings. The awareness programs should be comprehensive and be tailored according to the security and privacy awareness of the individuals. Therefore, the target individuals should be assessed in terms of their security and privacy habits and practices. The previous endeavors in this respect make use of questionnaire instruments that are specific to a particular type of individuals such as employees of an organization or tap onto use of a specific device. This study presents development of an instrument that gauges the security and privacy habits/practices of end users specifically students.

Keywords: Information security · Development of questionnaire · Information security awareness and training

1 Introduction

The cyber space is laden with cyber criminals and other miscreants who are on a constant look out for opportunities to carry out their nefarious purposes. Individuals—the weakest link in the security chain with their risky behavior not only put themselves at danger but also the overall IT systems and the people associated with them [1]. Their security and privacy habits while they surf the cyber space are questionable such as sharing passwords with others, neglecting basic security settings on mobile devices and downloading illegal software and utilities to name a few. Of all the segments of society, youngsters particularly students exhibit carefree and careless attitude towards their security [2]. The key to addressing human factors and competencies in information security is awareness, training, and education [3]. However, the awareness training should be designed keeping in view the need of the end user which entails that the security and privacy habits and practices of individuals should be assessed [4].

This study presents a comprehensive assessment instrument that can be used to gauge the security and privacy practices of the individuals especially students with the aim of better designing the security awareness programs. Section 2 presents the

background for carrying out this study followed by related work in Sect. 3. Section 4 presents the development of the instrument with Sect. 5 concluding the study and presents the future directions.

2 Background

2.1 Cyber Space Landscape

There are more than 4 billion Internet users out of which 3.3 billion are active on the social media [5]. These statistics reflect the fact that our lives are very much dependent on the cyber space. The expectation are that the Internet-based devices such as Internet of Things (IoT) will reach 50 billion devices by the year 2020 [6]. In addition to industrial, governmental and military uses of the internet, people now use the cyber-space to not only socialize but also perform professional, political, welfare and personal care related activities. The statistics confirm that claim, with 2 billion people actively using Facebook [7], 700 million bloggers on Tumblr [8] and more than 800 million people actively using Instagram. There are over 1 billion Youtubers with at least 38% female users [9]. This social activity over the Internet is reflective of enormous amount of personal data being generated and shared by millions of people every day. However, the ease and comfort of being able to share personal and professional data has some inherent risks. Unfortunately, the data generated by millions of people make it easy for the criminals and other miscreants to use it for their nefarious designs. Moreover, the annual losses from cybercrime is estimated to be close to \$500 billion to the global economy.

In order to safeguard against such attacks many methods/techniques have been developed both at a software and hardware level to secure the security loop holes that can be exploited by the cyber criminals [10]. These methods are technical in nature and they are aimed at achieving the security by exploiting latest cryptography algorithms and other latest sophisticated security measures. However, the claimed high security provided by these methods is contrary to the escalating number of security breaches occurring all around the world [11].

2.2 Human Factor of Security

One of the factors that is responsible for such high number of security breaches is the exclusive focus on the technical side of the external attacks by the security organizations. The security pertaining to individuals is fairly neglected. It should be noted that the technical methods are impotent and cannot work on their own without having to consider the human factor—the weakest link in the security chain [12]. So it is important in order to minimize the number of security incidents that security awareness weaknesses of individuals should be taken into account.

Schultz in (2001) addressed the need of the human factor in information security and pointed out not only users resisted the information security measures but also their use of different security tools and mechanism was not optimal [13]. This has been

acknowledged only recently that human factor has been researched to play a central role in the security by different studies [14].

End users who are non-malicious pose threat to their—as well as—the security of their peers due to noncompliance of security policies. They are said to be far more responsible for any theft/loss than the malicious users. These non-malicious end users are responsible for inappropriate use of technology due to their significant lower levels of information security awareness as a result of which they pose risks to themselves and others [10]. For instance, smartphone users are known to lack awareness of basic security knowledge and are not prepared for making appropriate security decision [15]. And then there is voluntary security misconduct on the part of the end user that undermines their security. The culture of self-disclosure on social network sites where people disclose sensitive information about themselves and people who are connected to them [16].

2.3 Students Security Awareness

Of all the segments of the society, students are considered quite vulnerable to the cyber threats due to their carelessness and carefree attitudes. Throughout college years, students leave a significant “digital footprint” visible to others about their personal and academic life. As a result, their privacy is very much at risk [17]. This is due to the fact that educational institutes provide a forum for easy exchange of information and knowledge. Teaching faculty frequently checks students’ records and communicates privileged information (names and campus identification numbers, grades, etc.) with students over email and websites. Students use Learning Management Systems (LMS) for their studies. As a result, these institutes are treasure trove of large amount of data related with students, including financial records, transcripts, credit histories, medical histories, contact information, social security numbers and other personally identifiable information [18]. The importance of information that can be extracted from tertiary institutions can be gauged by the fact that the price for such information in cyber black market can extract billions of dollars [19]. This concludes that if students are not knowledgeable of their institutions information security policy, they may be at a disadvantage of not understanding the risks of using information systems and the potential damage that can result [4]. As per evidence from the literature, although students from tertiary institutes are concerned about security [20], they lack complete knowledge of security practices and risks posed to them if they don’t follow those security practices [19]. It may be true that college students may be technologically well informed but it does not mean that they know how to protect their information and systems effectively [4]. Moreover, students frequently do not safeguard and unintentionally exchange personal information that should be protected. Not only they spend a copious amount of time on the internet but also are reckless in their technology usage [21]. Vulnerabilities, sometimes rooted in a “naive” student culture, can be observed in students’ practices of social networking, sharing passwords and student identification numbers with friends, and not protecting data on mobile devices and media [2]. The persistent need to be always connected put them at a clear online risks. Student in order to be accepted in the online social media society may feel pressured to reveal personal information about themselves [22].

2.4 Security Awareness and Training Is the Key

To fully exploit the potential of the cyberspace, it is important that the end users not only trust and have confidence on the Internet based economy but also are aware of best ways to protect themselves from the cyber threats. Different organizations put emphasis on training of their employees. According to Ernst & Young “*To protect information and systems effectively, organizations invest more in training and awareness programs that prevent users from being the weakest link in a security chain*”. “*One of the best uses of the information security budget is for comprehensive information security awareness programs for users because organizational management needs to realize the importance of information security awareness among users*” [23]. Similarly, computer security resource centers such as NIST, makes it mandatory requirement for organization employees to be trained and have awareness on basic security. As per NIST SP 800-16, this foundational knowledge is important to change the security attitudes of the employees [24]. Therefore, it can safely be concluded that the key to addressing human factors and competencies in information security is awareness, training, and education [3].

Similarly, due to the heightened need of information security and privacy awareness, it is considered an important theme in IT curricula and is part of IT Body of Knowledge [25]. The goal of a security awareness program is to heighten the importance of both information systems security and the possible negative effects of a security breach or failure [26, p. 1]. Therefore end-users specifically students should be trained on how to use the internationally shared space with responsibility by providing them security and privacy awareness training. Through the training, students can educate themselves and their peers about the necessary security concepts and skills needed to protect themselves. This need—to protect and train the citizens as a national priority is reflected in most cyber security policies of different nations around the world [20].

As per empirical evidence from the literature, a significant relationship exist between attending security training and belief about sufficient protection, it is imperative that the security awareness and training be comprehensive [4]. Moreover, the students should be given security training intermittently and should repeat on a regular basis as per change in the security landscape [4]. This entails that the security awareness training should be modified to incorporate the information about the latest and emerging threats.

3 Related Work

3.1 Previous Awareness and Training Programs

A systematic review on methods to assess cyber security awareness interventions synthesized results of 14 studies [27]. Most of the studies target audience was organizations and the assessment methods used involved questionnaires, game tools, focus groups and interviews to name a few. A recent survey of security awareness training was conducted to reveal that the methods for security training should be designed in such a way that the user find them attractive [28]. However, the review only covers the

trainings delivered to provide awareness on phishing and its variants. Another review was carried out to find out factors responsible for better information security awareness [29]. The study accumulated relevant literature on awareness programs along with other antecedents. The review was targeted towards studies that reported on the employees and organizations security behaviour and knowledge.

Some other studies provide security training and awareness to students and try to gauge its effectiveness by conducting experiments [30, 31], vocabulary test [32], conducting test by generating emails to ascertain the email security awareness [33], taking interviews to test the information security policy of schools [34]. While other studies were related to professional information security education [35, 36].

3.2 Limitations of the Security Awareness Program

The following limitations were observed in the security awareness programs.

- Programs tapping onto a particular aspect of cyber security
- Programs tilted more toward employees and organizations
- Programs do not incorporate change as per new cyber security threats
- No structured training program for youngsters specially students
- Lack of holistic content encompassing the different aspects of cyber space
- The assessments of the program is not done using program evaluation techniques.

In order for a structured security awareness and training program to be effective, one of the important steps is understanding the students' information security awareness level. There are few studies which have made contributions in measuring the security and privacy habits and practices of students [19]. These studies posit that the awareness programs need to incorporate the assessment of the target population. In doing so they should identify the short comings of security and privacy awareness of the target population on the basis of which a comprehensive awareness program can be designed. Therefore any endeavor should be geared towards constantly measuring students security habits and awareness in order to provide the current training contents the students need [4].

3.3 Previous Instruments

There has been few studies that make use of questionnaires to establish the online security and privacy habits of the end users. Some of the instruments given are specifically designed for the employees, while others tap onto a specific device and do not capture the holistic measures of cyber space threats. Table 1 shows the scale used and the studies.

4 Development of the Questionnaire

With the aim of developing an instrument that holistically captures the cyber security habits/practices of students, this study presents the instrument and its development method. The development process consisted of six steps. As step one, the literature was

Table 1. Previous instruments

Study	Scale name	References
Study 1	UISAQ	[37]
Study 2, study 3	SeBIS	[38, 39]
Study 4, study 5, study 6	(HAIS-Q)	[40–42]
Study 7	USP	[43]
Study 8	—	[4]
Study 9	Risky behavior scale	[44]
Study 10, study 11	Identity information disclosure scale	[45, 46]
Study 12	Security behavior scale	[47]
Study 13, study 14, study 15	Mobile security scale	[48–50]
Study 16	Privacy attitude scale	[51]

thoroughly searched for cyber security and privacy scales which have already been developed. The studies were thoroughly analyzed for relevant scales in step 2. The constructs of each scale were identified in step 3. It should be noted that there were studies that made use of previous scales and modified or adopted them, we incorporated those scales and constructs too in our instrument design process. Table 2 shows the details of the studies and the scales that were analyzed in detail. It should be noted that emphasis was made on how the study developed the scale and the carried out its

Table 2. Studies and scales analyzed for instrument development

Scales used/name	Scale making method	Validation/reliability
Risky behavior scale [10]	Experts interviews and based on related studies [53–57]	Cronbach's alpha is 0.934
Risky behavior scale [44]	Based on another research, reviewed by 3 experts	Cronbach's alpha 0.95, factor analysis
Personal information sharing awareness	Each scale was developed using previously validated research scales [58–60]	Cronbach's alpha for each set of construct item (0.891, 0.913, 0.911 for each scale) Internal validity of the content was done by literature review and by expert panels
Personal information sharing practices Personal information sharing habits [46]		
Security attitude scale Security behaviors Use of computer security tools Wireless security Data privacy [61]	Based on literature review CIA triad, pilot tested	Cronbach's alpha for whole scale is 0.69 Exploratory factor analysis
Users' potentially risky behaviour Users' awareness [62]	Security guidelines and reports	Cronbach's alpha of each item is calculated Factor analysis
Information security awareness level [4]	Review from the literature NIST SP 800-50 report	Cronbach's alpha = 0.779, principal component analysis

(continued)

Table 2. (*continued*)

Privacy attitude scale Privacy behaviour scale [51]	Based on existing literature review [63–65]	Cronbach's alpha 0.75 and 0.74 respectively
Information revelation, Concern for internet privacy, Concern about unwanted audiences, Profile visibility Privacy protection strategies [66]	Based on already presented scale [67, 68, 16]	No validation
Privacy behaviour scale Privacy concern scale Privacy attitude scale Identity information disclosure scale [60]	Based on already presented. Scale [51, 69, 45]	Cronbach's alpha is 0.80, 0.92, 0.94 and 0.82 respectively
Security related behaviour in specific areas [47]	Experts interviews	
UISAQ Potential risky behaviour Level of information security awareness Level of users' belief about IS Quality and security of passwords [70]	Made form security guidelines and reports	Cronbach's alpha of each item is calculated Factor analysis
Risk in practices, Countermeasures in practices Awareness of risks Awareness of countermeasures [71]	–	No validation
Mobile security scale [48–50]	Based on a review of extant literature of most frequently recommended security practices and guidelines from the Internet crime complaint center	Survey was piloted tested with 37 students

validation and reliability. For each scale the cronbach's alpha as a measure of reliability and the consequent Factor Analysis (FA) and Principle Component Analysis (PCA) were taken into account for its inclusion in our instrument.

The duplicate items and scales were removed in step 4 to form the instrument. In step 5, the instrument was given to experts for carrying out the face validity. The feedback was recorded and was analyzed and hence the face validity of the instrument was carried out. The experts gave their remarks which belonged to either of the following categories:

- Grammatical Mistakes—The questions had some English mistakes that included grammatical mistakes
- Multiple Interpretations—There were questions which had dual meaning
- Ambiguous and Needs Clarity—Some statements meaning was not clear and further explanation was required
- Needs Rewording—Experts gave feedback in terms of rewording some questions to better state them
- Lack of Knowledge—The questions which needed more explanation since they believed the participants may not have knowledge of particular questions posed
- Rearrangement of Questions—Social media questions needed to incorporate other questions to cater for general and specific social media.

In step 6, pilot was carried out with a small sample of students from a higher education institute to see if they understand the questionnaire. Forty three students took part in the pilot. The questionnaire was given in a paper format. It was explained to the participants, the objectives of the questionnaire and their feedback would be very beneficial in making the questionnaire readable and understandable to the target population.

Participants were instructed to ask about the statements they don't understand. The verbal feedback was noted down and their reservations were recorded for future amendments. Moreover, the participants were also instructed to annotate the questions that they don't understand and leave comments on each questions in order to leave their feedback. The participant's responses were properly analyzed and problems were found in the following five categories.

- Jargon—Participants did not understand the words and technical jargon used in the questionnaire
- Meaning—Some question statements were ambiguous and their meaning was not clear
- Lack of Knowledge—Participants did not know about some of the things/phenomenon related with the security practices
- Questions which seem to have same meaning—Few questions seemed like they have the same meaning
- Questions which need rewording—Some questions were identified by few students which could be reworded to better capture the true meaning of them.

After piloting the instrument, the amendments were made in the questionnaire. As per students' feedback, each and every question that they pointed out in terms of its jargon, meaning, lack of knowledge was explained more in detail. Furthermore, those questions were carefully reworded to convey the exact meaning of statement and hence any misinterpretation errors were weeded out. On the other hand, the feedback provided by the experts was analyzed comprehensively and actions were taken in response to their suggestions. According to the feedback, the questions were reworded while one question was removed due to its non-applicability. The social media related questions were re-arranged and four questions were added.

The final instrument consisted of six subscales i.e. social media security and privacy, security behavior scale, risky behavior scale, mobile device security, privacy

concern scale and level of information security awareness. The social media security and privacy subscale comprises of 18 items which are multiple choice and dichotomous in nature. The security behavior scale consists of 28 items (likert and dichotomous), risky behavior scale consists of 5 items, while the mobile security scale and privacy concern scale have 12 (dichotomous) and 14 (likert) items respectively. The level of information security awareness consists of 4 items (likert). Therefore the final instrument comprised of 81 items in total. The developed instrument can be found on https://drive.google.com/file/d/1gKFpl_yEAFdBNbtRlc7tjK0ZJ9n96Veq/view?usp=sharing.

The next stages of the development phase are yet to be carried out. The initial pilot work and item deletion and addition is followed by Factor Analysis that has to be administered to a sufficient size in the future. A rule of thumb is to have at least 5 respondents per item [52]. The reliability of the instrument is to be measured by Cronbach's alpha once the questionnaire is administered to a suitable sample size. Moreover appropriate validity measure apart from face validity is also to be carried out such as Factor Analysis and confirmation on an independent data set.

5 Conclusion and Future Work

This study presents the development of an instrument that is used to gauge the security and privacy habits and practices of end users especially students. Although, literature presents studies in the same vein as ours, the instruments are not comprehensive enough to tap onto different aspects of cyber space security. This instrument has tried to fill in the gap. Researchers can use this instrument to find the level of security and privacy practices of students. Moreover, it can be administered in developing countries where levels of security practices can be ascertained amid lack of resources and diverse social-culture environments.

This study is an ongoing work and still work needs to be done. The reliability and validity of the instrument will be carried out by administrating the instrument to the student's sample of considerable size. Moreover, we plan to exercise and ascertain the security and privacy habits of students as a country wide study in higher education institutes. The results of the study would be beneficial in designing a tailored yet comprehensive awareness program. Moreover, the instrument in this study can help to establish the effectiveness of the program.

References

- Thompson, H.: The human element of information security. *IEEE Secur. Priv.* **11**(1), 32–35 (2013)
- Allen, G.: Hitting the ground running. *Security* **48**(12), 44–45 (2011)
- Wilson, M., Hash, J.: Building an information technology security awareness and training program. *NIST Spec. Publ.* **800**(50), 1–39 (2003)
- Kim, E.B.: Recommendations for information security awareness training for college students. *Inf. Manag. Comput. Secur.* **22**(1), 115–126 (2014)

5. Statista: All products require an annual contract prices do not include sales tax, "Global digital population 2018|Statistic," Statista (Online). Available: <https://www.statista.com/statistics/617136/digital-population-worldwide/>. Last accessed 09 June 2018
6. Evans, D.: How the Next Evolution of the Internet Is Changing Everything, p. 11 (2011)
7. Business of Apps.: Facebook revenue and usage statistics (2018). Business of Apps. (Online). Available: <http://www.businessofapps.com/data/facebook-statistics/>. Last accessed 09 June 2018
8. Statista: All products require an annual contract prices do not include sales tax, "Leading global social networks 2018|Statistic," Statista (Online). Available: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>. Last accessed 09 June 2018
9. Aslam, S.: • YouTube by the numbers (2018): stats, demographics & fun facts, 05 Feb 2018
10. Öğütçü, G., Testik, Ö.M., Chouseinoglou, O.: Analysis of personal information security behavior and awareness. *Comput. Secur.* **56**, 83–93 (2016)
11. Pahnila, S., Siponen, M., Mahmood, A.: Employees' behavior towards IS security policy compliance. In: 40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007, pp. 156b-156b (2007)
12. Stanton, J., Mastrangelo, P., Stam, K., Jolton, J.: Behavioral information security: two end user survey studies of motivation and security practices. In: AMCIS 2004 Proceedings, p. 175 (2004)
13. Schultz, E.E., Proctor, R.W., Lien, M.-C., Salvendy, G.: Usability and security an appraisal of usability issues in information security methods. *Comput. Secur.* **20**(7), 620–634 (2001)
14. Trček, D., Trobec, R., Pavešić, N., Tasić, J.F.: Information systems security and human behaviour. *Behav. Inf. Technol.* **26**(2), 113–118 (2007)
15. Mylonas, A., Kastania, A., Gritzalis, D.: Delegate the smartphone user? Security awareness in smartphone platforms. *Comput. Secur.* **34**, 47–66 (2013)
16. Al-Saggaf, Y.: An exploratory study of attitudes towards privacy in social media and the threat of blackmail: the views of a group of Saudi women. *Electron. J. Inf. Syst. Dev. Ctries.* **75** (2016)
17. Mills, J.L.: Privacy: the Lost Right. Oxford University Press (2008)
18. Davidson, M.A.: Leading by example: the case for IT security in academia. *Educ. Rev.* **40**(1) (2005)
19. Chandarman, R., Van Niekerk, B.: Students' Cybersecurity Awareness at a Private Tertiary Educational Institution (2017)
20. Pramod, D., Raman, R.: A Study on the User Perception and Awareness of Smartphone Security (2014)
21. Aliyu, M., Abdallah, N.A., Lasisi, N.A., Diyar, D., Zeki, A.M.: Computer security and ethics awareness among IIUM students: an empirical study. In: 2010 International Conference on Information and Communication Technology for the Muslim World (ICT4M), pp. A52–A56 (2010)
22. Joinson, A.N., Reips, U.-D., Buchanan, T., Schofield, C.B.P.: Privacy, trust, and self-disclosure online. *Hum. Comput. Interact.* **25**(1), 1–24 (2010)
23. Von Solms, B., Von Solms, R.: The 10 deadly sins of information security management. *Comput. Secur.* **23**(5), 371–376 (2004)
24. Wilson, M., de Zafra, D.E., Pitcher, S.I., Tressler, J.D., Ippolito, J.B.: Information Technology Security Training Requirements: a Role- and Performance-Based Model. National Inst of Standards and Technology Gaithersburg MD Computer Security Div (1998)
25. Lunt, B.M., et al.: Curriculum Guidelines for Undergraduate Degree Programs in Information Technology, vol. 2, no. 2009. Retrieved Mar 2008
26. Susan Hansche, C.: Designing a Security Awareness Program: Part 1 (2001)

27. Rahim, N.H.A., Hamid, S., Mat Kiah, M.L., Shamshirband, S., Furnell, S.: A systematic review of approaches to assessing cybersecurity awareness. *Kybernetes* **44**(4), 606–622 (2015)
28. Al-Daeef, M.M., Basir, N., Saudi, M.M.: Security awareness training: a review. *Proc. World Congr. Eng.* **1**, 5–7 (2017)
29. Haeussinger, F., Kranz, J.: Antecedents of employees' information security awareness—review, synthesis, and directions for future research. In: Proceedings of the 25th European Conference on Information Systems (ECIS) (2017)
30. Fung, C.C., Khera, V., Depickere, A., Tantatsanawong, P., Boonbrahm, P.: Raising information security awareness in digital ecosystem with games—a pilot study in Thailand. In: 2nd IEEE International Conference on Digital Ecosystems and Technologies, 2008. DEST 2008, pp. 375–380 (2008)
31. Rahim, M.M., Cheo, A., Cheong, K.: IT security expert's presentation and attitude changes of end-users towards IT security aware behaviour: a pilot study. In: ACIS 2008 Proceedings, p. 33 (2008)
32. Kruger, H., Drevin, L., Steyn, T.: A vocabulary test to assess information security awareness. *Inf. Manag. Comput. Secur.* **18**(5), 316–327 (2010)
33. Kruger, H., Drevin, L., Steyn, T.: Email security awareness—a practical assessment of employee behaviour. In: Fifth World Conference on Information Security Education, pp. 33–40 (2007)
34. Hellqvist, F., Ibrahim, S., Jatko, R., Andersson, A., Hedström, K.: Getting their hands stuck in the cookie jar-students' security awareness in 1:1 laptop schools. *Int. J. Public Inf. Syst.* **9**(1) (2013)
35. Kam, H.-J., Katerattanakul, P.: Out-of-Class Learning: a Pedagogical Approach of Promoting Information Security Education (2014)
36. Dodge Jr., R.C., Carver, C., Ferguson, A.J.: Phishing for user security awareness. *Comput. Secur.* **26**(1), 73–80 (2007)
37. Solic, K., Velki, T., Galba, T.: Empirical study on ICT system's users' risky behavior and security awareness. In: 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1356–1359 (2015)
38. Egelman, S., Peer, E.: Scaling the security wall: developing a security behavior intentions scale (sebis). In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 2873–2882 (2015)
39. Egelman, S., Harbach, M., Peer, E.: Behavior ever follows intention. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '16, pp. 1–5 (2016)
40. Parsons, K., McCormac, A., Butavicius, M., Pattinson, M., Jerram, C.: Determining employee awareness using the human aspects of information security questionnaire (HAIS-Q). *Comput. Secur.* **42**, 165–176 (2014)
41. Pattinson, M., Parsons, K., Butavicius, M., McCormac, A., Calic, D.: Assessing information security attitudes: a comparison of two studies. *Inf. Comput. Secur.* **24**(2), 228–240 (2016)
42. Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A., Zwaans, T.: The human aspects of information security questionnaire (HAIS-Q): two further validation studies. *Comput. Secur.* **66**, 40–51 (2017)
43. Crossler, R., Bélanger, F.: An extended perspective on individual security behaviors: protection motivation theory and a unified security practices (USP) instrument. *ACM SIGMIS Database DATABASE Adv. Inf. Syst.* **45**(4), 51–71 (2014)
44. Gökçearslan, S., Seferoglu, S.S.: The use of the internet among middle school students: risky behaviors and opportunities. *Kastamonu Educ. J.* **24**(1), 383–404 (2016)
45. Stutzman, F.: An evaluation of identity-sharing behavior in social network communities. *Int. Digit. Media Arts J.* **3**(1), 10–18 (2006)

46. Ball, A.L., Ramim, M.M., Levy, Y.: Examining users' personal information sharing awareness, habits, and practices in social networking sites and e-learning systems. *Online J. Appl. Knowl. Manag.* **3**(1), 180–207 (2015)
47. Ion, I., Reeder, R., Consolvo, S.: "... No one can hack my mind": comparing expert and non-expert security practices. In: *Proceedings of SOUPS* (2015)
48. Jones, B.H., Chin, A.G.: On the efficacy of smartphone security: a critical analysis of modifications in business students' practices over time. *Int. J. Inf. Manag.* **35**(5), 561–571 (2015)
49. Jones, B.H., Heinrichs, L.R.: Do business students practice smartphone security? *J. Comput. Inf. Syst.* **53**(2), 22–30 (2012)
50. Jones, B.H., Chin, A.G., Aiken, P.: Risky business: students and smartphones. *TechTrends* **58**(6), 73–83 (2014)
51. Buchanan, T., Paine, C., Joinson, A.N., Reips, U.-D.: Development of measures of online privacy concern and protection for use on the Internet. *J. Am. Soc. Inf. Sci. Technol.* **58**(2), 157–165 (2007)
52. Bryman, A., Cramer, D.: Quantitative Data Analysis with SPSS Release 10 for Windows: a Guide for Social Scientists. Routledge (2002)
53. Aytes, K., Connolly, T.: Computer security and risky computing practices: a rational choice perspective. *J. Organ. End User Comput. JOEUC* **16**(3), 22–40 (2004)
54. Stanton, J.M., Stam, K.R., Mastrangelo, P., Jolton, J.: Analysis of end user security behaviors. *Comput. Secur.* **24**(2), 124–133 (2005)
55. Milne, G.R., Labrecque, L.I., Cromer, C.: Toward an understanding of the online consumer's risky behavior and protection practices. *J. Consum. Aff.* **43**(3), 449–473 (2009)
56. Ng, B.-Y., Kankanhalli, A., Xu, Y.C.: Studying users' computer security behavior: a health belief perspective. *Decis. Support Syst.* **46**(4), 815–825 (2009)
57. Tsohou, A., Karyda, M., Kokolakis, S., Kiountouzis, E.: Formulating information systems risk management strategies through cultural theory. *Inf. Manag. Comput. Secur.* **14**(3), 198–217 (2006)
58. Oceja, L., Ambrona, T., López-Pérez, B., Salgado, S., Villegas, M.: When the victim is one among others: empathy, awareness of others and motivational ambivalence. *Motiv. Emot.* **34**(2), 110–119 (2010)
59. Verplanken, B., Orbell, S.: Reflections on past behavior: a self-report index of habit strength¹. *J. Appl. Soc. Psychol.* **33**(6), 1313–1330 (2003)
60. Fogel, J., Nehmad, E.: Internet social network communities: risk taking, trust, and privacy concerns. *Comput. Hum. Behav.* **25**(1), 153–160 (2009)
61. Mensch, S., Wilkie, L.: Information security activities of college students: an exploratory study. *J. Manag. Inf. Decis. Sci.* **14**(2), 91 (2011)
62. Galba, T., Solic, K., Lukic, I.: An information security and privacy self-assessment (ISPSA) tool for internet users. *Acta Polytech. Hung.* **12**(7), 149–162 (2015)
63. Burgoon, J.K., Parrott, R., Le Poire, B.A., Kelley, D.L., Walther, J.B., Perry, D.: Maintaining and restoring privacy through communication in different types of relationships. *J. Soc. Pers. Relat.* **6**(2), 131–158 (1989)
64. DeCew, J.W.: *In Pursuit of Privacy: Law, Ethics, and the Rise of Technology*. Cornell University Press (1997)
65. Fox, S., Rainie, L., Horrigan, J., Lenhart, A., Spooner, T., Carter, C.: Trust and Privacy Online: Why Americans Want to Rewrite the Rules. Pew Internet Am. Life Proj., pp. 1–29 (2000)
66. Young, A.L., Quan-Haase, A.: Information revelation and internet privacy concerns on social network sites: a case study of Facebook. In: *Proceedings of the Fourth International Conference on Communities and Technologies*, pp. 265–274 (2009)

67. Govani, T., Pashley, H.: Student awareness of the privacy implications when using Facebook. Unpublished paper present. "Privacy Poster Fair" Carnegie Mellon Univ. Sch. Libr. Inf. Sci., vol. 9, pp. 1–17 (2005)
68. Tufekci, Z.: Can you see me now? Audience and disclosure regulation in online social network sites. *Bull. Sci. Technol. Soc.* **28**(1), 20–36 (2008)
69. Dinev, T., Hart, P.: Internet privacy concerns and their antecedents-measurement validity and a regression model. *Behav. Inf. Technol.* **23**(6), 413–422 (2004)
70. Velki, T., Solic, K., Ocevcic, H.: Development of users' information security awareness questionnaire (UISAQ)—ongoing work. In: 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1417–1421 (2014)
71. Slusky, L., Partow-Navid, P.: Students information security practices and awareness. *J. Inf. Priv. Secur.* **8**(4), 3–26 (2012)



Privacy and Security—Limits of Personal Information to Minimize Loss of Privacy

Hanif Ur Rahman^{1(✉)}, Ateeq Ur Rehman², Shah Nazir³,
Izaz Ur Rehman², and Nizam Uddin¹

¹ Department of Computer and System Sciences,
Stockholm University, Stockholm, Sweden
hanif.maidan@gmail.com

² Department of Computer Science, Abdul Wali Khan University Mardan,
Mardan, Pakistan

³ Department of Computer Science, University of Swabi, Swabi, Pakistan

Abstract. The current information age enabled the direct access to personal information that rapidly invades private space. With the advent of new information technology's equipment and innovations the time and distance limitations of communication got minimized that leads to more interaction among IT users. Thus with the expansion of information society the privacy violation becomes prominent as well as privacy dilemma got increased. IT users think that personal information is used for other purposes rather than security. The things are getting worse when the organization outsources data processing to other suppliers and sometime share information with third parties. All these steps are rapidly increasing users' attention towards their personal information. The aim of this research is to investigate the concerns of IT users about their personal information and to find out that how much personal information users would like to provide to government, secured organizations, private organizations' websites and social media. Data has been collected through a questionnaire from the students of Department of Computer and System Sciences (DSV), Stockholm University, Sweden and Royal Institute of Technology (KTH), Stockholm Sweden. The findings of the research study show that all the participants had deep knowledge about privacy. They were highly concerned about sharing their personal information with third parties. More importantly, respondents refused to provide complete personal information to nonprofit organizations and social media. On the other hand, the survey result reflects that people provide their complete personal information to organizations such as Swedish Migrationsverket, Skatteverket, and to DSV, KTH and Swedish banks. The IT users believed that these systems are quiet secure for their personal information and they will not use their information for other purposes.

Keywords: Privacy · Security · Privacy loss · Personal information

1 Introduction

Privacy has been a critical issue from the last few decades and especially it become very prominent for modern day technology and industry. Privacy is discarded at the cost of issues like national security and effective administration of corporations. The people in the developed countries and especially in the society of emerging information are explicitly or implicitly concerned over the increasing loss of their privacy. State corporations and private organizations are collecting personal information of customers and citizens in order to serve them effectively, but it is not confirmed that whether the collected information will be used for the said propose (for which information is collected) or it is just a way of more social control, where actual beneficiary is state and corporations. The personal information is used by organizations for gaining strategic advantages but recently very little efforts have been directed towards privacy issue [1]. New scientific innovations in data collection and management introduced new forms of control over personal information such as computer matching [2] and Dataveillance [3] which make easy the collection of personal information. Computer matching provides a method of conducting investigations of fraud, abuse and waste of government funds by screening all the records of federal employees. Likewise, Dataveillance monitors users' communication such as credit card transactions, emails and social networks and collect online data [4]. These technology capabilities such as Dataveillance and computer matching are making it much easier to collect personal information. Similarly, the merging of disparate bodies of information became possible due to the advancement in telecommunication systems and query languages. Furthermore, the technologies such as caller ID and automatic number identification can be used for the purpose to remove anonymity from communication. These technologies reducing time and distance limitations and bring rapidly increase in communication [5]. On the other hand, as the information society getting expansion the possibility to lose privacy getting increase and privacy dilemma occurs more severely.

Today people are more careful in providing their personal information to various government organizations, nonprofit organizations, private corporations and social media as there is an increase sense of privacy loss in them. Likewise, privacy problem will be larger in the future than today.

The rest of the paper is organized as follows; Sect. 2 shows the theoretical extensive background of the study. Section 3 shows the research method. In Sect. 4 the results and analysis are presented. Section 5 shows the details of discussions and the paper concludes at Sect. 6.

1.1 Description of the Research Area

Security community and state corporations are paying much attention to security, while privacy is focused a bit lightly. Individuals' activities are being monitored tremendously through advance technology both online and offline. Due to the rapid development of Internet, people's personal information can be accessed easily in a network environment that increases the possibility of privacy loss. Hence, network environment is becoming users' security concern [6]. A survey has been conducted of IT users to investigate privacy and security concerns which show that the majority of users were

not satisfied from the government laws and measurements taken for privacy [7]. Governments, financial institutions, hospitals and private businesses collect confidential information about their employees and customers. Most of this information is stored in computer and transmitted to other locations and computers across networks. All activities are interconnected through Internet and these collaborative activities need discloser of individual information (partially or fully identity revelation) and therefore it can be said that security partially intersects privacy. Some people believe that there is a trade-off between security and privacy. One can be achieved on the expense of other. However, this statement cannot be accepted as there are examples like arming pilots, reinforcing cockpit doors are example of security measures that have no effect on privacy [5, 8]. It can easily be refuted that security can be achieved at the cost of privacy. If privacy is taken away due to security measurements then it is the failure of system designer and bad security planning. If security is planned in beginning then it will not enforce individual to give up their privacy [5, 9].

Although there are regulations and laws to protect personal information but unfortunately the invasion of privacy was found inadvertently supported by technologies which grows the attention of users about their personal information [10]. Organizations are collecting personal information for specific purpose where the users believe that organizations ask for much personal information than the required. This situation is getting worse when the collected personal information is shared with third party which is a deviation from the specific use of personal information. This remarkably increases the feeling of users for losing their personal information [5, 11].

Privacy and security is an important issue and need to be investigated from users' perspective in order to find out how much personal information should be provided to various information systems. The provided personal information to various organizations, social media and for security implementation could be used to identify individuals. Therefore, the provided personal information causes the revelation of personal identity which grows the attention of users about the loss of their privacy. The purpose of this research is to investigate the concerns of IT users about their personal information. The users' minds will be studied and it will be analyzed that how much personal information users would like to provide to state corporations, private organizations websites and social media. More specifically, to what level, personal information should be provided with users' perspective to minimize the loss of privacy.

2 Theoretical Extensive Background

The following sub-sections describe the theoretical extensive background of the present research.

2.1 Privacy

There has been debate on privacy in late 1960s which did not come to a single conclusion and today still there is no collectively established definition of privacy [1]. There are a lot of controversies about the concept of privacy and every definition proposed by philosophers there can be found alike counterexample [1]. Privacy has

been recognized by the Council of Europe [10] as fundamental right in the 1950 which states that everyone has the right to respect for his private and family life, his home and his correspondence. The European countries introduced the privacy regulations laws much later although the right to privacy was vested in 1950. The need to harmonize information protection laws in Europe has been recognized by the European Union for the purpose of protection of citizen's privacy and personal information. Nowadays, within many countries there are rules and regulations for handling, collecting and for the use of personal information. The main purpose of these rules is to make sure the security of personal information which are using in different organizations and to protect loss of privacy [10]. According to Gavison [12] privacy loss occurs whenever, information about an individual is obtained and gained access to him. Privacy loss will be a bigger problem in the future than it is today and it is realized that there is a remarkable increase in the society for the loss of privacy [5].

As stated above privacy has been defined from various perspectives and there is no single accepted definition however, privacy definitions have been grouped into three distinct categories such as privacy as no access to person or the personal realm, privacy as control over personal information and privacy as freedom from judgment or scrutiny by others [1].

2.1.1 Privacy as no Access to Person or Personal Realm

According to Warren and Brandeis [13] privacy is "the right to be let alone". As there are some institutions, organizations and individuals who have the rights to access you such as tax service or your creditors therefore, it is a very limited definition. Haag [14] defined privacy as "privacy is the exclusive access of a person (or other legal entity) to a realm of his own". The right to privacy enables a person to avoid other people to watch, use and invade his personal information. The ambiguity here is the definition of private or personal realm. For example in some cultures bare torso are considered quiet personal while in some of the African tribes they are in the public domain. Gross [15] expressed his view as "the condition of human life in which acquaintance with a person or with affairs with his life which are personal to him limited". However, these definition do not make anybody clear to differentiate between the loss of privacy and whether somebody privacy has been violated or not because an individual can give access to others for accessing his personal realm. In such case his privacy has not been violated but it can be said that he is less private [1].

2.1.2 Privacy as Control Over Personal Information

Privacy is defined by Fried [16] "control over information or knowledge about one-self". Westin [17] also defined privacy on the same way "the claim of individuals, groups or institutions to determine for themselves when, how and to what extent information about them is communicated to others" or more generally by Parker [18] as "control over when and whom the various parts of us can be sensed by others". These ideas are explained by an example if anybody listens to or tap conversation. If they have installed coded system so he will not get anything from call listing, in this way they have control on their personal information but he might has violated their right to privacy. In this group of definitions privacy need is assumed implicitly [1].

2.1.3 Privacy as Freedom from Judgment or Scrutiny by Others

Johnson [19] discussed that real issue of privacy is the judgment by others and he expressed this as follows:

“Privacy is considered a conventional concept because the term private is defined socially and culturally that varies according to the context and society. In every culture the meaning of private is different and hardly found something that is considered private in every culture. Nevertheless, all example of privacy have a single common feature. They are aspects of a person’s life which are culturally, recognized as being immune from the judgment of others” [1].

The issue, judgment by others has been discussed by DeCew [20] as “an interest in privacy is at stake when intrusion by others is not legitimate because it jeopardizes or prohibits protection of a realm free from scrutiny, judgment and the pressure, distress or losses they can cause”.

Hence, it became clear from all the above discussion that there is no common and simple definition of privacy. However, the concept of privacy can be summarized from certain aspects as follows [1]:

- It is a relational concept.
- It is directed towards the personal domain.
- Actually to claim for the privacy is to claim for the right to control access to personal domain.
- For controlling access to personal realm, the best way is to control the distribution of information.
- For claiming privacy, should be claimed the right to private domain of immunity for protecting from the judgment of others.
- It is a relative concept and a matter of judgment.

The above discussion was about the privacy i.e. a brief discussion and overview of privacy and some important definitions of privacy.

2.2 Security

Security is related to computer hardware and software while privacy is individuals’ property. Any system which does not protect individuals’ privacy could be secure theoretically, but it is not wisdom to set up such as system in real world. The terms computer security and information security are often interchangeably used incorrectly. The confidentiality, integrity and availability of data are concerned with information security. Computer security makes sure the availability and to operate the computer system correctly regardless the information stored or processed [21]. The Institute for Security and Open Methodologies (ISECOM) define security as “a form of protection where a separation is created between the assets and threat” [21]. Security can be defined as the protection against any danger, loss, crime and damage whereas, information security is define as the protection of information system and information from unauthorized use, illegal access, disclosure, recording and changing of information and its destruction [21].

2.2.1 Personal Information

Personal information can be defined as “those facts or opinions that relevant to an individual and which it would be reasonable to expect him to regard as intimate or confidential and therefore to want to withhold or at least to restrict their circulation” [1]. The collection and processing of personal information should be according to the consent of the users. The personal information must be used for the legitimate purposes for which information has been collected. The service providers and corporations should not process information in ways which are not compatible with the specific purpose and especially the sharing of personal information with third parties must be restricted and the information should disclose on certain conditions. When IT system is designed the experts should make sure that personal information is being handled properly, according to the laws and regulations. As, nowadays there are rules and regulations for collecting and handling personal information in order to protect the loss of personal information [10].

In the following section the highest societal expectations about their privacy will be explained in the important areas like improper collection of personal information, new use of information, internal access and sharing of information [5].

2.2.2 Improper Information Collection

According to the Association for Computing Machinery (ACM), all their members should consider the privacy rules and try to collect minimum personal information. Personal information used by government corporation or agency and private organization for specific purpose does not pose a problem as information has been collected primarily for that purpose. Such use of information is reasonable and information can be used as input in decision process. On the other hand, the ways in which information is collected can be viewed unreasonable such as the collection of information by deceptive and secretive techniques [5].

2.2.3 Information Used for a New Purpose

Information which is held by an organization used for the other purpose than it was collected originally from users. It is another area of concern where policy attention needed. The use of such information is unreasonable even when the information handled carefully without any error and this area refers to new use of information [5].

2.2.4 Internal Access

All the users have a great concern about the improper access to computerized information. Who is allowed to access the personal information inside the organization? The users should have knowledge before to grant access to their personal information. ACM mentions that the access to the personal information should be limited as it is the responsibility of organization to protect individual’s privacy [5].

2.2.5 Information Shared in New Ways Across Entities

This is another area of concern where organizations share information in new ways. By sharing information through new ways not only concerns are becoming increasingly high for the unintended use of personal information but also the diffuse of

responsibilities problem arises. This area has received much attention from policy makers and there are extensive policies for sharing name, address and telephone information with third parties [5].

3 Research Method

In this research data has been collected online through a questionnaire from the students of Department of Computer and System Sciences (DSV), Stockholm University, Sweden and the Royal Institute of Technology (KTH), Stockholm Sweden. The students of DSV and KTH were surveyed only due to the limited time for this research and availability of their emails so they could respond quickly. The questionnaire for the research is based on the literature and scoping study. All the questions were made simple, easy and understandable for the purpose of avoiding ambiguity in giving answers to the questions. Before distributing the questionnaire it was tested on 4–5 students in order to revise the questions if needed for removing the ambiguity.

The conclusion of this research has been drawn after analysing the collected data. The limitations of the study are that the proposed study was conducted for Swedish Migrationsverket, Skatteverket, and to DSV, KTH and Swedish banks, and was not verified from the other organizations.

4 Results and Analysis

This section presents the result of the survey that conducted in the students of DSV and KTH, and its detailed analysis. According to the result question 2 and 3 which are about privacy and security shows that most of the people (18 out of 36) think that privacy and security are not opposite to each other and similarly, (26 out of 36) answered both are equally important. In response to questions 1 the respondent (35), having concerns about their privacy and in question 4 the majority of individuals (27), have expressed a sense of fear that more personal information would breach their privacy and only 4 individuals replied that more personal information does not affect privacy.

In question 10 it has been found that 12 participants revealed the use of 20% fake information and 10 participants 50% respectively, which show the tendency that the people do not want to provide much personal information.

In response to question 6, out of the total 26 individuals want to provide their personal information to the mentioned systems. Although the mentioned systems are secured but still 10 of the individuals do not want to provide personal information. In the question 7 which is related to question 6, majority of the respondent opted as secure, some thought as this was only for requirements and only 2 individuals checked both.

The answer to question 8 is, as it was expected because 32 individuals feared to provide personal information and only 4 individuals did not bother to provide personal information. Question 9 was included for the in-depth insights to question 8 which shows that 14 individuals answered that these service providers should not get more

personal information beside that 11 individual think these are insecure and the remaining 10 participants answered both.

In response to question 11, most of the participants (28) thought that personal information should not be shared with third parties, and only 2 individuals said yes. All the eleven questions of the questionnaire and their relevant obtained results are given diagrammatically in Fig. 1.

In the following section the analysis has been carried out in detailed in three parts i.e. 4.1, 4.2 and 4.3. The first part of the analysis consists of first five and the last two questions which are about the general concept of privacy, security and personal information. The second part of analysis comprised question 8, 9, 10 and 11 which are about providing personal information to nonprofit organizations websites, private organizations and social media while in part three, question 6 and 7 where users were asked about providing their information to Migrationsverket (Swedish migration board), Skatteverket (Swedish tax office), DSV, KTH and Swedish banks.

4.1 General Concept About Personal Information and Privacy

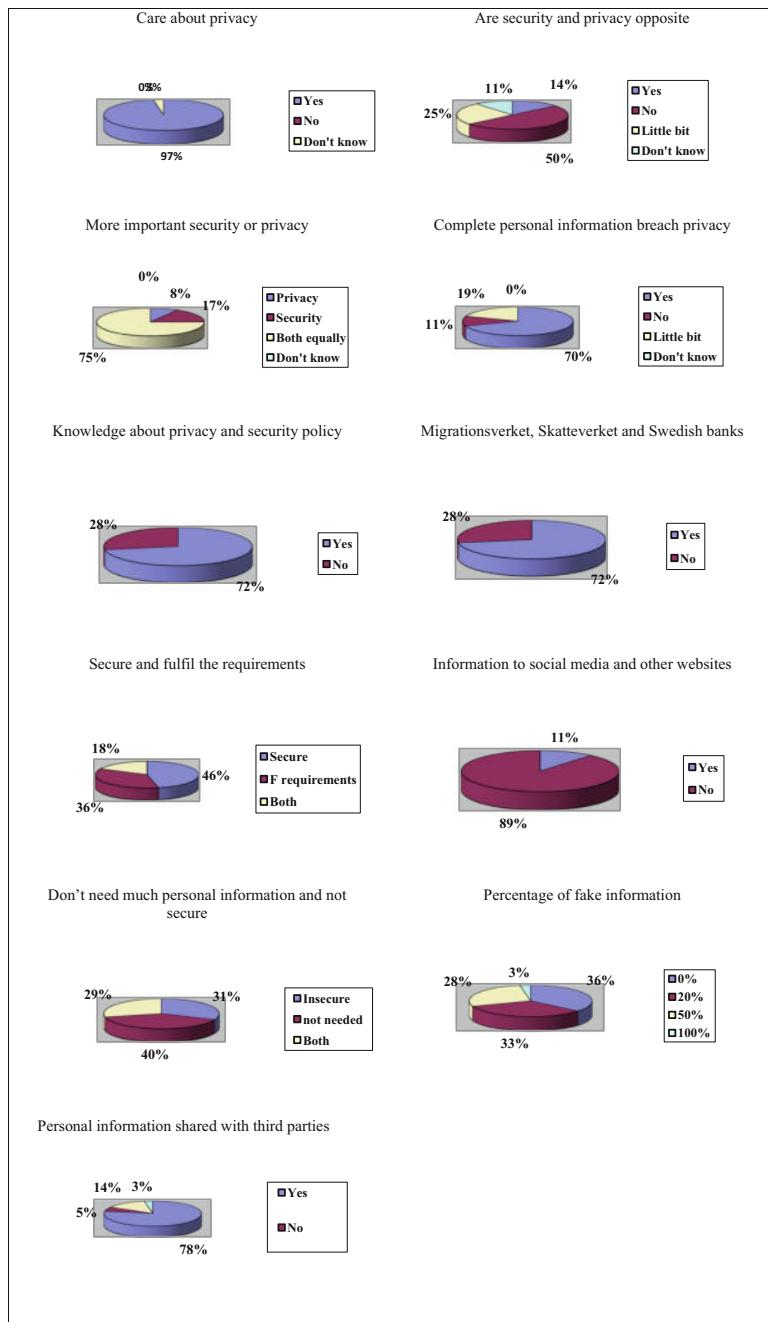
It has been realized from the results that the surveyed people have a deep knowledge about privacy, security and highly concerned while providing their personal information to government corporations, private corporations, and social media. In response to question 1 only one individual answered do not know, means does not care about personal information, he might not has knowledge about privacy. Similarly, most of the respondents to question 2 and 3 replied that privacy and security are not opposite to each other and both have equal importance.

4.2 Personal Information to Insecure Systems

The important point noted in question 8, is that the people have hesitation in providing their personal information to insecure systems such as nonprofit organizations websites and social media. The reason is that the provided information might be leaked or hacked that causes the privacy loss. To investigate this issue more thoroughly, when asked for the reason in question 9, majority of them replied service providers should not collect much personal information and also participants expressed that the social media and nonprofit websites are not secure enough to protect their personal information. Even some people thought that these systems are neither secure nor needed. Answer to the questions 10 and 11 has the same outcome, people do not want to provide much personal information, because they think that provided information would be shared with third parties and this is one of the reasons that some people provide fake information. This also shows that people do not trust on these insecure systems and social media regarding their privacy.

4.3 Personal Information to Secure Systems

On the other hand, it has been found from the survey that people's response was positive about question 6 and 7, although personal information included in the mentioned questions were fingerprints and personnumers (personal number) but still people

**Fig. 1.** Users' responses

are in the favour in providing all the sensitive and complete data. Furthermore, in question 7 the preceding question has been investigated deeply, when asked what is the reason of providing your complete information. The majority of respondents replied that these systems are secure. Therefore, it can be concluded that the systems such as Migrationsverket (Swedish migration board), Skatteverket (Swedish tax office), University and banks are secure and the people are confident that their provided personal information will not be accessed illegally and will not be used for other purposes. Hence people have no fear about the loss of their privacy.

5 Discussion

It is indicated by the study findings that some people provide fake information to social media and nonprofit organization websites because the users think that much information is collected on the name of security, in addition to this argument, individuals feel insecurity for their provided personal information. In order to minimize this gap among users and service providers, it is important to read the minds of users and to get feedback from them. This step will reduce the trust deficit among the users and will motivate the individuals to provide their personal information correctly.

6 Conclusion and Future Research

The purpose of this study is to investigate the level of personal information that should be provided to government organizations, nonprofit corporations, private organizations websites and social media. The findings of the research study show that all the participants had deep knowledge of privacy and security. The majority of respondents refused to provide complete personal information to nonprofit organizations and social media. They were highly concerned about sharing of their personal information with third parties. On the other hand, the participants agree to provide personal information to secured systems and government organizations such as Migrationsverket, Skatteverket, DSV, KTH and Swedish banks. The respondents believed that these systems are quiet secure as well as their personal information will not be used for other purposes. The following suggestions for future research have been given in this area:

- Is there any possible limit of personal information with organizations perspective?
- What could be the possible agreed limit of personal information of both users and organizations' perspective?

Acknowledgements. The authors would to acknowledge the Higher Education Commission of Pakistan for supporting the work and providing the grant, under the grant No. Ref. No.300.362/TG/R&D/HEC/2018/26774.

References

1. Introna, L.D.: Privacy and the computer: why we need privacy in the information society. *Metaphilosophy* **28**, 259–275 (1997)
2. Shattuck, J.: Computer matching is a serious threat to individual rights. *Commun. ACM* **27**, 538–541 (1984)
3. Clarke, R.: Information technology and dataveillance. *Commun. ACM* **31**, 498–512 (1988)
4. Van Dijck, J.: Datafication, dataism and dataveillance: big data between scientific paradigm and ideology. *Surveill. Soc.* **12**(2), 197 (2014)
5. Jeff, S.H.: Privacy policies and practices: inside the organizational maze. *Commun. ACM* **36**, 104–122 (1993)
6. Yu, Y., Wang, Q., Ke, Z.: Research on security for personal information and privacy under network environment. In: International Conference on Computational Intelligence and Natural Computing, IEEE, vol. 2, pp. 277–279 (2009)
7. Udo, G.J.: Privacy and security concerns as major barriers for e-commerce: a survey study. *Inf. Manag. Comput. Secur.* **9**(4), 165–174 (2001)
8. Schneier, B.: What our top spy doesn't get: security and privacy aren't opposites. https://www.wired.com/2008/01/securitymatters_0124/
9. Schneier, B.: Protecting privacy and liberty (2001) https://www.schneier.com/essays/archives/2001/10/protecting_privacy_a.html
10. Guarda, P., Zannone, N.: Towards the development of privacy-aware systems. *Inf. Softw. Technol.* **51**, 337–350 (2009)
11. Katz, J.E., Tassone, A.R.: A report: public opinion trends: privacy and information technology. *Public Opin. Q.* **54**, 125–143 (1990)
12. Gavison, R.: Privacy and the limits of law. *Yale Law J.* **89**, 421–471 (1980)
13. Warren, S.D., Brandeis, L.D.: The right to privacy. *Harvard Law Rev.* **4**, 193–220 (1890)
14. Haag, V.D.: On privacy. In: Pennock, J.R., Chapman, J.W. (eds) *Privacy (Nomos XIII: year book of the American Society for Political and Legal Philosophy)*, New York: Atherton (1971)
15. Gross, H.: The concept of privacy. *New York Univ. Law* **42**, 35–36 (1967)
16. Fried, C.: Privacy. *Yale Law J.* **77**(3), 475–493 (1968)
17. Westin, A.F.: Piracy and freedom. *Washington and Lee Law Review* **25** (1968)
18. Parker, R.B.: A definition of privacy. *California Law Rev.* **27**, 275 (1974)
19. Johnson, J.L.: Privacy and the judgement of others. *J. Value Inq.* **23**, 157–168 (1989)
20. Decew, J.W.: The scope of privacy in law and ethics. *Law Philos.* **5**, 145–173 (1986)
21. Baars, H., Hintzbergen, J., Smulders, A., Hintzbergen, K.: Foundations of information security based on ISO27001 and ISO27002. Van Haren Publishing (2010)



Towards Protection Against a USB Device Whose Firmware Has Been Compromised or Turned as ‘BadUSB’

Usman Shafique¹(✉) and Shorahbeel Bin Zahur²

¹ Department of Computer Science, Bahria University Islamabad,
Islamabad, Pakistan

usman.buic@bahria.edu.pk

² Department of Computer Science, Comsats University Islamabad,
Islamabad, Pakistan
binzahur@gmail.com

Abstract. A BadUSB is a Universal Serial Bus (USB) device (usually a mass storage device) whose firmware has been modified so as to spoof itself as another device (such as a keyboard) in order to avoid being scanned by an anti-virus. This way, a pre-written script runs, after the infected USB device is plugged-in, and keystrokes from a keyboard are simulated. This can cause an attacker to install backdoors, keyloggers, password sniffers etc. This paper attempts to solving this problem by presenting hardware—software coupled design which allows the user to have an additional layer of security so that such devices can be identified and stopped.

Keywords: USB · Firmware attack · Device spoofing · Hacking · Device compromised

1 Introduction

When Universal Serial Bus (USB) was introduced in the market, it was a revolution. It never happened before that one type of port could offer a connection with various types of devices. A computer could now be connected to any kind of device starting from mass storage to medical appliances. The requirement of a separate serial port for each kind of device on the computer had become obsolete.

While bringing standardization to plug and play connectivity, USB also brought along with the few faults. For example, drivers of each device were required separately (though most are included with the OS), Device IDs are read by the computer during handshake process and there is no way to authenticate the device etc. The biggest problem that was and is still faced with the use of USB mass storage devices is the hidden malware/viruses/Trojans that accompany and infect each system they visit. Patches and anti-virus software, provide solutions for these problems, but many more are created each day.

1.1 Problem Statement

Similar to these vulnerabilities that are associated with a USB device, a new attack has been identified and demonstrated by Nohl and Lell [1]. This attack vector utilizes the vulnerability in the firmware of a USB device and updates it by flashing and re-writing the firmware or writing scripts within the memory gaps that are generated once the firmware is written on the memory. These scripts, spoof the device as another device (such as a keyboard) and are able to simulate keystrokes. These simulated keystrokes are read as regular key-presses from a regular USB-Keyboard and thus no anti-virus scans them. This allows the attacker to record a sequence of key-strokes and perform various flavors of attacks for example creation of administrator accounts, opening a back-door, privilege escalation etc. Figure 1 shows one of the attack vectors that can be performed by an attacker. There are various other kind of attack scenarios that are identified apart from rootkit installation or back-door opening.

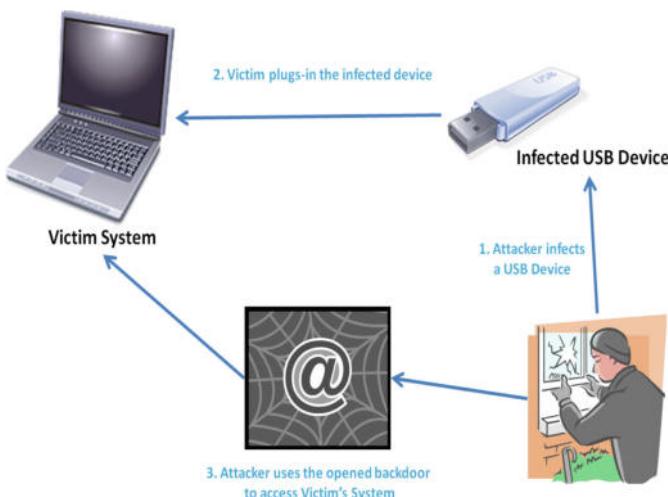


Fig. 1. The basic flow of attack

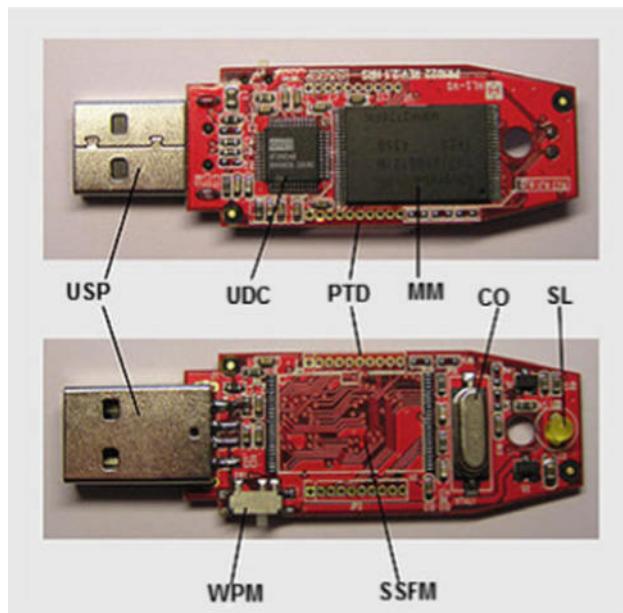
There is no effective defense from an attack over this vulnerability because of certain generic points [1]:

- (1) USB devices do not have standardized/unique serial numbers, therefore there is no way to identify if the connected device is spoofed.
- (2) Even the basic devices such as keyboards, web-cams etc. can be abused.
- (3) Firmware cannot be scanned as the firmware is required to access the device.
- (4) A solution that involves firmware-update lock would work but existing billions of devices would remain vulnerable.

This research paper focuses on the provision of a solution for the above-mentioned problem. The solution shall focus on proposing a design of a physical device that acts as a verifier, between a computer system and a USB device, being plugged in.

2 Structure of USB Storage Device

Before diving into the details of how a USB device is connected to a host, it is necessary to first understand what the basic structure of a USB device. A USB device consists of a few major components which are included with every device [2]. Figure 2 shows the internal structure of the USB device, Details of components are given below.



Abbreviation:

- USP: USB standard plug
- UDC: USB device controller
- PTD: Points to test the device
- MM: Main memory (in case of mass storage devices)
- CO: Crystal oscillator
- SL: Small light (usually an LED)
- WPM: Write-protect switch for main memory (in case of storage devices)
- SSFM: Space for second flash memory chip

Fig. 2. Structure of mass storage device [2]

2.1 Default Pinout for Most USB Devices

USB uses 4 wires shielded, two of them are used for power (GND and +5 V) and two of them are used for data signals (D+ and D-). Data signals are differential, has no termination needed and are transmitted on a twisted pair. To reduce the noise of electromagnetic effect on protracted lines, Half-duplex differential signaling is used.

Both data signals are not separate simplex connections they operate together. The default pinout for most USB devices is given below in Table 1 [3–5]. Each pin is associated with different functions.

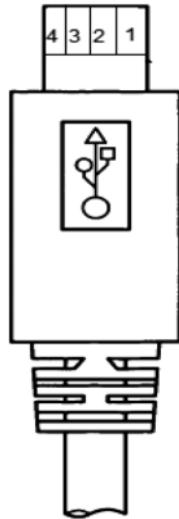


Table 1. Pin function

Pin number	Function
1	+5 V
2	Data –
3	Data +
4	Ground

2.2 USB Communication

USB device driver [7] interact with the application are explained with different steps Fig. 3 shows communication, Firstly it opens a file by calling WIN32 application programming interface (API), to create a connection with a device. After this another module KERNEL32.DLL implements, this API by requesting other system platform-dependent services to reach kernel-mode routine. According to application interact to the device, I/O manager create input-output request packet (IRP) which is the entry point in some device driver. Once USB device driver receive IRP, it constructs corresponding USB request block (URB), which provides information about how USB client drivers can use windows driver model (WDM), in next phase USB bus driver subdivide this URB into data packets and it sends to USB host controller driver, here it execute kernel mode and after this can directly talk to hardware.

Davies [6] introduces the working of USB and Mingli Li [7] explains briefly the communication between the application, USB controller, device and system.

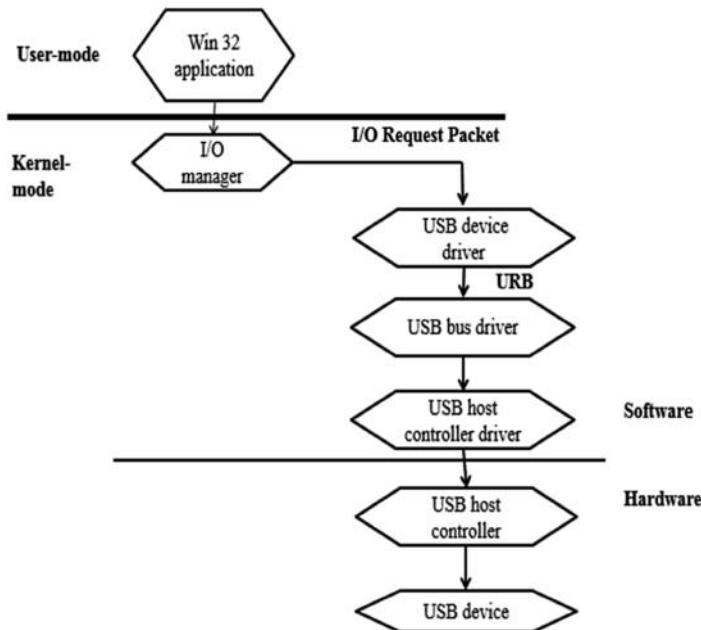


Fig. 3. Flowchart of USB device communication

3 Related Work

BadUSB is a novel attack and not much has been done or published as its solution. Where many claim that this attack is only focused on certain types of vendor-chipsets, another claim that any USB device is susceptible to this attack. This attack, however, utilizes the re-programming of firmware that has been discussed by Ang and Basnight [8, 9]. For this reason, this attack is very difficult to detect and protect against (once infection has occurred). There had been a history of USB devices (especially mass storage) to hold and disseminate viruses/Trojans. The example of Stuxnet worm [10] is very common and one that displays the power of these viruses if used for wrongful purposes.

3.1 Known USB Related Threats

Various kind of threats has been seen after the advent of USB. These include Trojans, batch files, zip bombs, viruses, worms etc. Few of these are listed below:

- (1) *Password stealer* [11]: Copies all the passwords from various sources in a system and saves/sends them.
- (2) *Zip bombs*: Attempts to crash the anti-virus program by causing an out-of-memory error.
- (3) *Folder Flooder*: Creates many folders in every directory and attempts to open them. Usually to annoy the user.

- (4) *Application flooders*: Opens various applications in an attempt to cause ‘out-of-memory’ state over a system.
- (5) *Recycler virus*: It copies Autorun.inf file to all the directories on a computer. Keylogger is also used.
- (6) *RAT (Remote Access tools) installer including Rootkit*: Rootkits [12] attempt to open backdoor channels (usually by running a trojan) to allow a remote hacker to take control, spy, or disable a system.

All of these viruses are propagated through the use of USB devices from one system to another. They reside in the main memory of the device and are hidden. They give out their initial instructions through batch files (normally autorun.inf). Many anti-virus programs catch these viruses even before they start propagating.

3.2 Firmware Upgrade

Vendors keep the firmware unlocked so as to allow firmware upgrades to be installed on the devices. This is one of the reasons that by firmware re-programming [13], a device may be turned malicious. The firmware update is a process that itself unlocks the firmware, which is why, BadUSB attack utilizes even the vendor-specific upgrade software to unlock, modify and re-flash the original firmware. The USB device firmware [14] upgrade document shows what steps are taken during upgrade [15].

3.3 Known Vulnerable Chipsets

Nohl et al. [14] and his team are thoroughly going through various types of chipsets/vendor-devices available in the market and are keeping a record of which all are vulnerable and to what extent. The data is concise and gives a clear picture of the fact that locking the firmware on a device is extremely necessary. Table 2 presents a detail of each chipset along with their firmware and their vulnerability analysis.

Even more, devices that include adapters and USB hubs are vulnerable as well.

3.4 Remedies for BadUSB

As the information related to BadUSB have spread around, the ideas on its remedies have also started emerging. These include various theories to actual products but none out of these have a complete solution as of yet.

- (7) *Good USB* [23]: Dave et al. proposed a solution is a Linux based one which stops the USB device registration and forces it to connect it to a honey-pot while the user is asked to verify the device authentic behavior.
- (8) *Center Tools DriveLock* [24]: Activates in early stages of boot and provides user the control to block/allow each peripheral device separately. This way any new connecting device requires approval from the user to be unlocked for use.
- (9) *Endpoint Protector* [25]: Similar to Drive Lock, this software tool provides user to control access of each peripheral device connected to the system. Whenever a new device is detected on the system, it is blocked till the time allowed specifically by the user.

Table 2. Chips detail

Device type	Chipset	Firmware update tool	Vulnerability analysis
USB storage	ALCOR AU698X	ALCOR MP_v14.01.24.00.zip [16]	Probably
USB storage	SMI SM325X/SM326X	RecoverTool_V2.00.33_L1224.exe [17]	Most likely
USB storage	Skymedi SK62XX SK66XX	SK6211_PDT_20090828.rar [18]	Probably
USB storage	Solid State System SSS6677, SSS6690 and SSS6691	SSS_MP.Utility_v2162.rar [19]	Probably
USB storage	Innstor IS903-A2, IS903-A3	Innstor_IS903_MP_Packae_V105_04_1303281.7z [20]	Probably
USB storage	Phison Chips	Psychson [21]	Most likely
Input/HID	Truly Ergonomic keyboard	Switch to allow firmware update	Probably
Input/HID	Apple USB Mighty Mouse Model-No. A1152	EEPROM which can only be written once	Not vulnerable
Input/HID	Logitech RX250 optical mouse	USB bootloader can re-program it	Most likely
Input/HID	USB Mouse Tchibo	Chip hardware bound for mouse application	Not likely
Input/HID	Logitech G5 mouse	G5Update12.exe [22]	Most likely
USB storage	Phison Chips	Psychson [21]	Most likely
Input/HID	Truly Ergonomic keyboard	Switch to allow firmware update	Probably
Input/HID	Apple USB Mighty Mouse Model-No. A1152	EEPROM which can only be written once	Not vulnerable
Input/HID	Logitech RX250 optical mouse	USB bootloader can re-program it	Most likely
Input/HID	Logitech G502 Proteus Core Gaming Mouse	G502Update_v16.exe	Most likely
Webcam	Cheap SpeedLink Reflect LED Webcam	No external flashable ROM found	Not likely
Webcam	Creative Labs Live! Cam Sync HD Model VFO770	External serial flash interface found that can re-flash the internal ROM	Most likely
Input/HID	Logitech G502 Proteus Core Gaming Mouse	G502Update_v16.exe	Most likely

- (10) *Ironkey by Imation* [26]: Ironkey is a product by Imation which claims to have a secure firmware by using digital signatures. This enables it to remain immune to firmware re-programming.
- (11) *Hardware (Physical) write-protect switches* are available but only for the main memory portion that is visible to the user. Same should also be made available for firmware updates [27]. Only plug-in the devices from the vendors/individuals that you trust.

Griscioli et al. [35] present a similar idea which requires intervention and confirmation from the user when a new device is connected. However, the registration and authorization process requires additional steps which may not be acceptable by every user. As this type of attack is limited to specific type of hardware and firmware, therefore in contrast, our proposed design automatically detects all the type of devices that are broadcasted by the device and shows it on a separate hardware. This device can be used individually as well as centrally by an enterprise to confirm if the USB has turned “bad” or not.

3.5 USB Protocol Analyzers

USB protocol analyzers work as sniffers [28] and read packets flowing between the USB device and the host system. These can be a hardware-software device [29, 30, 33] or a software-only product [31]. The hardware products [32] are used in conjunction with software products to analyze live traffic that travels on the USB bus. The usual layout for these devices works as shown in Fig. 4.

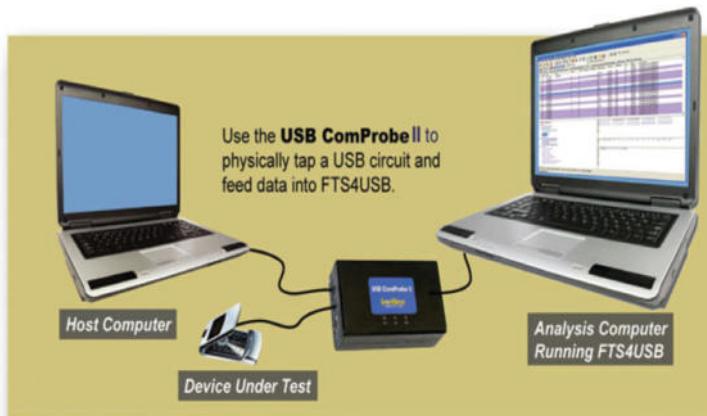


Fig. 4. Layout of hardware analyzers

These analyzers are mentioned here to point out that USB sniffers do exist and are available. Similar devices with supporting software can be made to counter the menace of BadUSB attack is explained in the preceding section.

4 Proposed Solution

As mentioned in the previous section, BadUSB is an attack that is quite difficult to detect once the malicious device's driver has been loaded in the system. Software only solutions can only protect the system to a certain extent and therefore a more robust solution may be followed. The proposed solution in this section utilizes a combination of hardware protocol sniffer and a software analyzer.

The proposed solution in this paper is aimed to provide a base design and considerations for an elaborate and practical design in the future. Although, before the solution can be presented, it is imperative to understand the basics of a USB device handshake.

As shown in Fig. 5, the USB device's firmware sends descriptors to identify what kind of a device it is. These descriptors are sent within a packet ID (as given in Table 3) [34]. The solution is to sniff and analyze the packets going from the device to the system, mid-way. Each packet may be let through except the ones that contain descriptors, especially during handshakes. These special packets must be analyzed and the device may be allowed to register itself to host, only if user permits. A preliminary design of the physical device is given in Fig. 6.

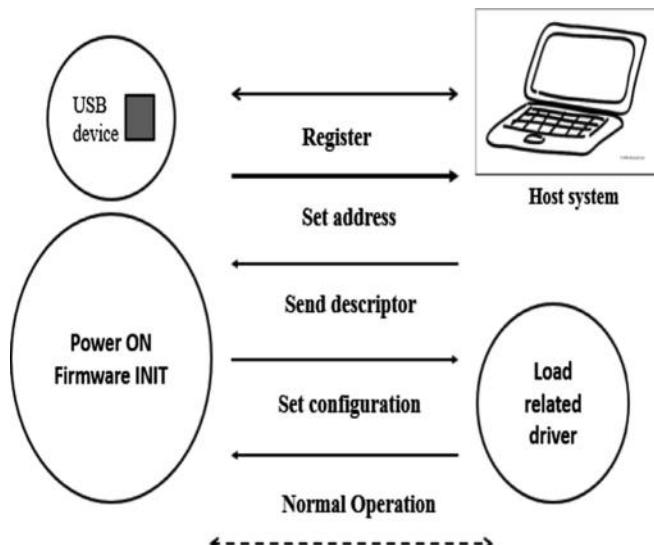
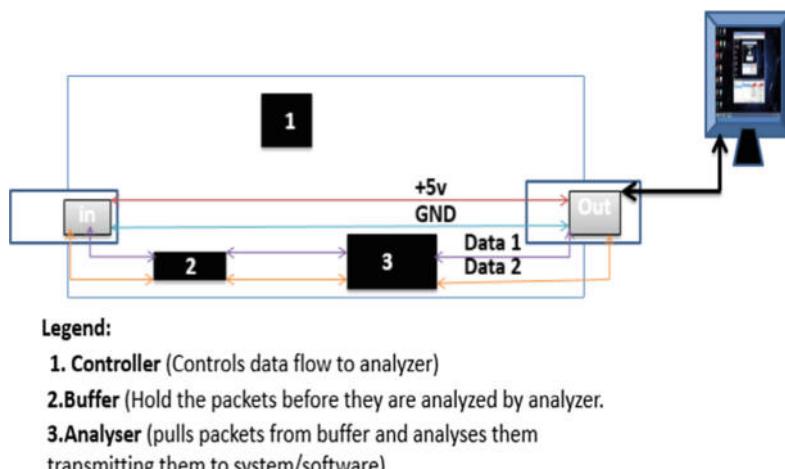


Fig. 5. USB device handshake/registration steps

Table 3. Field details

Field	Value	Description
<i>bLength</i>	Number	<i>Size of the Descriptor</i>
<i>bDescriptorType</i>	Constant	<i>Device Descriptor</i>
<i>bcdUSB</i>	BCD	<i>USB Specification Number</i>
<i>bDeviceClass</i>	Class	<i>Class Code (Assigned by USB Org)</i> <i>If equal to Zero, each interface specifies its own class code</i> <i>If equal to 0xFF, the class code is vendor specified.</i> <i>Otherwise, field is valid Class Code.</i>
<i>bDeviceSubClass</i>	SubClass	<i>Subclass Code (Assigned by USB Org)</i>
<i>bDeviceProtocol</i>	Protocol	<i>Protocol Code (Assigned by USB Org)</i>

**Fig. 6.** Preliminary design for a physical device

The physical device is designed on the lines of USB protocol analyzer which buffers the data going from USB device to the system, analyses it, and then sends it across. The lag in the communication can be reduced by only analyzing the handshake packets thoroughly and interrupting the communication when a new device handshake is observed. The flowchart in Fig. 7 explains these steps.

There can be two types of designs that can be implemented.

4.1 Hardware Sniffer/Software Analyzer

In this design, the physical device will only act as a sniffer and a buffer. Each packet will be sent for analysis to the corresponding software. The software will decide if the packet is trying to register a new device. If so, user permission will be obtained and if not, the packet will be allowed.

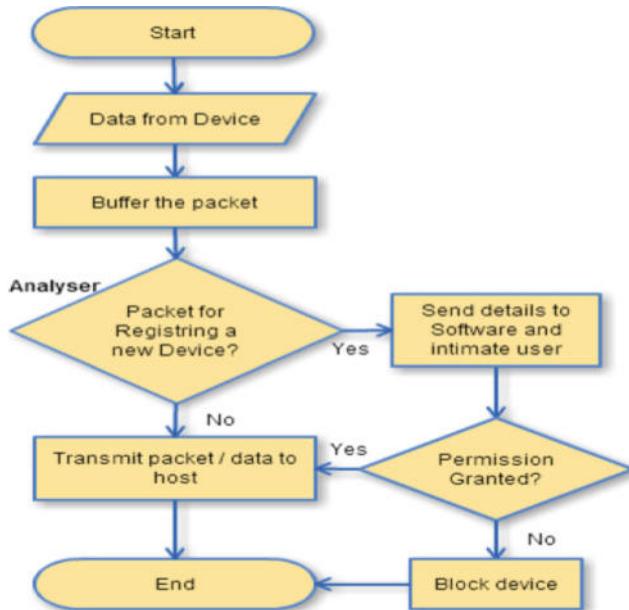


Fig. 7. Data flow of the proposed device

4.2 Hardware Sniffer and Analyzer/Software Analyzer

In this design, the hardware (like physical USB protocol analyzers [29, 30]) will have the capability to monitor all the packets and analyze and block/buffer only handshake packets while they are passed to software for thorough analysis and user's permission. There can be two types of layouts that can be designed for various kinds of situations.

1. The device connected to a system on the edge and all the USB devices that enter the premises may be scanned thoroughly. The working of all the devices may be authenticated and ensured that no device is acting maliciously.
2. The device may be connected to each system separately and any device may be connected/monitored through them.

5 Conclusion

With the world now evolving and adopting technology for connection of various devices, prevention and remedies to keep our systems safe from BadUSB menace are now more imperative than ever.

Daily, millions of corporations/offices/institutions/private users etc., rely on USB devices for transit/sharing of data and information. USB devices not only carry the sensitive data/information, but they are prone to various security vulnerabilities as well, BadUSB attack targets to exploit a crucial vulnerability in this tool. The only way to be safe from this attack is if the vendors lock the device firmware before releasing the

product in the market or a user take measures that block the device ‘before’ it has the chance to reach the system. The aim of this paper was to give out a generic design and layout for designers to implement in the future and to help individual users and organizations to remain protected.

6 Future Work

In future work, we will implement our design on an Arduino or Raspberry Pi based platform to design a separate scanning box. We will use a library like libusb to implement the proposed design which can be deployed in an enterprise environment. In order to validate the implementation, we would use a variety of USBs from different vendors and with different firmware. In our validation process, we will provide certain test cases like GoodUSB and BadUSB which ensure the accuracy of our proposed design. Our design can also be extended by the conjunction of an anti-virus solution on the same hardware in order to decrease USB based virus threats along with the BadUSB issue for an organization.

References

1. Nohl, K., Lell, J., Kri, S.: Turning USB peripherals into BadUSB (2014) [Online]. Available: <https://srlabs.de/badusb/>
2. Nohl, K., Kri, S., Lell, J.: BadUSB—on accessories that turn evil (2014)
3. USB Mass Storage Device (2011) [Online] <http://docshare01.docshare.tips/files/5761/57611265.pdf>
4. Caudill, Adam, Wilson, Brandon: Making BadUSB work for you. Derbycon, Location (2014)
5. USB in a Nutshell. Making Sense of the USB Standard
6. Davies, Z.: “USB,” Ziff Davies Inc (2010)
7. Li, G., Li, M., Zhao, G., Zang, J.: Research on USB driver for data acquisition. In: 2010 2nd International Conference on Future Computer and Communication (ICFCC), pp. V2-74-V2-78 (2010)
8. Cui, A., Costello, M., Stolfo, S.J.: When firmware modifications attack: a case study of embedded exploitation. In: Presented at the 20th Annual Network and Distributed System Security Symposium (2013)
9. Basnight, Z., Butts, J., Lopez, J., Dube, T.: Firmware modification attacks on programmable logic controllers. Int. J. Crit. Infrastruct. Prot. **6**, 76–84 (2013)
10. Denning, D.E.: Stuxnet: what has changed? Future Internet **4**, 672–687 (2012)
11. Password Stealing USB [Online]. Available: <http://www.gohacking.com/hack-passwords-using-usb-drive/>
12. Beegle, L.E.: Rootkits and their effects on information security. Inf. Syst. Secur. **16**, 164–176 (2007)
13. M. B. Solutions “User’s Guide,” no. February 2004
14. Project BadUSB [Online]. Available: <https://opensource.srlabs.de/projects/badusb>
15. Universal serial bus device class specification for device firmware upgrade, pp. 1–44 (1999)
16. Alcor: Alcor MP AU698x 100517 firmware [Online]. Available: <http://www.flashdrive-repair.com/2013/06/download-alcor-mp-au698x-100517-firmware.html>

17. Flashboot.ru: RecoverTool [Online]. Available: <http://flashboot.ru/iflash/page5/>
18. F. D. Repair, “SK6211_PDT_20090828.” [Online]. Available: <http://www.flashdrive-repair.com/2014/09/download-skymedi-sk6211-pdt-20090828.html>
19. Flashboot.ru, “3S_MP.Utility_v2162.” [Online]. Available: <http://flashboot.ru/files/file/270/>
20. Flashboot.ru, “Innostor_IS903_MP.Package.” [Online]. Available: <http://flashboot.ru/files/file/379/>
21. Caudill A.: Psychson—BadUSB code [Online]. Available: <https://github.com/adamcaudill/Psychson/>
22. Logitech, “G5Update12.exe.” [Online]. Available: <http://www.logitech.com/pub/techsupport/mouse/G5Update12.exe>
23. Tian, D.J., Bates, A., Butler, K.: Defending against malicious USB firmware with GoodUSB. *Acsac*, pp. 261–270 (2015)
24. D. Control and A. Control, “BadUSB- sticks locked out DriveLock Device Control protects against BadUSB Ludwigsburg, August 2014. Companies that want to protect against infection of a so-called BadUSB sticks have an effective solution with the award winning DriveLock Device Control,” 2014
25. Endpoint Protector [Online]. Available: <http://www.endpointprotector.com/solutions/badusb-threats-risks-and-how-to-protect-yourself>
26. Imation, “Ironkey.” [Online]. Available: <http://www.ironkey.com/en-US/solutions/protect-against-badusb.html>
27. Ducklin, P.: Never trust a USB device again [Online]. Available: <https://nakedsecurity.sophos.com/2014/08/02/badusb-what-if-you-could-never-trust-a-usb-device-again/>
28. USB Debug Techniques [Online]. Available: http://processors.wiki.ti.com/index.php/USB_Debug_Techniques#USB_protocol_analyze
29. Totalphase, “Beagle USB 12 Protocol Analyser” [Online]. Available: <http://www.totalphase.com/products/beagle-usb12/>
30. Ellisys, “USB Explorer 200, USB Protocol Analyser” [Online]. Available: <http://www.ellisys.com/products/usbex200/>
31. Virtual USB Analyser [Online]. Available: <http://vusb-analyzer.sourceforge.net/>
32. Teledyne, Mercury T2 Protocol analyser [Online]. Available: <http://teledynelecroy.com/protocolanalyzer/protocoloverview.aspx?seriesid=414>
33. Frontline, ComProbe USB [Online]. Available: <http://www.fte.com/products/FTS4USB-details.aspx>
34. B. Logic, USB a NutShell.” [Online]. Available: <http://www.beyondlogic.org/usbnutshell>
35. Griscioli, F., Pizzonia, M., Sacchetti, M.: USBCheckIn: Preventing BadUSB attacks by forcing human-device interaction. 2016 14th Annual Conference on Privacy, Security and Trust (PST). IEEE (2016)



A Multi-dimensional Adversary Analysis of RSA and ECC in Blockchain Encryption

Sonali Chandel^{1(✉)}, Wenzuan Cao¹, Zijing Sun¹, Jiayi Yang¹,
Bailu Zhang¹, and Tian-Yi Ni²

¹ New York Institute of Technology, Nanjing, China
`{schandel, wcao04, zsun12, jyang33, bzhang17}@nyit.edu`
² Arizona State University, Tempe, USA
`tianyinl@asu.edu`

Abstract. During this current age of big data, the security of sensitive data in the cyberspace has become utmost important. Blockchain, as a new age technology, provides the necessary tools to ensure data integrity and data protection using some encryption. Smaller transaction size and higher transaction efficiency are the essential requirements of the blockchain. However, these requirements are closely related to the efficiency of the encryption algorithms that blockchain uses. In this paper, we have analyzed and compared the performance of Rivest-Shamir-Adleman (RSA) algorithm and Elliptic Curve Cryptography (ECC) algorithm that is most commonly used in blockchain by having a general consideration of a transaction size and transaction efficiency. We aim to provide a better understanding of the reason behind their extensive use in the blockchain. We hope that our evaluation and analysis of these two encryption algorithms can help promote the further development of blockchain.

Keywords: Blockchain · Encryption algorithms · ECC · RSA · Privacy

1 Introduction

In the age of big data, one of the most important and valuable assets in the world is the data [1]. Almost every business in every field make most of their decisions after computing and analyzing the available data. Therefore, ensuring the safety and reliability of the data during its transmission and management has become a pressing and important issue. This also includes ensuring the authenticity of data sources and preventing malicious alteration of original data.

Blockchain as one of the most innovative technology in the present times is known for providing the integrity, authenticity, and confidentiality of the data in a decentralized way. Achieving all these security features successfully in a blockchain requires smaller transaction size and higher transaction efficiency, which in turn is strictly dependent on the encryption algorithms used for blockchain's implementation. Several encryption algorithms are used to implement blockchain, and they all have their pros and cons when it comes to their overall performance.

In this paper, we will compare and evaluate the performance of the two commonly used encryption algorithms in most of the applications of blockchain namely, Rivest-

Shamir-Adleman (RSA), and Elliptic Curve Cryptography (ECC), to find out the most efficient one. In Sect. 2, we introduce the structure of blockchain and a few other mainstream encryption algorithms used to implement this concept. In Sect. 3, we introduce the concept of RSA and ECC. In Sect. 4, we present an experiment to evaluate and compare the differences between RSA and ECC algorithms. Section 5 talks about the conclusion of our research and some future research prospects in this direction. The data we used in our experiment to analyze and compare the two algorithms mainly comes from [2, 3].

Compared to the papers we found as our reference, we found that most of them have separately discussed RSA and ECC from different aspects such as key size, key generation performance, signature verification performance, processing time and processor usage. However, no algorithm can show prominence in all aspects, so it is hard to give a general idea of which one is better. In this paper, we have designed our mathematical model to present a multi-dimensional analysis of these two algorithms so that we can find the most efficient one with an answer that could be straightforward and more comprehensive. The result of this study can help promote the further research and development of this fast-growing domain of blockchain. At the same time, it can help people in improving their awareness of privacy protection and making a better choice of encryption techniques while transferring their data in the cyberspace.

2 Related Background

2.1 Blockchain Overview

As a decentralized technology, the primary function of blockchain is to store data and information. It is a sequence of blocks, which holds the complete record of transactions like a public ledger [4]. As shown in Fig. 1 [5], each block in the chain carries a list of transactions and a hash to the previous block.

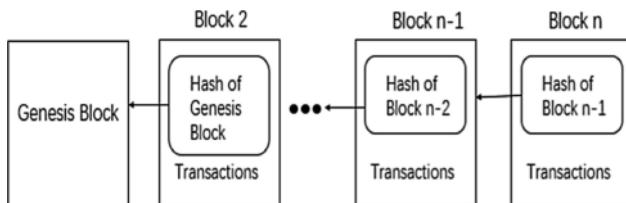


Fig. 1. A block diagram showing a Blockchain structure

A block serves as the basic unit of a blockchain and is composed of two parts - the block header and the block body. As shown in Fig. 2, the block header encapsulates block version, parent block hash, nBits, Nonce, Merkle tree root hash, and timestamp. The block body consists of a transaction counter and transactions. In a block, parent block hash and Merkle tree root hash mainly use encryption algorithms to protect data and transactions. Parent block hash records the hash value of the parent

block. The hash value of the current block must be smaller than that of the other one. Merkle tree root hash records all transactions. [5]

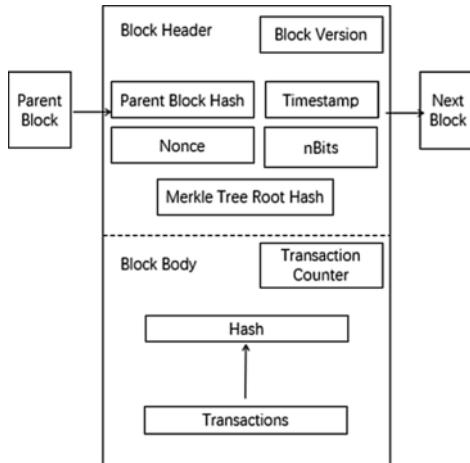


Fig. 2. The structure of a block in a Blockchain

When each transaction enters the block, the field needs to be recalculated for an update. The digital signature mainly functions at Merkle tree root hash. A digital signature uses various encryption algorithms to prevent malicious users from changing the data on purpose. This results in making the process of tampering very difficult for the hackers because in order to avoid any detection of their presence they would need to tamper the block containing that record as well as the ones linked to it as well [6].

2.2 Hash Function

A hash function is the fundamental component of a blockchain. All the data in the block is converted into hash value. A blockchain is a form of the hash chain where the latter block contains the hash of the former block. Such data structure guarantees that the data in the block cannot be interpolated. In simple terms, hashing means taking an input string of any length and giving out an output of a fixed length [7]. Some common hash functions are SHA (Secure Hashing Algorithm) series and MD5 (Message Digest Algorithm) series. A hash function can be used to hide information being transmitted and make it more secure. However, it does not allow for reverse engineering.

2.3 Cryptography

The records on a blockchain are secured through cryptography. Network participants have their private keys that are assigned to the transactions they make and act as a personal digital signature. If a record is altered, the signature will become invalid, and the peer network will know right away that something has happened [6]. Using cryptography in blockchain secures the identity of the sender of transactions and ensures that the records cannot be tampered with [8].

In this section, we will discuss the cryptography techniques that contain both encryption and decryption algorithms. Cryptography is a process of using advanced mathematical principles in storing and transmitting data in a way that only those, whom it is intended for, can read and process it [9]. Cryptography can be divided into two categories: Symmetric Cryptography (Private Key Cryptography) and Asymmetric Cryptography (Public Key Cryptography).

- *Symmetric Cryptography*—One of the oldest, most straightforward and popular encryption techniques that need only one secret key to be shared between the sender and the receiver of the message to cipher and decipher the information (as shown in Fig. 3) [9]. The secret key can be anything from a number to a word or a string of random letters, and it is needed to decipher the plain text.

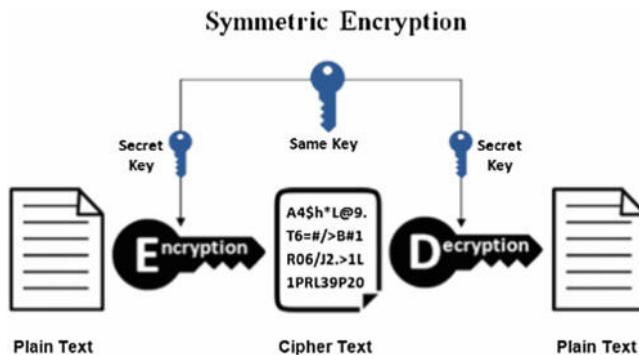


Fig. 3. The process of Symmetric Cryptography

- *Asymmetric Cryptography*—As seen in Fig. 4 [9], the process of encryption and decryption share a pair of private key and public key for one plain text. The public key can be used by anyone who wants to send a message, but the secret key is private for the decryption process. A message that is encrypted using a private key can be decrypted using a public key. While a message encrypted using a public key can only be decrypted using a private key. This ensures more security from the malicious users during the data transfer.

Table 1 shows the results of our comparison of the two basic types of cryptography namely symmetric and asymmetric. Although symmetric cryptography shows higher efficiency in encryption but to use this method, all the senders and receivers share the same encryption algorithm. This means that once one of the receiver's key is hacked or leaked, all of the receivers' message will be in danger. On the contrary, in asymmetric cryptography, senders and receivers share a pair of public key and private key (only belongs to the sender) which guarantees the safety of the message. Thus, most applications in the blockchain such as Bitcoin, Ethereum or some other cryptocurrencies use asymmetric cryptography for encryption. Despite being founded upon a similar framework, the type of cryptography used in the blockchain, namely public-key cryptography, is considerably better suited to the functions associated with the technology than symmetric-key cryptography [8].

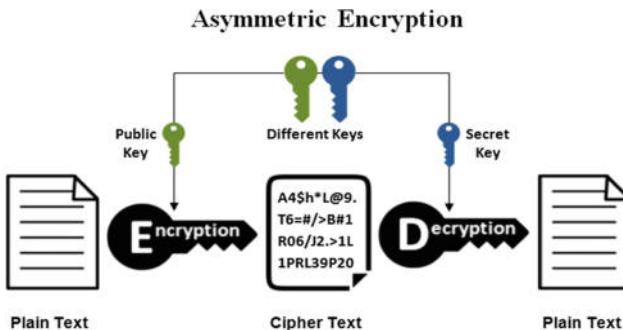


Fig. 4. The process of Asymmetric Cryptography

Table 1. Strengths and weaknesses of the two cryptography techniques

Type	Symmetric	Asymmetric
Strengths	Good effectiveness	Do not need to share the key in advance
Weaknesses	Need to share the key in advance (The key is easy to leak)	Low effectiveness
Typical algorithms	DES, 3DES, AES, IDEA	RSA, ECC, DSA

3 Encryption Algorithms Used in Blockchain

Blockchain encryption protects and prevents sensitive data from fraud or being manipulated by malicious users. The way blockchain and encryption security works are based on math, through a mining network. Every transaction that is verified by solving an algorithm, if approved by a consensus of over 51% of the blockchain network, is added to the blockchain. Once that transaction is added, it can never be changed [10].

Several algorithms of asymmetric cryptography are applied to applications of the blockchain. Moreover, the most commonly used algorithms are RSA and ECC. However, in recent years, the ECC algorithm has gradually replaced RSA as the mainstream of encryption algorithm in the blockchain. The most famous applications of the blockchain, such as Bitcoin and Ethereum use ECC algorithm. Therefore, we will compare and discuss different aspects of these two algorithms when they are applied to the blockchain. We will use data to demonstrate the reason why ECC algorithm is more popular and a popular choice for blockchain nowadays.

3.1 RSA Algorithm

RSA is one of the most influential public key encryption algorithms, proposed by Ron Rivest, Adi Shamir, and Leonard Adleman. This algorithm is based on the extremely difficult decomposition of large integers and can be used for both key encryption, and digital signature referred to as the factoring problem [11]. The product can be published as the encryption key, namely the public key, and the combination of two large prime

numbers can be set as the private key. The difficulty of getting the plaintext message back from the ciphertext and the public key is dependent on the difficulty of factoring the massive product of two prime numbers [7]. Algorithm 1 and 2 represents the algorithm for encryption and decryption in RSA respectively [12].

Algorithm 1: RSA Encryption

Input: RSA public key (n, e) , Plain text $m \in [0, n - 1]$

Output: Ciphertext c

Begin

1. Compute $c = m^e \bmod n$

2. Return c .

End

Algorithm 2: RSA Decryption

Input: Public key (n, e) , Private key d , Ciphertext c

Output: Plain text m

Begin

1. Compute $c = c^d \bmod n$

2. Return m .

End

3.2 ECC Algorithm

In 1985, N. Koblitz and Miller used an elliptical curve to implement the cryptographic algorithm called Elliptic Curve Cryptography. ECC is a kind of public key cryptography, where each user or device has a pair of keys, a public key, and a private key. The mathematical operations of ECC are defined over the elliptic curve $y^2 = x^3 + ax + b$, where $4a^3 + 27b^2 \neq 0$. Each value of the ‘a’ and ‘b’ gives a different elliptic curve. A public key is a point in the curve, and a private key is a random number. The public key is created by multiplying the private key with the generator point G in the curve. The generator point G , the curve parameters ‘a’ and ‘b’, together with a few more constants constitutes the domain parameter of ECC [7]. Algorithm 3 and 4 represents the algorithm for encryption and decryption in ElGamal ECC respectively [12].

Algorithm 3: ElGamal Elliptic Curve Cryptography Encryption

Input: Parameters field of an elliptic curve (p, E, P, n) , Public key Q , Plain text m

Output: Ciphertext (C_1, C_2)

Begin

1. Represent the message m as a point M in $E(F_p)$

2. Select $k \in R^{[1, n-1]}$.

3. Compute $C_1 = K_p$
4. Compute $C_2 = M + kQ$
5. Return (C_1, C_2)

End

Algorithm 4: ElGamal Elliptic Curve Cryptography Decryption

Input: Parameters field of an elliptic curve (p, E, P, n) , Private key d , Ciphertext (C_1, C_2)

Output: Plain text m

Begin

1. Compute $M = C_2 - dC_1$, and m from M .
2. Return (m) .

End

3.3 Time Complexity of RSA and ECC Algorithms

As shown in Fig. 5, the time complexity of RSA and ECC is $O((\log_2 x))^3$ and $O(\sqrt{x})$ respectively [13, 14]. Compared to RSA, ECC has a lower growth rate. RSA focuses on fast and straightforward encryption and verification [13], which is easier to implement and understand. However, the process of key generation, signing, and decryption is slower [15]. Under the same safety level, ECC has a smaller key size, faster key generation performance and more complex encryption algorithm that is difficult to decipher. However, ECC is complicated and tricky to implement securely [15]. ECC runs more efficiently and has better performance when used in encryption. In the following section, we will compare RSA and ECC in detail to prove that ECC has better performance than RSA.

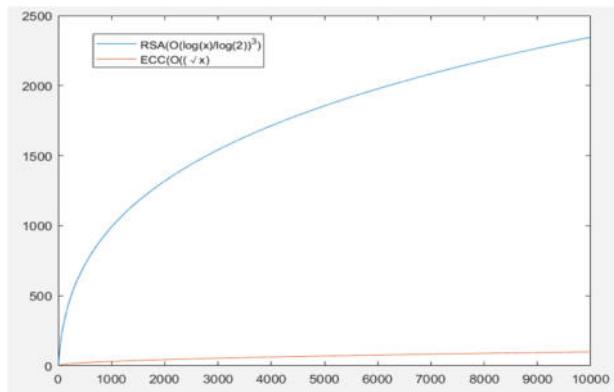


Fig. 5. The time complexity of RSA and ECC

4 Comparison of RSA and ECC Algorithms

4.1 Evaluation of RSA and ECC Algorithms

The continuous application of various transactions in the blockchain increasingly demands smaller transaction size and higher transaction efficiency. All these requirements are closely related to the encryption algorithms used during the transaction. Small key size will occupy less memory. Good key generation performance will spend less time and provide a higher speed of producing a transaction. Outstanding key verification performance will spend less time and provide higher speed for verifying the transaction [16].

In this paper in our model, we use MATLAB to do the calculation. We have picked the following three aspects namely; key size, key generation performance, and signature verification performance to make the comparison between RSA and ECC algorithms. Firstly, we will use the “Coefficient of Variation” method to determine the index weight. Then, the weight will be applied to the “TOPSIS” method and the “Grey Relational Analysis” (GRA) method to calculate the similarity and the grey relational grade. Finally, we will compare the results of these two models and conclude.

4.2 Determination of Index Weight

To determine the index weight of the three aspects—key size, key generation performance, and signature verification performance, we have used the “Coefficient of Variation Method” (CVM) [17]. Unlike most of the other subjective weight methods like the Analytic Hierarchy Process (AHP), CVM determines weight by using specifics statistics.

We define key size, key generation performance, and signature verification performance as x_1, x_2, x_3 respectively to be the criteria. We have two algorithms as alternatives to evaluate. We use ' i ' alternatives and ' j ' criteria.

Calculating the mean value \bar{x}_j and the standard deviation σ_j for each evaluation index:

$$\bar{x}_j = \frac{1}{2} \sum_{i=1}^2 x_{ij} \quad (j = 1, 2, 3) \quad (1)$$

$$\sigma_j = \sqrt{\sum_{i=1}^2 (x_{ij} - \bar{x}_j)^2} \quad (j = 1, 2, 3) \quad (2)$$

Calculating the variation coefficient for each index:

$$c_j = \sigma_j / \bar{x}_j \quad (j = 1, 2, 3) \quad (3)$$

Normalizing the variation coefficient and obtaining the weight of each index:

$$\omega_j = c_j / \sum_{j=1}^3 c_j \quad (j = 1, 2, 3) \quad (4)$$

4.3 Model 1 for Comparison—TOPSIS

Hwang and Yoon first developed TOPSIS. It is one of the ideal and classical methods for multi-criteria decision-making (MCDM) methods known for reliable evaluation results, quick computing process, ease of use, and understanding [18]. Moreover, it is based upon the concept that the chosen alternative should have the shortest distance from the ‘Positive Ideal Solution’ (PIS) and the furthest from the ‘Negative Ideal Solution’ (NIS) [19]. TOPSIS model is also useful for both the qualitative and the quantitative data, mainly because it normalizes the data at the beginning of the process that removes the effect of unit or magnitude. In their paper, Bafandehkar et al. [16] have only compared the advantages and disadvantages of ECC and RSA based on various criteria, but they have not given a conclusion about which one is better under the consideration of all the three aspects that we are considering for our model.

In our paper, MATLAB is used to help solve our model. Firstly, we keep the positive correlation coefficient constant and inverse the negative correlation coefficient. With the intersection of each alternative and criteria given as a_{ij} , we have a matrix $A = (a_{ij}) 2 \times 3$.

Then, we normalize the data:

$$a_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^2 a_{ij}^2}} \quad i = 1, 2; j = 1, 2, 3 \quad (5)$$

Multiply matrix A and weight of every criterion $W = (\omega_1, \omega_2, \omega_3)$.

$$b_{ij} = \omega_i a_{ij} \quad i = 1, 2; j = 1, 2, 3 \quad (6)$$

Then, we get a matrix $B = (b_{ij}) 2 \times 3$.

4.3.1 Determining the Worst Alternative C^0 and the Best Alternative C^*

The jth best alternative as c_j^* , the jth worst alternative as c_j^0 .

$$\text{Best alternative : } c_j^* = \max_{1 \leq i \leq 2} b_{ij} \quad j = 1, 2, 3 \quad (7)$$

$$\text{Worst alternative : } c_j^0 = \min_{1 \leq i \leq 2} b_{ij} \quad j = 1, 2, 3 \quad (8)$$

4.3.2 Calculating the Distance Between the Alternative and Best Alternative and the Distance Between the Alternative and the Worst Alternative

The distance between the alternative and the best alternative:

$$d_i^* = \sqrt{\sum_{j=1}^3 (b_{ij} - c_j^*)^2} \quad i = 1, 2 \quad (9)$$

The distance between the alternative and the worst alternative:

$$d_i^0 = \sqrt{\sum_{j=1}^3 (b_{ij} - c_j^*)^2} \quad i = 1, 2 \quad (10)$$

4.3.3 Calculate the Similarity to the Best Condition

$$f_i^* = \frac{d_i^*}{d_i^0 + d_i^*} \quad i = 1, 2 \quad (11)$$

4.4 Model 2 for Comparison—Grey Relational Analysis

Grey Relational Analysis (GRA) is based on the grey system theory. It calculates the relational coefficient and correlation degree between objects by comparing the geometric relations between the system's statistical data [20]. GRA has proved to be an effective method, especially for inexact, incomplete, multi-input and discrete data [14] [21]. In GRA, points are always considered as objects, and the distance between points or the concave and convex degree are mostly used to measure the correlations. By analyzing the correlation between the influencing factors (input variables) and selected objects (output variables), the correlation rank can be obtained.

GRA is widely applied in various fields because it does not require objects strictly observing specific mathematical laws or linear relationships [14]. It is an advisable solution to intricate correlations between multitudinous factors and variables [21]. GRA is superior in dealing with small samples and poor information system. Also, it only needs a small amount of calculation and has results consistent with the qualitative analysis [20]. By using this technique, we can relate influencing parameters and capability [21].

4.4.1 Defining Alternatives Array (Algorithms Evaluated)

$$x_i = \{x_i(k) | k = 1, 2, 3\}, \{i = 1, 2\} \quad (12)$$

We use formula (5) to normalize the matrix and formula (7) to get the best criteria array.

$$x_0 = \{x_0(k) | k = 1, 2, 3\}. \quad (13)$$

4.4.2 Calculating Grey Relational Coefficient $\rho \in [0, 1]$ Represents the Distinguishing Coefficient

We assume $\rho = 0.5$

$$\varepsilon_i(k) = \frac{\min_i \min_t |x_0(t) - x_i(t)| + \max_i \max_t |x_0(t) - x_i(t)|}{|x_0(k) - x_i(k)| + \rho \max_i \max_t |x_0(t) - x_i(t)|} \quad (14)$$

4.4.3 Calculating the Grey Relational Grade

$$\gamma_i = \sum_{k=1}^n \omega_i \varepsilon_i(k) \quad (15)$$

We use γ_i to describe the correlation degree between x_i and x_0 , namely to describe the influence on x_0 caused by the change of x_i .

4.5 Data Testing and Analysis

As already mentioned in Sect. 4.2, we will analyze the performance of RSA and ECC on three aspects. Moreover, for each aspect, we will analyze the data at the same security level. This means that both the algorithms will be compared under the same level of encryption. Security level is usually expressed in “bits,” where n-bit security means that the attacker would have to perform 2^n operations to break it [22].

From Table 2, we can see that ECC requires smaller key sizes than RSA under the same security level. The minimum key size required for a secured cryptosystem of ECC is 160 bits or more [16]. Therefore we have chosen the key size 160 bits for ECC and 1024 bits for RSA as the starting point. In addition, to see how RSA and ECC performance develops under these three factors as the key sizes increase, we must choose at least five different key sizes. As seen from Table 4, some factors like signature verification performance for RSA do not show any noticeable difference until the key size increases to 15,360 bits.

- *Key size:* Table 2 shows the security level (bits) and the ratio of cost for RSA and ECC with equivalent security level [2].
- *Key generation performance:* Table 3 shows the performance comparison of RSA and ECC in key generation performance [3].
- *Signature Verification performance:* Table 4 shows the performance comparison of RSA and ECC in signature verification performance [3].

Table 2. Security level (bits) and the ratio of cost for RSA and ECC with the equivalent security level of Encryption

Key size		Security level (bits)		Ratio of cost
RSA/DSA	ECC			
1024	160	80		3:1
2048	224	112		6:1
3072	256	128		10:1
7680	384	192		32:1
15360	521	256		64:1

Table 3. Performance comparison of RSA and ECC in key generation performance

Key length		Time(s)	
ECC	RSA	ECC	RSA
163	1024	0.08	0.16
233	2240	0.18	7.74
283	3072	0.27	9.80
409	7680	0.64	113.90
571	15,360	1.44	679.06

Table 4. Performance comparison of RSA and ECC in signature verification performance

Key length		Time(s)	
ECC	RSA	ECC	RSA
163	1024	0.23	0.01
233	2240	0.51	0.01
283	3072	0.86	0.01
409	7680	1.8	0.01
571	15,360	4.53	0.03

4.6 Results of Index Weight

Table 5 shows the weight of the three criteria for five key sizes.

We can see from Tables 1, 2 and 3 respectively that the contrast between the performance of ECC and RSA under three aspects becomes increasingly evident as the key size increases. We also see that with the value of the 5th key size, the difference between the two algorithms is at the highest level. This is also good for our mathematical analysis to evaluate the performance of ECC and RSA.

Table 5. Weights of three criteria

Weight	Key size	Key generation performance	Signature verification performance
1	0.367,203,327	0.16,874,578	0.464,050,894
2	0.293,597,723	0.351,615,031	0.354,787,246
3	0.30,552,136	0.341,708,417	0.352,770,223
4	0.313,696,153	0.343,418,172	0.342,885,676
5	0.320,324,987	0.341,367,455	0.338,307,557

4.7 TOPSIS Model Results

An initial decision matrix as shown in (16) is created by taking the values of the best key size, key generation performance, and signature verification performance from Tables 2,3 and 4 for ECC (row 1) and RSA (row 2) respectively:

$$A = \begin{pmatrix} 521 & 1.44 & 4.53 \\ 15360 & 679.06 & 0.03 \end{pmatrix} \quad (16)$$

Using formula (5), we calculate normalized matrix A in (17).

$$A = \begin{pmatrix} 0.9994 & 1.0000 & 0.0066 \\ 0.0339 & 0.0021 & 1.000 \end{pmatrix} \quad (17)$$

The weighted normalized decision matrix B using formula (6) is shown in (18).

$$B = \begin{pmatrix} 0.3201 & 0.3414 & 0.0022 \\ 0.0109 & 0.0007 & 0.3383 \end{pmatrix} \quad (18)$$

Using formula (9) and (10), the vector of positive ideal solution c_j^* and negative ideal solution c_j^0 are shown in (19) and (20).

$$c_j^* = (0.3201 \quad 0.3414 \quad 0.3383) \quad (19)$$

$$c_j^0 = (0.0109 \quad 0.0007 \quad 0.0022) \quad (20)$$

Table 6 shows the results of our evaluation using the TOPSIS model. Using formula (11), we get the values of f_i^* . The smaller the value of f_i^* , the closer the positive ideal solution gets.

Table 6. Closeness coefficients and rank

Alternative	d_i^*	d_i^0	f_i^*	Rank
ECC	0.3361	0.4601	0.4421	1
RSA	0.4601	0.3361	0.5779	2

4.8 GRA Model Results

After using formula (5) and formula (7), we get the best criteria array in (21).

$$x_0 = (0.9994 \quad 1.0000 \quad 1.0000) \quad (21)$$

Table 7 shows the result of our evaluation using GRA. We calculate the grey relational grade using formula (13), (14) and (15). The bigger the γ_i value is, the better alternative it becomes.

$$\varepsilon_i(k) = \begin{pmatrix} 1.0000 & 1.0000 & 0.3343 \\ 0.3407 & 0.3333 & 1.0000 \end{pmatrix} \quad (22)$$

Table 7. Grey relational grade and rank

Alternative	γ_i	Rank
ECC	0.7748	1
RSA	0.5612	2

4.9 Comprehensive Analysis of RSA and ECC Algorithms

Figure 6 shows that under the same level of security, ECC performs better than RSA in most cases. The evaluation results using both TOPSIS and GRA methods confirms that as well. The longer the key length is, the better the performance of ECC becomes. Compared with RSA, ECC has many advantages in many aspects. It has a stronger resistance to attack, lower CPU and content usage, lower network consumption, and faster encryption. To achieve the same level of security, the key length required by the ECC algorithm is much lower than the RSA algorithm. Therefore, it can achieve a balance between key length and level of security very effectively.

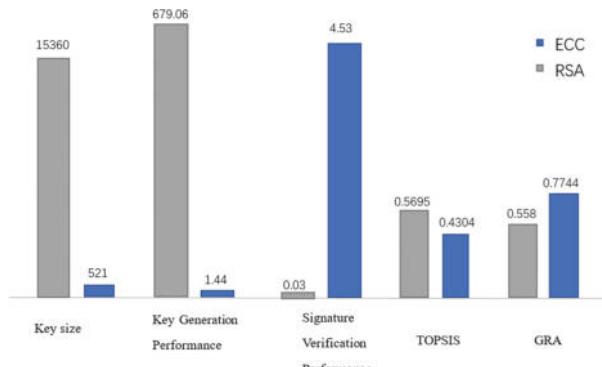


Fig. 6. Performance of RSA and ECC using TOPSIS and GRA (under the same level of security)

5 Conclusion and Future Work

In this paper, we have analyzed RSA and ECC algorithms based on their key size, key generation performance and signature verification performance to conclude that ECC performs better than RSA when used under the same encryption level. To avoid any contingency, we used TOPSIS and GRA models as our comparison models to verify our conclusion. According to the results, it is undeniable that RSA is superior to ECC in signature verification performance. However, TOPSIS and GRA models both confirm that ECC performs better than RSA in general as an encryption algorithm. This also explains the reason why ECC is preferred over RSA in designing blockchain. As mentioned earlier in our paper, blockchain relies heavily on the encryption algorithm used for its security. According to our research results, all the characteristics to satisfy the security needs of blockchain are met better by ECC than RSA.

In the future, research on the same topic will need us to find more factors that might be affecting the performance of RSA and ECC, as well as better ways to determine their weight in our comprehensive analysis. In future we would also like to compare the performance of other algorithms used in the blockchain.

References

1. Schwab, K., Marcus, A., Oyola, JO., Hoffman, W., Luzi, M.: Personal data: the emergence of a new asset class. In an Initiative of the World Economic Forum (2011)
2. National security agency: The Case For Elliptic Curve Cryptography, ([nsa.gov](http://www.nsa.gov/business/programs/elliptic_curve.html)). http://www.nsa.gov/business/programs/elliptic_curve.html (2009). Last accessed 15 Jan 2009
3. Jansma, N., Arrendondo, B.: Performance Comparison of Elliptic Curve and RSA Digital Signatures. (2004). Accessed on 28 Apr 2004
4. Lee Kuo Chuen, D. (ed.) Handbook of digital currency, 1st edn, Elsevier (2015)
5. Zheng, Z., Xie, S., Dai, H., Chen, X., Wang, H.: An overview of blockchain technology: architecture, consensus, and future trends. In: 2017 IEEE 6th International Congress on Big Data, June 2017
6. <https://www.ibm.com/blogs/blockchain/2017/12/blockchain-security-what-keeps-your-transaction-data-safe/>
7. Kumar, A., Tyagi, S.S., Rana, M., Aggarwal, N., Bhadana, P.: A comparative study of public key cryptosystem based on ECC and RSA, (May 2011)
8. <https://lisk.io/academy/blockchain-basics/how-does-blockchain-work/blockchain-cryptography-explained>
9. Symmetric vs Asymmetric Encryption—What are the differences? <https://www.ssl2buy.com/wiki/symmetric-vs-asymmetric-encryption-what-are-differences>
10. <https://www.techrepublic.com/article/how-blockchain-encryption-works-its-all-about-math/>
11. Savari, M., Montazerolzohour, M., Thiam, Y.E.: Comparison of ECC and RSA algorithm in multipurpose smart card application
12. da Silva Quirin, G., Moreno, ED.: Architectural evaluation of algorithms RSA, ECC and MQQ in arm processors. Int. J. Comput. Net. Commun. (IJCNC). 5(2), (2013)
13. Ting-ting, G., Tao, L.: The implementation of RSA public-key algorithm and RSA signature algorithm. Department of Computer Science, Sichuan University (1999)
14. Han, M., Zhang, R., Qiu, T., Xu, M., Ren, W.: Multivariate, chaotic time series prediction based on improved grey relational analysis, Senior Member, IEEE (2017)

15. Michael Hamburg “Which one is better: elliptic curve cryptography or RSA algorithm and why?” <https://www.quora.com/Which-one-is-better-elliptic-curve-cryptography-or-RSA-algorithm-and-why>
16. Bafandehkar, M., Yasin, S.M., Mahmood, R., Hanapi, Z.M.: Comparison of ECC and RSA algorithm in resource-constrained devices. In: International Conference on It Convergence & Security, 3 Jan 2013
17. Dong, D., Yan, Y., Wang, Z.: Application of grey correlative model for surface quality evaluation of strip steel based on the variation coefficient method. In: 2011 International Conference on Electronic & Mechanical Engineering and Information Technology
18. Primasari, C.H., Setyohadi, D.B.: Financial analysis and TOPSIS implementation for selecting the most profitable investment proposal in goat farming. In: 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (2017)
19. García-Cascales, M.S., Lamata, M.T.: On rank reversal and TOPSIS method. *Math. Comput. Model.* **56**(5), 123–132 (2012)
20. Zhou, L., Wang, Y., Wang, F., Yan, C., Bi, J.: A transformer fault diagnosis method based on grey relational analysis and integrated weight determination, IEEE, (June 2017)
21. Khurana, A., Sharma, T., Shukla, K.K.: Optimization of parameters affecting the performance of wind turbine blade using grey relational analysis, IEEE (2017)
22. Lenstra, Arjen K, Key Lengths: Contribution to The Handbook of Information Security, Citibank, N.A., and Technische Universiteit Eindhoven, 1 North Gate Road, Mendham, NJ 07945–3104, U.S.A



Assessing the Performance of a Biometric Mobile Application for Workdays Registration

Cristian Zambrano-Vega^(✉), Byron Oviedo, and Oscar Moncayo Carreño

Universidad Técnica Estatal de Quevedo, Quevedo, Los Ríos, Ecuador
`{czambrano,boviedo,omoncayo}@uteq.edu.ec`
<http://www.uteq.edu.ec>

Abstract. The professors in the State Technical University of Quevedo - Ecuador (UTEQ) must register the workdays (workday entries and workday exits) in the attendance management software provided by the Human Resources department through static biometric devices. In some cases, the biometric devices are not close to their offices or classrooms, so they forget to register their workdays, wrong workdays registrations. With the aim of improving this registration process we have developed bioFACE, a novel mobile application for biometric authentication by face recognition, which allows to convert the user smartphones in biometric devices, connected to the attendance management software, avoiding large crowds in rush hours moments, especially. With the aim to assess its performance, we have carried out some experiments measuring the features accuracy and workdays registration time. Despite the limited CPU and memory capabilities of today's mobile phones, the obtained results are very promising, shows a high accuracy facial identification and a faster and easy alternative to the workday registration.

Keywords: Mobile biometric authentication · Face recognition · Mobile smart applications

1 Introduction

Mobile devices are rapidly becoming a key computing platform, transforming how people access business and personal information. The rich set of input sensors on mobile devices, including cameras, microphones, touch screens, and GPS, enable sophisticated multimedia interactions. Biometric authentication methods using these sensors could offer a natural alternative to password schemes, since the sensors are familiar and already used for a variety of mobile tasks [1].

An increasing attention is given to a new application field of face detection and recognition dealing with mobile phones. This attention is not due to some blind chance but it grew with several practical needs [2, 3]. Since more and more mobile phones provide wireless access to the Internet and high computational

performance, is then possible to perform a mobile biometric authentication by face recognition.

Biometric time and attendance system has brought more precise system to measure group or individual's activities and attendance as well. Biometric attendance machine captures your unique biological/physical feature such as your hand or fingerprint, iris pattern, face and sometimes even your voice as a record for identity verification and allows you to perform something that you are authorized to do.

In this paper we present bioFACE, a novel mobile application for the biometric authentication by face recognition of the users of the attendance management software provided by the Human Resources department of the State Technical University of Quevedo. This application converts the smartphones in a biometric device that allows register the workday entries and workday exits from any place inside of the university campus. The user-location is validated by the GPS coordinates using the Android Geofence API [4] and the biometric authentication of the users (employees and professors) is carried out by face recognition performed by Microsoft Face API features [5]. Furthermore, with the aim of assess its performance, we have carried out some experiments measuring the features accuracy and workdays registration time. For more details of the mobile application, the bioFACE website is: <https://alex-jr-1994.wixsite.com/bioface>.

The rest of this paper is organized as follows: First, the Biometrics definitions are described in Sect. 2. In Sects. 3 and 4 we detail the specification of the APIs to Face Recognition (Microsoft Face API) and GeoLocation (Android Geofencing), respectively. The features, threats and Benefits of our proposal bioFACE are detailed in Sect. 5. The application performance analysis is detailed in Sect. 6 and finally, Sect. 7 outlines some concluding remarks and suggest some lines of future work.

2 Biometrics Definitions

Biometrics are distinctive, measurable characteristics that are used to identify us. These identifiers are categorized as physical or behavioral characteristics, related the shape of specific body parts or a pattern of behavior. Physical biometrics include fingerprints, facial structure and shape, hand geometry, iris patterns, even the shape of your ear. Behavioral biometrics refer to how we type, how we swipe our fingers on a touchscreen, our gait, or how we speak or sing [6].

Unlike passwords, biometrics are part of us. They cant be forgotten, lost or borrowed. And they are not easily hackable. Thats why biometrics are more secure than passwords and convenient to use.

2.1 Face Recognition

Face recognition has been an active research topic since the 1970s [7]. Given an input image with multiple faces, face recognition systems typically first run

face detection to isolate the faces. Each face is preprocessed and then a low-dimensional representation (or embedding) is obtained. A low-dimensional representation is important for efficient classification [8].

The shape of the face is a very unique physical characteristic. Using computer vision technology, facial structure and shape can be identified and categorized, with specific “landmark” features, including the relative position and shape of the nose, eyes, jaw, and mouth. The earliest work in face recognition was feature-based and sought to explicitly define a low-dimensional face representation based on ratios of distances, areas, and angles [9]. Today’s top-performing face recognition techniques are based on convolutional neural networks. Facebook’s DeepFace [10] and Google’s FaceNet [11] systems yield the highest accuracy. However, these deep neural network-based techniques are trained with private datasets containing millions of social media images that are orders of magnitude larger than available datasets for research [9]. Advances in facial imaging today allow for 3D images, increasing the complexity and accuracy of face recognition algorithms

3 Microsoft Face API (FKA ‘Project Oxford’)

Microsoft Face API (Application Program Interface) is part of Microsoft’s Cognitive Services Platform, is a cloud-based service that provides advanced face algorithms to perform two main functions: face detection with attributes and face recognition.

3.1 Face Detection

Detect one or more human faces in an image and get back face rectangles for where in the image the faces are, along with face attributes which contain machine learning-based predictions of facial features. The face attribute features available are: Age, Emotion, Gender, Pose, Smile, and Facial Hair along with 27 landmarks for each face in the image that can be specified by file in bytes or valid URL. One example of the result is illustrated by Fig. 1 [5].

3.2 Face Recognition

Face recognition is widely used in many scenarios including security, natural user interface, image content analysis and management, mobile apps, and robotics. Face-Recognition provides four face recognition functions: face verification, finding similar faces, face grouping, and person identification. To perform the facial recognition in our application we have used the features of the Face-Identify API.

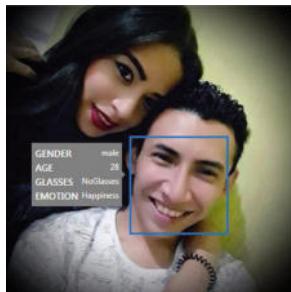


Fig. 1. Face detection example with face attributes details

Face-Identify API Face-Identify API identify people based on a detected face and a people database (defined as a LargePersonGroup/PersonGroup). Each group may contain up to 1,000,000/10,000 person objects. Meanwhile, each person object can have up to 248 faces registered. For each face in the faceIds array, Face Identify will compute similarities between the query face and all the faces in the person group (given by personGroupId) or large person group (given by largePersonGroupId), and return candidate person(s) for that face ranked by similarity confidence (percentage). One example of the use of this API can be found at [12] and is illustrated by Fig. 2.

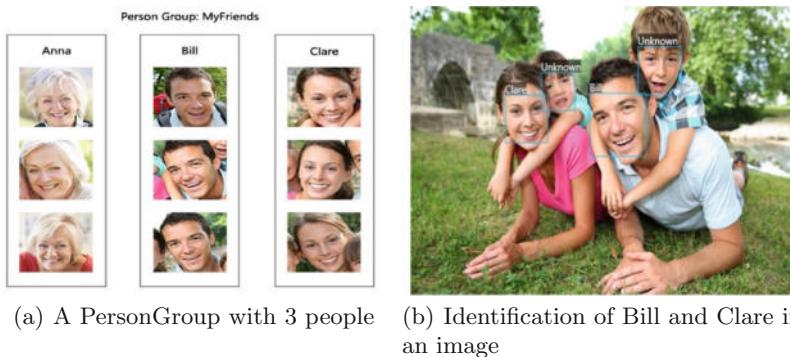


Fig. 2. Face recognition example using the Face-Identify API [12]

4 Geofencing

Geofencing is a location-based service, that sends a notification to a smartphone who enter a defined geographic area. For example, human resource department to monitor employees working in special locations or child location services can

notify parents if a child leaves a designated area. To mark a location of interest, you specify its latitude and longitude. To adjust the proximity for the location, you add a radius. The latitude, longitude, and radius define a geofence, creating a circular area, or fence, around the location of interest [4, 13]. Figure 3 illustrates an example of geofencing on Android.

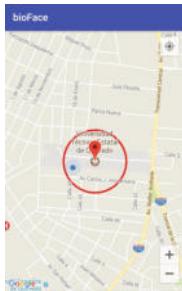


Fig. 3. Geofencing on android example at Quevedo State Technical University

In Android, for each geofence, you can ask Location Services to send you entrance and exit events, or you can specify a duration within the geofence area to wait, or dwell, before triggering an event. You can limit the duration of any geofence by specifying an expiration duration in milliseconds. After the geofence expires, Location Services automatically removes it [4].

5 BioFace



BioFace is a mobile application for the control of biometric assistance from mobile devices that uses facial recognition as an authentication measure and the geographic coordinates of the GPS (Global Positioning System) as a validation of the location of employees and professors in their respective dependencies or classrooms. BioFace has a previous storage of the company's employees: personal and biometric (face) data and the GPS coordinates of the geographical location of the company. The URL of the website is: <https://alex-jr-1994.wixsite.com/bioface>.

The principal features, threats and benefits are described to follow:

5.1 Features

- Biometric authentication of employees through facial recognition.
- Validate the company's rank in real-time geographic location.
- Report of the attendance record of employees with: Day, Time, Date, and compatibility level of the authenticated face.
- Administrator profile to save basic data of the employees of the company.
- Compatible with Android operating systems.
- Interoperability with attendance management systems. Synchronize the registration through JSON Web Services.

5.2 Benefits

- Fast registration of the workday starts and workday exists of your working day from your same work location.
- Avoid long crowds in the place where the control of attendance Rush hours entry or exit of the majority of employees.
- Ideal in companies where work offices are very distant from the biometric devices.
- Reduce costs by purchasing several assistance control equipment, maybe none is necessary.
- The physical characteristics of a face are much more difficult to counterfeit.
- Reasonable prices, flexible licensing and free customer support.

5.3 Threats

- There are third-party applications that manage to modify the actual geographical location of GPS of mobile devices, but their operation requires technical knowledge such as: activate ROOT mode, activate developer options or apply location changes.
- Low internet connection.
- Facial recognition also requires updates to maintain accuracy, while fingerprints, for example, never need it, age affects our appearance, so, as we get older, the images will have to be updated regularly to ensure that the data is accurate.

5.4 How It Works

A bioFace flowchart is described to follow:

1. First, we must create a person group for each career o group of employees, with an specified: personGroupId, name, and user-provided user Data. After creation, we have to add persons (employees or professors) into each group, taking the face images and saving it into the employee personal data, then train the PersonGroups to get ready for Face - Identify. Figure 4 illustrates an application screenshot on this step.

2. The first time that the application is executed, request to user assign the permissions to access of: internet, camera, media storage and geolocation. Figure 5 illustrates an application screenshot on this step.
3. Login User: All the employees and professors must log in to the system using their identification card number, so that the application identifies and recognize their registered dependency or faculty. Figure 6 illustrates an application screenshot with the login interface.
4. Main Interface: This is the interface of the application. Among the options are: the function to register the start and end of work, configuration of the personal information and a report with the registers. Figure 7 illustrates an application screenshot with the main interface.
5. Personal Data: The user (employees or professors) can modify their personal data, including: Name, E-mail and registered career. Figure 8 illustrates an application screenshot with the user personal data.
6. Register the workday start or workday exit: When a user requests register its start or end of the workday, first, the application validates that he is inside the university campus, then the camera is opened for the biometric authentication of the employee and captures the face as an image. The facial recognition is carried out with the help of the Face API. Once confirmed that the face corresponds to the person who registers, the application registers the event to the system sending by JSON Web Service the personal data: EmployeeID or ProfessorID, Date and Time and GPS coordinates. Figure 9 illustrates an application screenshot on this step.
7. Reports: The application shows a report with all the registrations performed by user, the fields are: Date, Time and the confidence percentage with which the facial recognition was performed. Figure 10 illustrates an application screenshot with the report data.

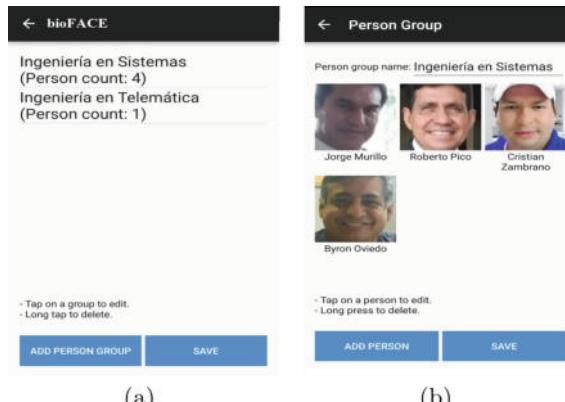


Fig. 4. Creating person groups for each dependency or career. **a** List of careers or dependencies and **b** List of professors of one career (group)



Fig. 5. Permissions requests: **a** GPS geolocation permissions request, **b** Camera permissions request



Fig. 6. User login

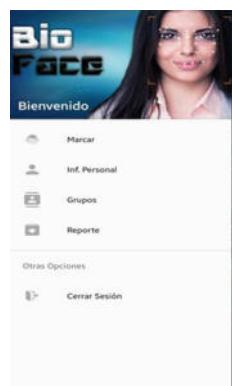


Fig. 7. Options of the main menu of the application



Fig. 8. Personal data setting

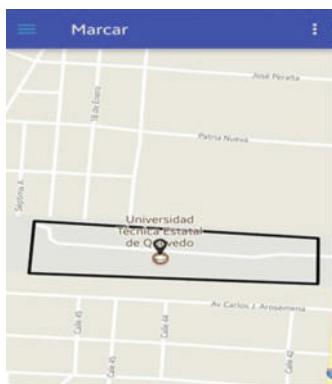


Fig. 9. Register the workday entry or workday exit

This screenshot shows the 'Reporte' (Report) section of the app. The title 'REPORTE' is at the top. The report lists four entries, each with a registration number, date, time, and confidence level:

# de Registro	
Fecha	2018/01/30
Hora	23:12
Confianza	0.88733
# de Registro	
Fecha	2018/02/02
Hora	00:32
Confianza	0.77107
# de Registro	
Fecha	2018/02/08
Hora	21:42
Confianza	0.92118
# de Registro	
Fecha	2018/02/08

Fig. 10. Report of the registrations

6 Performance Analysis

With the aim of assess the performance of our application bioFace, we have carried out two experiments, measuring the accuracy of its main features and the required time (in minutes) to perform the workdays registration. The considered features are: face identification, face recognition, geolocalization and workday registration. We have selected 10 users (5 professors and 5 employees), each one of them used their own smartphones connected to internet using a wi-fi or mobile data connection. These experiments were executed for 4 days.

Table 1 shows the task error rates (%) of each feature and Table 2 shows the elapsed time (in minutes) to perform a workday registration, where the *T1* column represents the current biometric device time, the *T2* column represents the time using our mobile application and *Diff* column is the difference between both times.

The results illustrated by Table 1 shows that the face identifications, face recognition and the workdays registration features have a low task error rate, less than 5%, but the GeoLocalization feature have a slightly higher error rate, less than 8%, produced by the low GPS service precisions of the smartphone.

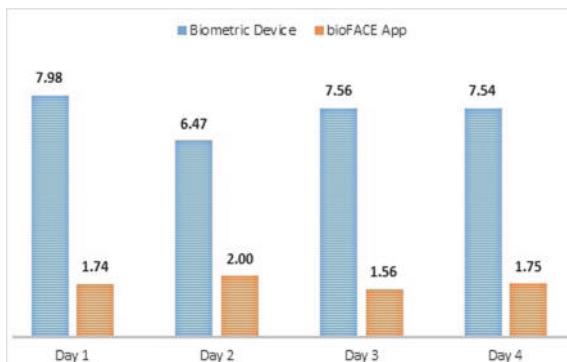
The elapsed time illustrated by Table 2 shows that our mobile application allows to professors and employees reduce the required time to perform a workday registration just using their smartphone with internet even from their personal workplace. These results are summarized in Figure 11, illustrating the average of the workday registration time (in minutes) for all the ten users (5 professors and 5 employees) during 4 days.

Table 1. Error rates (%) of the features of bioFACE

User	F. Identify (%)	F. Recognition (%)	GeoLocalization (%)	Workday reg. (%)
Professor 1	3	5	9	3
Professor 2	4	3	7	5
Professor 3	1	6	7	3
Professor 4	3	4	8	3
Professor 5	1	4	8	4
Employee 1	5	2	6	3
Employee 2	1	1	3	3
Employee 3	3	5	9	7
Employee 4	2	3	5	5
Employee 5	4	6	5	7
Average	3	4	7	4

Table 2. Elapsed time (minutes) to perform a workday registration

User	Day 1			Day 2			Day 3			Day 4		
	T1	T2	Diff									
Professor 1	5.85	1.24	4.61	5.94	1.66	4.28	7.01	2.25	4.76	6.52	1.33	5.19
Professor 2	9.25	1.30	7.95	6.91	1.55	5.36	7.49	1.66	5.83	8.44	1.82	6.62
Professor 3	9.62	1.19	8.43	7.56	1.21	6.35	9.06	1.04	8.02	7.50	1.30	6.20
Professor 4	7.25	2.75	4.50	6.53	1.94	4.58	9.10	1.64	7.45	9.50	2.83	6.68
Professor 5	9.53	2.18	7.35	5.71	2.84	2.87	5.15	1.02	4.13	8.83	1.56	7.27
Employee 1	7.32	1.75	5.56	5.45	2.83	2.62	8.35	1.55	6.80	6.24	2.92	3.32
Employee 2	9.39	1.04	8.35	5.69	1.64	4.05	6.32	1.17	5.15	5.18	1.55	3.63
Employee 3	5.66	2.68	2.98	6.73	2.13	4.59	8.67	1.71	6.95	6.46	1.03	5.43
Employee 4	6.27	1.12	5.16	8.93	1.79	7.14	9.00	1.65	7.35	9.82	1.83	7.98
Employee 5	9.64	2.11	7.53	5.25	2.41	2.84	5.44	1.93	3.51	6.88	1.36	5.52
Average	7.98	1.74	6.24	6.47	2.00	4.47	7.56	1.56	6.00	7.54	1.75	5.79

**Fig. 11.** Average of the workday registration time (in minutes) for all the ten users during 4 days

7 Conclusions

The mobile application bioFACE represent a faster and more efficient method to register the workdays of the UTEQ employees and professors. bioFACE converts the smartphone to biometric devices allowing register the workday starts and the workday exists being at any place inside of the university campus. With the aim of assess its performance, we have carried out some experiments measuring the features accuracy and workdays registration time performed by bioFACE. The obtained results are very promising, show a high accuracy level of its features (face detection, face identification and workday registration) and a faster and easy alternative method to perform the workdays registration.

In the future work, extensive experiments on more users under will be done to guarantee the performance accuracy of our application. Furthermore, we will implement machine learning algorithms (deep neural networks) embedded in the same application avoiding the requirement of internet access to use the Microsoft Face APIs.

References

1. Trewin, S., Swart, C., Koved, L., Martino, J., Singh, K., Ben-David, S.: Biometric authentication on a mobile device: a study of user effort, error and task disruption. In: Proceedings of the 28th Annual Computer Security Applications Conference, ACSAC '12, pp. 159–168. New York, NY, USA (2012) ACM
2. Hadid, A., Heikkila, J.Y., Silvén, O., Pietikainen, M.: Face and eye detection for person authentication in mobile phones. In: First ACM/IEEE International Conference on Distributed Smart Cameras, 2007. ICDSC'07. pp. 101–108. IEEE (2007)
3. Liu, H., Xie, X., Ma, W.-Y., Zhang, H.-J.: Automatic browsing of large pictures on mobile devices. In: Proceedings of the Eleventh ACM International Conference on Multimedia, MULTIMEDIA '03, pp. 148–155, New York, NY, USA, 2003. ACM
4. Helmy, J., Helmy, A.: Demo abstract: alzimio: a mobile app with geofencing, activity-recognition and safety features for dementia patients. In: 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 994–995 (May 2017)
5. Microsoft Face API. <https://docs.microsoft.com/en-us/azure/cognitive-services/face/>, 2018. [Online; accessed 20-May-2018]
6. Veridium: Biometrics Definitions. <https://www.veridumid.com/biometrics/>, 2018. [Online; accessed 20-May-2018]
7. Kanade, T.: Picture processing system by computer complex and recognition of human faces (1974)
8. Jebara, T.S.: 3d Pose Estimation and Normalization for Face Recognition. McGill University, Centre for Intelligent Machines (1995)
9. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: a general-purpose face recognition library with mobile applications. CMU Sch. Comput. Sci (2016)
10. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1701–1708 (2014)
11. Schroff, F., Kalenichenko, D., Philbin, J., Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823 (2015)
12. Ruffieux, S., Ruffieux, N., Caldara, R., Lalanne, D.: Iknowu - exploring the potential of multimodal ar smart glasses for the decoding and rehabilitation of face processing in clinical populations. In: Bernhaupt, R., Dalvi, G., Joshi, A., Balkrishnan, D.K., O'Neill, J., Winckler, M. (eds.) Human-Computer Interaction - INTERACT 2017, pp. 423–432. Springer International Publishing, Cham (2017)
13. Carr, N., McCullagh, P.: Geofencing on a mobile platform with alert escalation. In: Pecchia, L., Chen, L.L., Nugent, C., Bravo, J. (eds.) Ambient Assisted Living and Daily Activities, pp. 261–265. Springer International Publishing, Cham (2014)



Closer Look at Mobile Hybrid Apps Configurations: Statistics and Implications

Abeer AlJarrah^(✉) and Mohamed Shehab

University of North Carolina at Charlotte, Charlotte, NC 28262, USA
{aaljarra,mshehab}@uncc.edu

Abstract. We are witnessing a transition in the development of mobile operating systems from native custom architectures to web-based cross-platforms. There are several security implications of bringing the web code to smartphones. In this paper, we present a large-scale study that is centered on mobile hybrid apps configurations and permissions usage patterns. We study the platform configuration model and its' evolution. We find that while the platform is adding more security features, there is a demonstrable misconfiguration trend. The result of analyzing a set of 2111 hybrid apps uncovered several alarming observations. We have found that 80% of the apps are vulnerable to injection attacks because of an absence or a poor usage of the security model provided by the platform. We also detect a trend of keeping risky default configuration settings which results in having over-privileged apps that may expose device APIs to malicious code. On the system side, we realize that most of the apps have access to the platform's INTERNET and GEOLOCATION permissions. Google messaging is also recognized as the most widely used third-party service. In addition, we detect suspicious set of domains including spying, payment, Adware, and military that are white-listed. This study has the following contributions: (1) Systematizing our knowledge about mobile hybrid apps configuration model. (2) Providing an evidence of configuration misuse and developers tendency to use defaults. (3) Discussing possible reasons of misconfiguration practices and suggesting recommendations that address both the platform and the developer.

Keywords: Mobile hybrid apps · Configurations · Security · Cross-platforms · HTML5 apps

1 Introduction

The explosive growth in the number of smartphone users in the global market has led to a tremendous increase in the number of apps, thus an increasing apps'-driven revenue. By 2020, smartphone users are expected to reach 6.1 billion users¹ with gross annual revenue exceeding \$189 billion.² Thus, choosing which mobile application platform to adopt is becoming a strategic decision as it affects: the market share to target, cost of

¹<https://techcrunch.com/2015/06/02/6-1b-smartphone-users-globally-by-2020-overtaking-basic-fixed-phone-subscriptions/>.

²<https://www.smashingmagazine.com/2017/02/current-trends-future-prospects-mobile-app-market/>.

development, and the features supported. There are three main mobile development approaches for mobile apps, native (Android, iOS, etc.), web-based (HTML5) and hybrid. Mobile cross platforms or mobile hybrid platforms enable the development of mobile apps that run on several platforms using one code base written in standard web technologies (HTML5, JavaScript and CSS). The web-based code is wrapped in a thin native layer. This is supported by libraries that enable access to device native features such as CAMERA, GEOLOCATION, and NETWORK using JavaScript APIs.

Although native apps dominate the market, hybrid apps have succeeded to maintain a stable niche market since its' creation in 2008.³ Emerging trends such as enterprise mobile app development, virtual and augmented reality, dominance of the Internet-of-Things, and multi-platform ownership^{4,5} have played a strong role in motivating its popularity. In fact, hybrid apps are expected to "trump" native apps for several reasons, such as code reuse, dynamic updates, low development cost and potential growth.⁶

Vendors of hybrid platforms are continually improving in terms of features and tooling support. Examples of mobile hybrid platforms include React native, PhoneGap, Titanium, Xamarin, and Sencha Touch in addition to several other options. Adobe's PhoneGap is by far the fastest growing platform compared to all other ones.⁷ This platform uses Adobe's library Cordova which implements the bridge between web-based code and native container so that the app can access device resources using JavaScript APIs.

New technologies continue to be a favorable target to malicious code authors. Migrating the web code to smartphone devices have created new attack channels. Moreover, the platform's immature security model has made hybrid apps a low hanging fruit. Several studies [1, 2] have discussed injection attacks made possible through device sensors as the bridge provided by the hybrid platform libraries exposes the device APIs to JavaScript code access.

Cordova library have a set of configurations to control JavaScript code access to the system functionalities (APIs) in addition to other aspects of the app behavior. Configurations include what features to use and what domains are allowed to download resources from into the app. Ideally, the configuration model should serve as a security mechanism to control the external access to device resources and interactions with internal components. Proper configurations that follow the principle of *Least Privilege* can be the first line of defense against injection attacks. Strict settings can revoke privilege from malicious code, thus nullify its' impact. In this work, we shed the light on configurations and native permissions usage patterns of hybrid apps. We also investigate effectiveness of the security model provided by the platform.

³<https://www.g2crowd.com/categories/mobile-development-platforms>.

⁴<http://www.business2community.com/mobile-apps/2017-mobile-app-market-statistics-trends-analysis-01750346#HyFxJzqDdhgIGqkp.97>.

⁵<http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/>.

⁶<https://saucelabs.com/blog/hybrid-apps-and-the-future-of-mobile-computing>.

⁷<https://www.pixelcrayons.com/blog/mobile/cross-platform-mobile-development-trends-tactics-and-tools/>.

Contributions

We provide the following contributions:

- Discuss and analyze mobile hybrid apps configuration model and highlight its security related issues.
- Describe configuration model evolution.
- Identify and quantify configuration practices and explain the security implications.
- Provide analysis and suggest changes to the middle-ware and development process.

Organization

Section 2 explains the configuration model, its evolution, and the security limitations it suffers from. Then, in Sect. 3 we present the results of a market scan of 2111 apps configurations, permissions and security measures. Based on the results we further analyze the patterns and suggest recommendations in Sect. 4. We discuss related work in Sect. 5 and we finally conclude in Sect. 6.

2 Mobile Hybrid Apps Configurations

Cordova based applications configurations can be found in a file named **config.xml** file. It is a platform-agnostic XML file based on the W3C widget specification. Moreover, it allows developers to specify metadata about the application and also controls many aspects of an application’s behavior. This file is automatically created when creating the project itself. It can be edited manually or using certain commands provided by the command line interface (CLI). This file can be found at the root of the application as a global configuration source for different platforms’ apps. Configurations are parsed and the content is then translated into a set of features (see Fig. 1). Those features form the policy of accessing the device native APIs and define the app local and external interactions with web domains. For instance in a cordova Android-based app, certain components in the library, mainly PluginManager and CordovaActivity are responsible for enforcing the policy.

2.1 Configurations Items

Default configuration settings are shown in Fig. 2. Configurations are mostly common among different platforms. We summarize the description of the common configuration items in Table 1.^{8,9}

We categorize the configuration items into two categories based on their functionality and impact:

- Descriptive configurations: configurations related to providing meta-data about the app and the author. Examples include widget, name, description, author, and platform.

⁸<http://docs.phonegap.com/phonegap-build/configuring/>.

⁹<https://cordova.apache.org/docs/en/latest/configref/index.html/>.

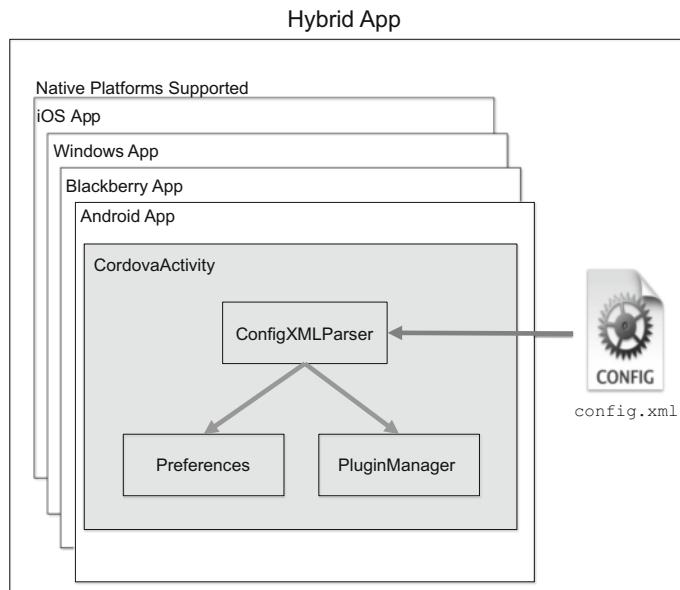


Fig. 1. Mobile hybrid app architecture

```

<?xml version='1.0' encoding='utf-8'?>
<widget id="com.example.hello" version="1.0.0" xmlns="http://www.w3.org/ns/widgets"
  xmlns:cdv="http://cordova.apache.org/ns/1.0">
  <name>HelloWorld</name>
  <description>
    A sample Apache Cordova application that responds to the deviceready event.
  </description>
  <author email="dev@cordova.apache.org" href="http://cordova.io">
    Apache Cordova Team
  </author>
  <content src="index.html" />
  <plugin name="cordova-plugin-whitelist" spec="1" />
  <access origin="*" />
  <allow-intent href="http:///*/*" />
  <allow-intent href="https:///*/*" />
  <allow-intent href="tel:/*" />
  <allow-intent href="sms:/*" />
  <allow-intent href="mailto:/*" />
  <allow-intent href="geo:/*" />
  <platform name="android">
    <allow-intent href="market:/*" />
  </platform>
  <platform name="ios">
    <allow-intent href="itms:/*" />
    <allow-intent href="itms-apps:/*" />
  </platform>
</widget>

```

Fig. 2. Default configurations as of version 8.x

- Access-control configurations: configurations related to controlling app access to device APIs, interacting with external domains and installed apps, and the starting page. Examples include content, access, allow-navigation, allow-intent, and feature.

Descriptive configurations are simple and self-explanatory as explained in Table 1. We focus on the app's access-control configurations as they can have a substantial impact on the app security and behavior. We explain the main access-control configurations:

Table 1. Description of the common configuration items

Item	Description
Widget	Root element of the config.xml document
Name	Specifies the app's formal name, as it appears on the device's home screen and within app-store interfaces
Description	Specifies metadata that may appear within app-store listings
Author	Specifies contact information that may appear within app-store listings such as author's email and website
Content	Defines the app's starting page in the top-level web assets directory. The default value is index.html, which also appears in a project's top-level www directory
Access	Defines the set of external domains the app is allowed to communicate with. The default value allows it to access any server
Allow-navigation	Controls which URLs the WebView itself can be navigated to. Applies to top-level navigations only
Allow-intent	Controls which URLs the app is allowed to ask the system to open. By default, no external URLs are allowed
Preference	Sets various options as pairs of name/value attributes
Feature	Specifies the device APIs the app is allowed to use. Examples include camera, geolocation, etc.
Platform	Specifies preferences or other elements specific to a particular platform

Plugins APIs: Hybrid apps can access device native features such as CAMERA, through a JavaScript API provided by the library. A developer needs to include the plugin name in the configurations using the tags <feature> or <plugin>, depending on the version. Developers can use command lines to add plugins which also updates the native side automatically to include the required system permissions and the native code needed to use the plugin. For instance, including the CAMERA API to an Android app, changes the AndroidManifest file to include the permission: 'android.permission.CAMERA' and the config file to include a declaration of the camera plugin. Earlier versions of the library include by default a set of 13 core plugins (examples include CAMERA, INTERNET, GEOLOCATION) in the configurations along with all required system permissions.

Domain White-Listing for Accessing Network Resources: that is the security mechanism that controls access to external domains over which the app have no control. Cordova provides a configurable security policy to define which external sites can be accessed. The library "recommends" developers to customize the white-list to allow access to specific domains and subdomains before moving the app to production. For Android, Cordova's security policy (as of its 4.0 release) is

extensible via a plugin interface cordova-plugin-whitelist, as it provides better security and configurability than earlier versions of Cordova. The library configurations use `<access>` element within the app's config.xml file to enable network access to specific domains. For example, to access domain XYZ.com, access should be configured as: `<access origin="http://XYZ.com">`. The default value for this configuration item is “*” which allows access to **any domain**. To overcome this, with the release of version 5.0, the library recommended using content security policy (CSP) on a page level to mitigate code privilege and to control triggering malicious code. Apps developed using the versions before CSP is introduced can be easily vulnerable to any malicious domain if not configured properly. Moreover, up to the latest version, there is a limitation in the enforcement of this item, as described in the library documentation.¹⁰ The issue is that white-listing alone cannot block network redirects from a white-listed remote website (i.e. http or https) to a non-whitelisted website which also may jeopardized the app to several attacks. Examples include injection attacks, XSS, and untrusted JavaScript code inclusion.

Content Security Policy: A fundamental component of the platform security model. Its main purpose is to mitigate XSS and injection attacks. It controls which network requests (images, XHRs, etc.) are allowed to be made (via webview directly) on a page level. CSP is set by including a `<meta>` tag listing the rules of content download. A CSP consists of a set of policy directives and corresponding values. Examples of directives include default-src, script-src, and img-src. These are responsible of controlling default content, script and images respectively. default-src is a default directive that applies in case other directives are not specified. Policies help controlling allowed URL sources, executing inline code, and enabling eval() function. In Cordova, a default CSP policy: (1) disables eval() and in-line scripts style (2) allows network requests of types CSS, AJAX, and frame (3) allows only local code execution. CSP is added to the library starting version 5.0.0 and is included by default in the generated index.html file.

Domain White-Listing for Intents: Like any native app, a hybrid app may interact with other system components including other installed apps. In Android, this is handled by Intents which is designed to manage inter-application communication. Hybrid apps may access other intents such as dialer, email or messages via URIs and may also pass data to these intents. Thus, cordova library provides rules via allow-intent to govern what intents can be called. As it is demonstrated in Fig. 2, default configurations allow interaction with the following apps: dialer, messages, maps, market, and the browser. This white-list applies to calls via hyperlinks and the method window.open().

Domain White-Listing for Navigation: controls which URLs the WebView itself can be navigated to. Applies to top-level navigations only. On Android, it also applies to iframes for non-http(s) schemes. By default, navigations are allowed only to local files. To allow other URLs, the item `<allow-navigation>` can be used to set the list of URLs.

¹⁰<https://cordova.apache.org/docs/en/latest/reference/cordova-plugin-whitelist/>.

2.2 Configurations Evolution

Initially in version 1.5.0, the library supported 6 platforms (Android, Blackberry, iOS, Symbian, WebOS and Windows Phone) and offered APIs to access 13 device resources (Accelerometer, Camera, Capture, Compass, Connection, Contacts, Device, Events, File, Geolocation, Media, Notification, and Storage). The most recent version of the library (to the time of writing this paper) is 8.x. This version supports 7 platforms (adding WP8), ships with 16 core APIs in addition to supporting a repository of 3646 plugins developed and shared by the community¹. We track and report the changes on the library starting from version 1.5.0 till the most recent version in Table 2. We only highlight the changes related to configurations and security in addition to any security awareness effort made by the library. In version 1.5.0, all core plugins are included by default in the configurations. This would also effect the native side of the app because it would be configured to include the system permissions needed to use all the included APIs. An Android app for instance would have 15 system permissions by default, including CAMERA, ACCESS_COARSE_LOCATION, and INTERNET. At that time, there was no consideration of any security principle such as controlling app interaction with external domains. It was till the release of version 1.8.0., the concept “Domain White-listing” was introduced to support controlling access to

Table 2. Cordova configuration history summary

Version	Change(s)
1.5.0	<ul style="list-style-type: none"> • Core plugins are included by default • Plugins declarations in file plugin.xml • No security model
1.8.0	<ul style="list-style-type: none"> • Domain whit-listing is introduced. • Documentation added a section about • Access rules in file cordova.xml
1.9.0	CLI is shipped with the library
2.0.0	<ul style="list-style-type: none"> • Plugin developers guide is introduced • Access rules in file config.xml
2.8.0	Documentation added a privacy guide
3.0.0	<ul style="list-style-type: none"> • No plugins are included by default • CLI enables adding plugins • Documentation added configuration reference • White-list plugin is introduced
3.5.0	Documentation added security guide
5.0.0	<ul style="list-style-type: none"> • Content-Security Policy (CSP) is added • White-list plugin introduces <allow-navigation> • White-list plugin introduces <allow-intent>

external domains.¹¹ At that time, domain white-listing is implemented for 3 platforms (Android, iOS, and Blackberry). Access rules to outside domains are specified in an XML format using the element <access>. It was set to origin="*" which allows access to any server by default. These rules are stated in a file named cordova.xml. Later in version 1.9.0 the library presented Command Line Interface (CLI) which is a standard set of command-line tools. The purpose is to simplify interaction with the library and make it easier to develop cross-platform applications. It was supported for 3 platforms, Android, iOS, and Blackberry. In version 2.0.0 the library started to encourage developers to develop their own plugins, hence the library released a plugin development guide in their manual; explaining the architecture of the Cordova API on both sides JavaScript and native. In version 2.8.0, the library dedicated a section in their documentation for privacy. The privacy guide at the time encouraged developers to follow practices to preserve users' privacy such as, having a privacy policy, avoiding collecting users' information, and giving users more control on allowing collecting information and accessing device APIs. Version 3.0.0 had a major change compared to previous versions. First, plugins are not included by default. Second, adding plugins is supported commands through CLI. Third, domain white-listing policy is implemented in the file config.xml.

Not until the release of version 3.5.0, a security guide was added in the documentation.¹² The guide addressed security issues that are discussed in research [3–5]. It focused on several security breaches including bringing awareness to issues related to setting access to specific servers rather than allowing access to all servers.

To provide more modularity and extensibility to the security model, the library provided white-list plugin for Android and iOS in version 5.0.0. This plugin provides better security and configurability compared to earlier versions of Cordova as it added more security related configurations such as <allow-navigation> and <allow-intent>. In addition to that, the library implemented Content Security Policy (CSP) to allow more control of content downloads on a page level.

2.3 Adversary Model

There are several adversary models that are relevant to mobile hybrid apps. We discuss two models that proper configurations would have helped curb the impact if not abort the attack.

Injection Attacks. Malicious code can be injected to mobile hybrid apps through several channels such camera, WIFI access points and simple text fields [2, 6].

Compromised Third Party. Including JavaScript from remote servers is a common practice among web developers. Remote code have the same privilege as local code which means that compromising these servers would render the app vulnerable to remote attacks.

¹¹<https://cordova.apache.org/docs/en/latest/reference/cordova-plugin-whitelist/>.

¹²<https://cordova.apache.org/docs/en/3.5.0/guide/appdev/security/index.html>.

In the previous models proper configurations can be the first line of defense as follows: (1) proper domain-whitelisting can prevent malicious URLs to download malicious payloads. (2) proper CSP can prevent triggering malicious injected code to mention limiting code access on a page level. (3) using minimal plugins APIs declarations minimize the attack surface thus the damage if malicious code is activated.

2.4 Configurations and Security Consideration for Hybrid Apps

Efforts to make the library more secure (driven either by the research community or the library itself) are evident. Yet, the library still suffers security limitations in their configuration model.

- (1) *Coarse-Grained*: Configuration rules are enforced on the app regardless the context. Policy fails to address context-aware details such as location and time. Moreover, configurations are global to the whole app. The app consists of one or more pages or states. Each of which requires different permissions based on the app behavior in a specific state. However, the configurations grant permissions to the app as a whole.
- (2) *Risky defaults*: Default settings of the configurations are liberal. Most default values grant non-restricted access. Values, such as “*” indicating non-restricted access, is common. For instance, if we examined the default settings for domain white-listing for network resource access is set to “*” which allows access to any server—in the absence of CSP. One consequence is the capability of including remote JavaScript from remote untrusted servers into the app. Moreover, intents white-listing default settings allows access to several build-in apps including dialer, SMS, maps, and browser. Risky defaults is a serious issue given the high probability that developers are likely to keep default values as is.
- (3) *Inefficient Security Model*: Content Security Policy is added to control access on a page level and to recover limitations of domain white-listing. However, this security model is not efficient because its usage is not enforced. In other words, the absence of a CSP does not prevent building the app, it would only log a warning message. This may not necessarily draw developers’ attention to include it at all. Not having a proper CSP and keeping the network resource setting to default, shall expose device resources to any untrusted server. This jeopardizes the security of the app to enable injection attacks.

3 Related Work

3.1 Hybrid Mobile Platforms Security

Security of mobile hybrid platforms is discussed in the literature in several parallel tracks. Once is mainly focused on security issues related to the platform itself. Issues resulting from the improper implementation or enforcement of security principles. Martin et al. [7] are the among the first to systematically evaluate the PhoneGap/Cordova platform security model. They analyze the software stack created

by the platform and demonstrate that it doesn't properly compose the access-control policies governing web code and local code. They demonstrate that the platform origin checks are not working properly which have to lead to documenting the vulnerabilities CVE. They also introduce the difference between accessing device features through a bridge and loading web content (like ads) into the app which leads to the introduction of the new configuration item (allow-navigation). Kapil [8] discuss the issue of having a coarse-grained permission model for hybrid apps. He is more interested in the native permissions granted to the hybrid app. He argues that the current native permission model of accessing device sensors and data follows an all-or-nothing approach and does not satisfy the Least-Privilege principle. Thus, he suggests the modification to the platform to implement fine-grained and context-dependent policies. Similarly, Shehab and AlJarrah [9] have addressed the coarse-grained configuration scheme and have proposed a page-level policy to be set to every plugin declaration such that a plugin is only accessed from specific pages determined by the developer. Phun et al. [10] have recognized the library limitation of tracking the JavaScript code origin, thus the incompetence of setting policies per origin which may have serious implications if malicious third-party origins including advertising domains are white-listed. Their solution doesn't require changing the platform, instead, they provide a JavaScript API that enables setting principal-based policies. Another track focuses on identifying attacks vectors applicable to hybrid apps and proposing solutions. Jin et al. [1] has identified code injection channels that apply to mobile devices such as 2D barcodes, RFID tags, media files, the ID field of Bluetooth devices and Wi-Fi access points. Code injection causes attacks similar to XSS attacks but is more damaging because they are running on the mobile platform that offers bridges to access devices features. Along the same line, Jin et al. in [2] have identified more attack channels such as content providers, file system, and intents. Also, injection using HTML5 textbox input type along with "document.getElementById("TagID").value" [6]. They [1, 2, 6] also proposed solutions to detect injection attacks that all involve changing the platform. More work has been directed toward systematizing security issues related to mobile hybrid apps. Hale and Hanson [11] have defined common hybrid attack vectors and provided a testbed platform for analyzing vulnerabilities. Yan et al. [12] provide a system that use static and dynamic analysis modules to discern potentially vulnerable apps.

3.2 Abusing Privilege in Mobile Apps

Over-privileged apps is a common problem in mobile apps. Granting privileges to apps through system permissions or relaxed configurations may expose the user and the device to vulnerabilities and bugs which may compromise the device and users data. In the context of Android apps, permissions controls privilege. Felt et al. [13] have indicated that apps usually have more permissions than what they actually need. They developed a tool for detecting over-privileged apps that perform static analysis on the code and generate the maximum set of needed permissions. According to them, API documentation errors and lack of developer understanding are the main reasons for such issue. Similarly Bartel et al. [14] have highlighted the same problem and proposed to solve it by computing a mapping of permissions based on the code. More complex solutions have been proposed to help handle insecure data access permissions. Zhu

et al. [15] have designed a mobile app recommender system which automatically detects and evaluate the security risk of an app based on apps' popularity and the set of permissions used. Similar to what Sarma et al. [16] have discussed in using permissions an app requests to better inform users whether the risks of installing an app if commensurate with its expected benefit. Wang et al. [17] have proposed a framework to quantify security risk assessment of both android permissions and apps based on permission request patterns from benign apps and malware.

4 Scan Results of the Configurations of Market Apps

4.1 Data Collection

We compile a set of cordova based package names from different resources including previous research work² in addition to package names manually extracted from the platform website³. To ensure that the apps are still active in Google Play market, we have implemented and used an automated tool to search the app name in Google Play, install the apk on a mobile device, then copy the apk file from the device to a computer. We started with a list of 20,000 names and ended up with a 2111 app; rest are either not found or do not install properly.

4.2 Starting Page <src>

A hybrid app consists of one or more pages. Pages can be hosted locally inside the app folder or remotely. We have found that most of the apps (98%) start from a local page. The rest of the apps (2%) have the following settings:

- 17 apps are set to start from a remote page that uses http connection
- 3 apps are set to start from a remote page that uses https connection.

4.3 Network Resources Access <access>

This is a key setting as it indicates the white-listed domains that can download resources (including JavaScript files) into the app. The remote JavaScript code from a white-listed domain has the privilege to access device APIs and execute on the device just like the local code. The developer may use one or more rules to white-list the domains. The library adds a default rule of a “*” which restricts no domain. Developers may remove it and specify domains according to the app needs. We first report how many rules are found for this configuration item. Results of scanning the data set indicate that most apps (96%) use 1–5 rules (see Fig. 3). A maximum number of rules per app reached up to 29 access rules. In terms of rules’ values, we categorize them into the system provided settings that are suggested by the platform and custom settings that developers set depending on the app specific access needs. Note that because of certain versions of the library used this item to control navigation and intents, system provided settings may include values other than “*”, these values are to control intents and navigation in the absence of corresponding configurations items in earlier versions of the library.

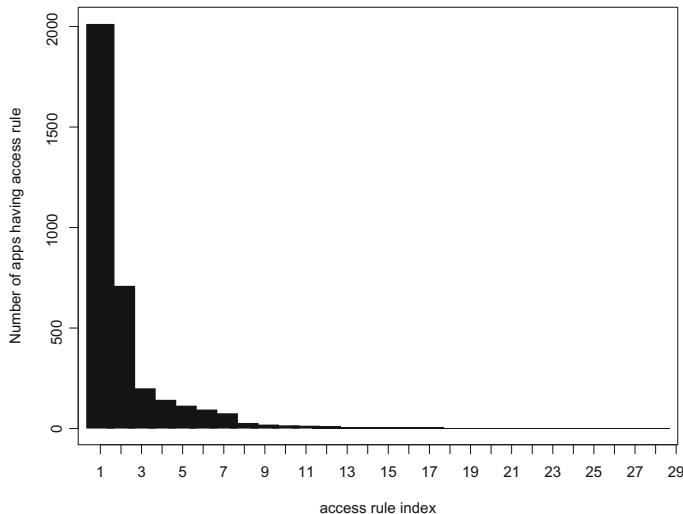


Fig. 3. Access rules usage

Figure 4 demonstrates the percentages of apps using default values provided by the platform. The value “*” has the highest percentage of occurrences as **67%** of the apps include this value. This means that almost **1342** apps are configured to accept downloads from *any* server. Localhost value which is found in 37% indicates that the app can accept accesses from local files. This low percentage can be attributed to the fact that setting the access to “*” implies that it can download from itself so developers may not need to add it. Values such as “tel:/*” and “http:///*/*” are found because, in versions before 5.0.0, the items allow-navigation and allow-inent were not added yet. Hence, <access> was the only configuration item related to domain white-listing regarding network resource access, intents, and navigation. Other values that are related to intents and navigation are found in less than 5%. Developers may also use specific domains which is a recommended practice by the library. Yet, as it is demonstrated in Fig. 5, custom URLs are found in less than 5% of the apps. We categorize URLs based on their purpose. Social media URLs such as Facebook, Twitter, and Youtube are found in 3.4% of the apps. Next, Google APIs, such as “*.googleapis.com”, “<http://googleanalytics.com>”, and “https://play.google.com/store/apps/*”, are found in 2.4%. We have also found that almost 25 apps (1.1%) white-listed Gstatic domain which according to the official documentation¹³ is needed for the TalkBack function. This function is offered by Google as an accessibility service that helps vision-impaired users interact with their devices. Moreover, we have identified several URLs (“https://*.netspend.com”, “https://*.mycontrolcard.com”, “https://*.wuprepaid.com”, “https://*.acebusinessselectcard.com”, “https://*.paypalprepaid.com”) related to online

¹³<https://cordova.apache.org/docs/en/latest/reference/cordova-plugin-whitelist/>.

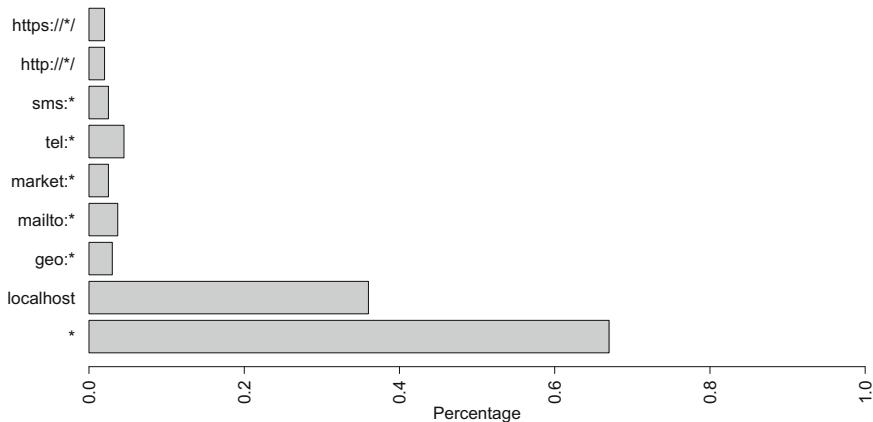


Fig. 4. Network resource access: platform provided settings

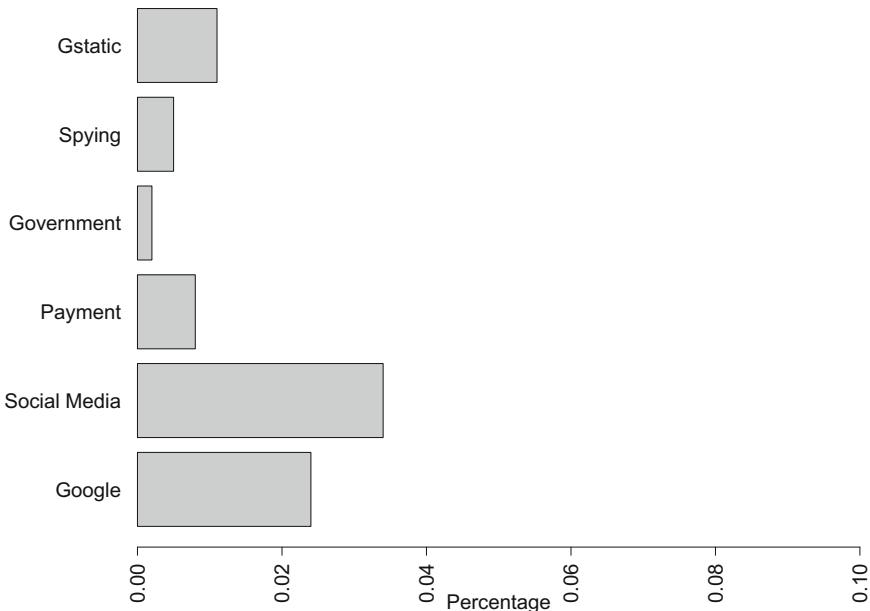


Fig. 5. Network resource access: custom settings

payment/banking in 0.8%. Additionally, spying URLs (“https://*.spykemobile.net”, “https://*.iesnare.com”) are found in 0.5%. Finally, we identify military and government domains in 0.2% of the apps.

4.4 Domain Whitelisting for Intents<allow-intent>

Starting from version 5.0.0, this configuration item is added to control the app inter-process communication between the app itself and other apps and also what data can be passed. We have found that only 15% of the apps contained settings related to intents' white-listing. For that set, we demonstrate the configuration values distribution in Fig. 6. We can see that among those apps, 39% white-list "geo:/*" which offers maps service, followed by 35% using "*" which allows interaction with all apps. The intent that white-lists Google play ("market:/*") is found in 13%. Then, the rest of the values are found in less than 5%.

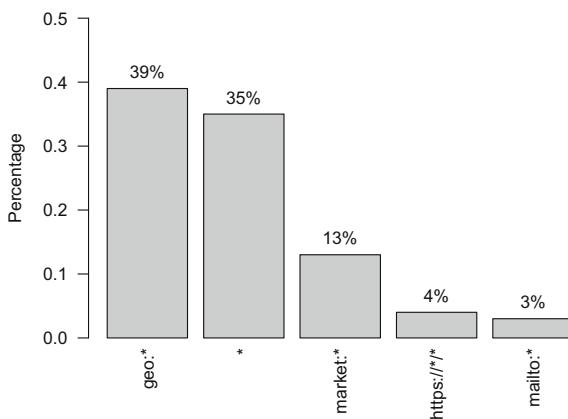


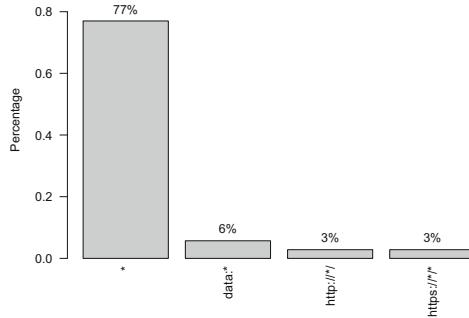
Fig. 6. Intents white-list settings

4.5 URL Navigation <allow-navigation>

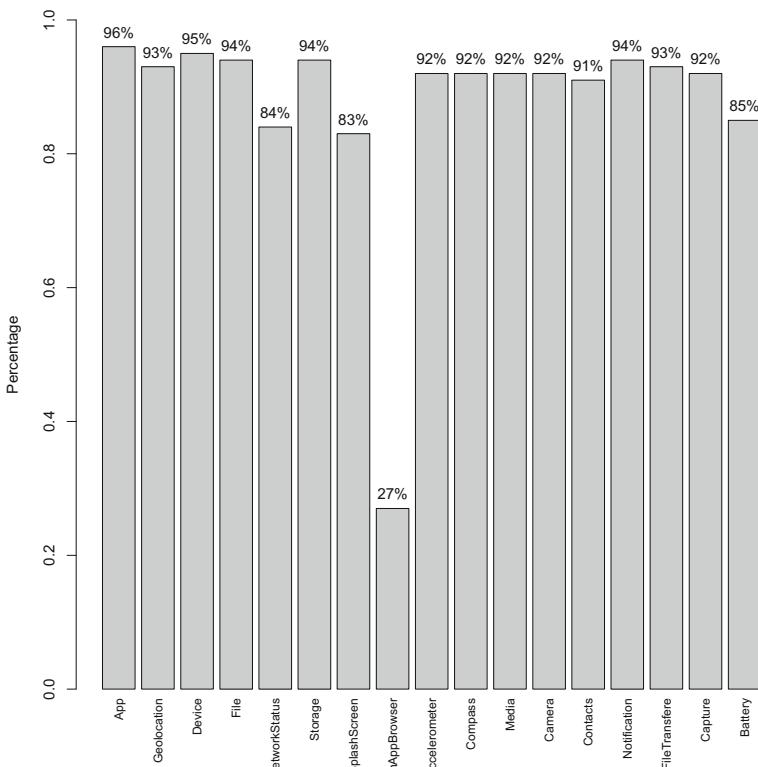
We have found that only 10% of the apps contain navigation white-listing settings. In that set we scan the settings and demonstrate the distribution in Fig. 7. For this configuration item, there is no default configurations. This means that any added configuration is added by the developer. The value "*" is found in 77% of the apps which allows navigation to any protocol/domain. This might explain the low percentages of all other values, such as "http:///*/*", "https:///*/*" and "data:///*/*", since the value "*" include them all.

4.6 Plugin Usage <plugin|feature>

Having a plugin declaration means that the corresponding device functionality can be accessed by the JavaScript code. Thus, developers should ensure to include only the plugins needed by the app. We scan plugins configurations tags and divide them into core and custom. It is important to highlight that plugins declaration in the configurations does not necessarily mean their usage. The plugin tag may exist because it is added by default, which may explain the high percentages of default core plugins usage

**Fig. 7.** Navigation whitelist settings

as depicted in Fig. 8. This also explains why 88% of the apps use core plugins. This demonstrate a wide attack surface that malicious code can abuse. On the other side, 80% of the apps use custom plugins in addition to the core ones. We scan and classify custom plugins based on their purpose. Figure 9 shows custom plugins usage percentages. Plugins related to accessing device features are found in 79% of the apps.

**Fig. 8.** Core plugins APIs

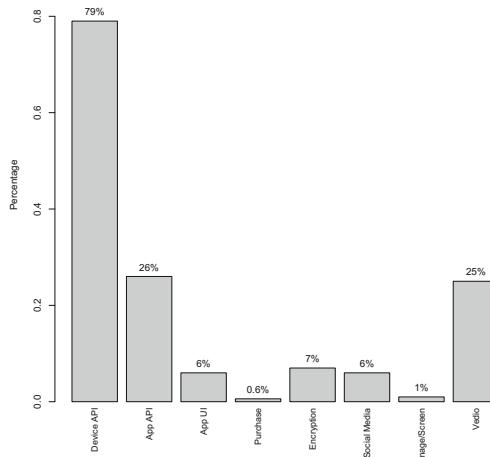


Fig. 9. Custom plugins APIs

Examples of device-related plugins are bar-code scanners, location-based, and SMS plugins. Application APIs mostly related to web browsing are found in 26%. Video related APIs such as video players are found in 25% of the apps. We also identify several encryption-APIs (Crypto, i4crypt, and cryptUtil) in 7% of the apps. Application user interfaces interaction APIs such as push notifications and progress dialogues are found in 6%. Same percentage for social media APIs such as facebook, twitter and social sharing. Advertisement APIs and screen image APIs are detected in almost the same percentage (1%).

4.7 Native Permissions

In a hybrid app, using a plugin may require one or more platform permissions on the native side. For instance, using the CAMERA plugin on an Android app requires the app to use the permission “`android.permission.CAMERA`”. Including a plugin in the configuration without using the proper corresponding native permissions shall void any call to that plugin because the native platform won’t be able to fulfill the request. Permissions in Android is a security mechanism by which the platform protects access to sensitive device data and sensors. Developers need to include proper permissions to use certain functionalities. In addition to platform defined permissions, a developer may define her own custom permissions to control access to a component from other apps. Android permissions are divided into several protection levels. The two most common levels are normal and dangerous permissions. Normal permissions are used to grant access to data or resources outside the app’s sandbox, but where there’s very little risk to the user’s privacy or the operation of other apps. However, dangerous permissions grant access to data or resources that involve the user’s private information, or could potentially affect the user’s stored data or the operation of other apps [18]. We are particularly interested to scan the permissions used to see if there exists a particular set of permissions that are often used. We are also interested to check if these native

permissions are the required permissions for the default plugins. To extract native permissions, we scan the file AndroidManifest.xml file. We have found that 90% of the permissions used are standard Android permissions. The 10% are developer's customized permissions to use third party services such as google cloud messaging and In-app billing. For standard permissions we show in Fig. 10 the top permissions that are used color coded based on the protection level. Blue bars represent normal permissions while red ones represent dangerous permissions. Solid bars are for permissions needed for core plugins, while striped ones are permissions needed for other purposes. The permission to use the INTERNET is used in almost all the apps. Then comes a set of dangerous permissions. Examples include those related to accessing location such as ACCESS FINE LOCATION and ACCESS COARSE LOCATION. Also, the WRITE EXTERNAL STORAGE permission that is needed to save data in the device, the permission READ PHONE STATE that allows read-only access to phone state, including the phone number of the device, current cellular network information, the status of any ongoing calls, and a list of any Phone Accounts registered on the device. While most permissions are those needed to use the core Cordova plugins, we found are two permissions; WAKE LOCK (32%) and RECEIVE BOOT that there COMPLETED (15%) that are not needed for the core plugins, yet they are among the top used permissions. Both are normal permissions. The first one allows using PowerManager WakeLocks to keep processor from sleeping or screen from dimming. The second one allows an application to receive the broadcast after the system finishes booting [19]. We also found that the permission RECORD VEDIO that exist in 10% of the apps does not actually exist on Android permission list. Its' existence is probably due to being a default permission added by the library in previous versions. Regarding custom permissions, we have found that the most used permission is the "com.google.android.c2dm.permission.RECEIVE" which is used in 36%. This permission is needed

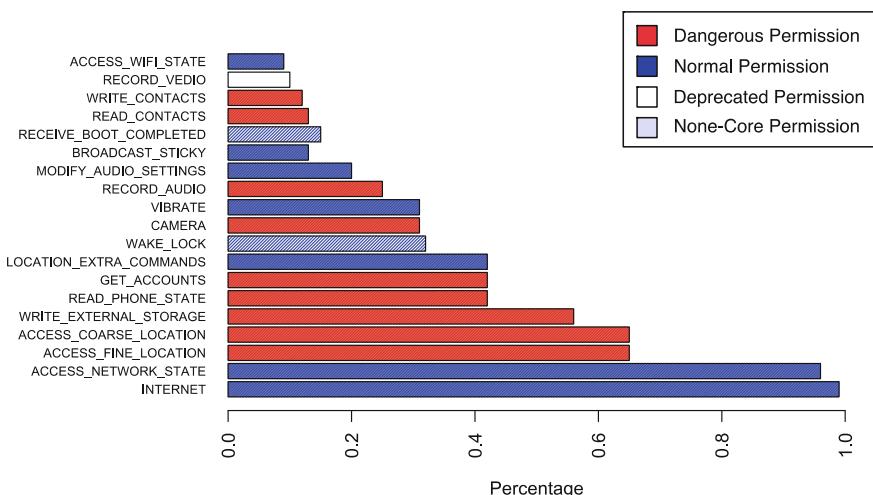


Fig. 10. Platform native permissions usage

to use Google's cloud messaging service. Next permission is the in-app billing permission "com.android.vending.BILLING" which is used in 2%.

It is observed that the permissions that are added by default are found in high percentages (average of 91%). This also supports the reflection that developers tend not change the default settings. This is also an indication that these granted permissions exist not because they are necessarily needed, rather because they are simply added by default. This also indicates an issue of permission misuse and violating Least-Privilege principle.

4.8 Content Security Policy (CSP)

Unlike previous configuration items, CSP is not set in the config.xml. It is set in the html pages. The purpose of setting CSP per page is to control content download. We scan the index page and extract CSP meta tags—if any. Results depicted in Fig. 11 show that majority (75%) did not contain any CSP. Moreover, 10% of the apps contained a commented out CSP. These are probably the default generated CSP tags but for some reason developers decided to comment them out. We also identified empty CSP rules in 1% of the apps. This leaves only 15% of the apps having an active CSP. This means that 85% of the apps have no effective CSP enforcement. Thus, no control of content download on a page level which means that these apps are vulnerable to injection attacks. To inspect the way policies are specified in the 15% of the apps (178 policy), we break down the policies and report results in Table 3. Each percentage is the number of occurrences of the policy found for the corresponding policy directive divided by the total number of policies (178). For instance, 79% of the policies allows content download from any domain. Considering the way policies are set, we argue that developers are using liberal values in setting their policies. Having * as the most commonly used value rather specifying URLs is an example. Moreover, allowing unsafe-inline and unsafe-eval enables malicious code execution. Strict policies such as using 'self' or specifying URLs are the least used. A combination of * as a source of script download, allowing unsafe-inline or unsafe-eval voids the whole purpose of having CSP. This combination is found in 62% of the policies.

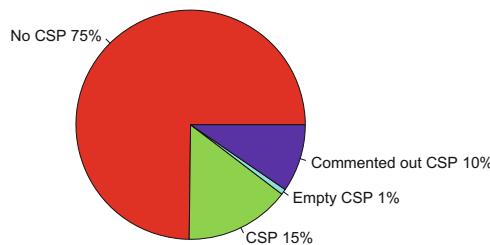


Fig. 11. Content security policy usage

Table 3. CSP break down

Policy directive	Values				
	* (%)	Unsafe-inline (%)	Unsafe-eval. (%)	Self only (%)	URLs (%)
default-src	79	65	70	2	2
script-src	63	59	1	11	0
style-src	5	27	0	24	0
img-src	6	1	0	2	1
media-src	9	0	0	0	0
frame-src	1	0	0	0	0

5 Analysis and Recommendations

5.1 Developers' *BAD* Practices

There has been several studies discussing developers tendency to errors that may effect the security of the applications [20, 21]. Human aspect in developing secure code has been receiving a lot of attention in literature. The increasing trend towards appification allows inexperienced layman developers to produce complex and sensitive apps. Developers make mistakes due to lack of understanding or an attitude that implies that it's not their responsibility. Security and configurations are usually a low priority for most developers. This can be due to the common development conditions including heavy loads and tight deadlines. Default configurations for this library are liberal. Although this may be considered usable by minimizing developers involvement in configurations. But, it comes at the expense of security. In this work, we quantify for each configurations item, the total percentage of default settings used (summarized in Table 4).

Table 4. Default settings usage

Configuration item	Default used (%)
Starting page: src='index'	95
Network resource whitelisting: origin='*'	67
Plugins (Core APIs)	91

5.2 Recommendations

- (1) *Improve Documentation:* A study on Android developers have examined how information resources that are used by developers to produce code, is negatively impacting the security of the code, as most developers use online resources such as StackOverflow rather than the platform standard manual [22]. We believe this is a similar situation. Unfortunately, the platform documentation fails to provide an

easy to follow content. We recognize that the documentation has improved substantially by including separate sections discussing security and privacy implications. Still, the technical documentation related to CSP for instance is brief and not easy to follow.

- (2) *Supporting Tools*: The process of configuring the app is manual and error-prone. Developers add plugins they think they need. They may also not change default settings especially if the app is working properly. There is an evident problem of a gap between developers conceptual understanding of security and their attitudes regarding their responsibility [20]. Thus, automated tools are needed to support developers while developing apps to help provide more secure code and to help detect over-privileged and misaligned configured ones. For Android apps, several tools have been designed to detect over- privileged and risky apps on Android and suggest the needed permissions based on the apps needs [13–17]. Similar tools can be designed to detect poorly configured hybrid apps. Static and Dynamic analysis of the JavaScript code would yield an indication of the required configurations, thus having more aligned configurations with the app needs.
- (3) *Fine-Grained & Context-Aware Configurations*: Smart-phone users prefer to have more control in terms of what resources are being accessed. Most users would not mind being confronted with security decisions [23]. Android adapted to this need by changing the way permissions are requested. Starting from Android 6.0, users grant permissions to apps while the app is running, not when they install the app.¹⁴ This gives the user more control over which permissions to grant and when. A similar approach can be adopted by hybrid apps. Hybrid apps may ask the user at run-time to grant access to a plugin. This may give the user not only more control but also provide a context so that the user may take a better decision. This is especially important if we considered injection attacks that may change app behavior.

6 Conclusion and Discussion

In this work, we discuss hybrid apps configuration model. We base our work on Cordova library as the open nature of the project offers an insight into its design, and allows for the introduction of extensions to its security architecture. Considering the current configuration model and analyzing a set of 2111 apps, we bring attention to the following:

- Security limitations of the configuration scheme including risky default values and coarse-grained model.
- Ineffective adoption of the security model, CSP in specific (80% of the apps are vulnerable and the rest have poorly written policies).
- Tendency to keep default settings on both native side and web side resulting in having over-privileged apps.

¹⁴<https://developer.android.com/training/permissions/requesting.html>.

This work provides a founding evidence of insecure configuration issues that may substantially compromise hybrid apps' security.

The goal of this work is to highlight the importance of the security decisions made by these libraries. Developers mainly rely on the default security settings and are very unlikely going to change them. Another goal is to emphasize the lack of security awareness among developers. Future research should investigate solutions to help developers have more aligned configurations and more secure content policies. Moreover, further work is needed to change the way security model is enforced. A main limitation of this work is our inability to discern the version of the library used while developing the apps. If we were able to recognize the version, our analysis would have been more accurate in terms of explaining the existence/non-existence of certain values. However, we argue that there is a stressing need to uncover developers practices in general to expose the gap between what libraries recommends and promises users in terms of maintaining security and privacy and how developers actually use configure apps and use these libraries features.

References

1. Jin, X., Luo, T., Tsui, D.G., Du, W.: Code injection attacks on html5-based mobile apps (2014). arXiv preprint [arXiv:1410.7756](https://arxiv.org/abs/1410.7756)
2. Jin, X., Hu, X., Ying, K., Du, W., Yin, H., Peri, G.N.: Code injection attacks on html5-based mobile apps: characterization, detection and mitigation. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 66–77. ACM (2014)
3. Luo, T., Hao, H., Du, W., Wang, Y., Yin, H.: Attacks on webview in the android system. In: Proceedings of the 27th Annual Computer Security Applications Conference, pp. 343–352. ACM (2011)
4. “Phonegap platform security,” <https://github.com/phonegap/phonegap/wiki/Platform-Security>
5. “Html5 security cheat sheet,” <https://www.owasp.org/index.php/>
6. Chen, Y.-L., Lee, H.-M., Jeng, A.B., Wei, T.-E.: Droidcia: a novel detection method of code injection attacks on html5-based mobile apps. In: Trustcom/BigDataSE/ISPA, vol. 1, pp. 1014–1021 (2015). IEEE
7. Georgiev, M., Jana, S., Shmatikov, V.: Breaking and fixing origin-based access control in hybrid web/mobile application frameworks. In: NDSS Symposium, vol. 2014. NIH Public Access, p. 1 (2014)
8. Singh, K.: Practical context-aware permission control for hybrid mobile applications. In: International Workshop on Recent Advances in Intrusion Detection. Springer, Berlin, pp. 307–327 (2013)
9. Shehab, M., AlJarrah, A.: Reducing attack surface on Cordova-based hybrid mobile apps. In: Proceedings of the 2nd International Workshop on Mobile Development Lifecycle, pp. 1–8. ACM (2014)
10. Phung, P.H., Mohanty, A., Rachapalli, R., Sridhar, M.: Hybridguard: a principal-based permission and fine-grained policy enforcement framework for web-based mobile applications

11. Hale, M.L., Hanson, S.: A testbed and process for analyzing attack vectors and vulnerabilities in hybrid mobile apps connected to restful web services. In: 2015 IEEE World Congress on Services (SERVICES), pp. 181–188 (2015). IEEE
12. Yang, L., Cui, X., Wang, C., Guo, S., Xu, X.: Risk analysis of exposed methods to javascript in hybrid apps. In: Trustcom/BigDataSE/I SPA, pp. 458–464. IEEE (2016)
13. Felt, A.P., Chin, E., Hanna, S., Song, D., Wagner, D.: Android permissions demystified. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, pp. 627–638. ACM (2011)
14. Bartel, A., Klein, J., Le Traon, Y., Monperrus, M.: Automatically securing permission-based software by reducing the attack surface: an application to android. In: Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering, pp. 274–277. ACM (2012)
15. Zhu, H., Xiong, H., Ge, Y., Chen, E.: Mobile app recommendations with security and privacy awareness. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 951–960. ACM (2014)
16. Sarma, B.P., Li, N., Gates, C., Potharaju, R., Nita-Rotaru, C., Molloy, I.: Android permissions: a perspective combining risks and benefits. In: Proceedings of the 17th ACM Symposium on Access Control Models and Technologies, pp. 13–22. ACM (2012)
17. Wang, Y., Zheng, J., Sun, C., Mukkamala, S.: Quantitative security risk assessment of android permissions and applications. In: IFIP Annual Conference on Data and Applications Security and Privacy, pp. 226–241. Springer, Berlin (2013)
18. “Android api guide <permission>,” <https://developer.android.com/guide/topics/manifest/permission-element.html>
19. “Android normal permissions,” <https://developer.android.com/guide/topics/permissions/normal-permissions.html>
20. Xie, J., Lipford, H.R., Chu, B.: Why do programmers make security errors? In: 2011 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp. 161–164. IEEE (2011)
21. Xie, J., Chu, B., Lipford, H.R., Melton, J.T.: Aside: IDE support for web application security. In: Proceedings of the 27th Annual Computer Security Applications Conference, pp. 267–276. ACM (2011)
22. Acar, Y., Backes, M., Fahl, S., Kim, D., Mazurek, M.L., Stransky, C.: You get where you’re looking for: The impact of information sources on code security. In: 2016 IEEE Symposium on Security and Privacy (SP), pp. 289–305. IEEE (2016)
23. Wijesekera, P., Baokar, A., Hosseini, A., Egelman, S., Wagner, D., Beznosov, K.: Android permissions remystified: a field study on contextual integrity. In: USENIX Security Symposium, pp. 499–514 (2015)



Security and Privacy Issues for Business Intelligence in IoT

Mohan Krishna Kagita^(✉)

Charles Sturt University, Melbourne, Australia
mohankrishna4k@gmail.com

Abstract. IoT is a revolution in the business sector and adoption of IoT devices in the business organization leads to the success management of business data along with the success of the organization. Implementation of IoT in the business organization helps in building an Enterprise Competitive advantages and increases the efficiency of operational activities. Big data and cloud computing has provided a significant impact on improving the business intelligence by analyzing, identifying and discovering the business-related data. Data is biggest asset of every business organization and proper data maintenance is the major concern for every organization. Internet of Things (IoT) is using in every sector of business for efficient utilization of data and the organization expected to have enriched data from the use of IoT devices. The security issues of IoT are a biggest concern in managing and maintaining business data. Business intelligence is maintained dependent on the historical and present data in order to estimate and perform calculation on future data. Due to security issues of IoT the organizational data can be in danger due to potential threats, loss of data and data manipulation. Business intelligence mainly depends on the plugins which needs internet connectivity. Attacker can perform on the browser of the system and destroy the plugins associated with it, if browser is not protected from antivirus and firewalls. Lack in security features of the device will allows attackers to trace the details of the system and perform attack by manipulating or stealing the data. These types of attacks are intentionally performed to exploit the business data of the organization by hampering the business intelligence.

Keywords: IoT · Big data · Cloud computing

1 Introduction

Business Intelligence is an up growing technology in the business industries which uses an approach of the technology-driven process to make any strategic planning and decision. Business intelligence helps in analyzing the data and the information to assist the business executives and the managers in making an effective business decision. Business intelligence comprises uses of various tools and methods which enables and assist and organization in the collection of relevant data and information from the internal sources and the external sources. Internet of Thing is widely used in the Business Intelligence process for making a strategic decision and to make an effective planning. Internet of thing allows employees working in the organization to connect to multiple devices which are operating on the same network. Implementation of IoT

devices in the business sector provides a huge advantage in managing and planning the business operation. Considering the advantages of IoT in business intelligence, it has been observed that IoT is very effective in estimating and evaluating the sales and understanding the market strategy. The operation activity of the business can be improved with the implementation of IoT in the business industry. In terms of business intelligence, IoT can be very much effective and inefficient in the utilization of the data and the information to assist the top management employees in getting the enriched data for making any business plan or strategy. Business intelligence is basically dependent on the fact that, it uses both the present and the historical data in order to differentiate the change and the progress. IoT is very much effective in making the calculation and estimating the future data of the business organization. Big data and the cloud computing is extensively used in the Business intelligence and it has created a significant impact on improving the business intelligence. IoT is supported by the cloud computing network therefore, it has a great impact on the Business intelligence with the uses of big data by identifying, analyzing and discovering the data related to the business. The IoT uses a collection of data from various sources and it can be internal and the external.

In the background literature review, several types of attacks of IoT with respect to business intelligence will occur. In addition to that different types of attack in the IoT platform which can hamper on business intelligence will also be cover in this section. Several types of issues and the challenges may occur in the organization when the IoT platform is deploy for the business intelligence. Security challenges and the other threats will be discussed in this section. The solution to various types of security challenges, issues and the threats will also describe in the report Different types of research methodologies will be cover in the report. The advantages and the disadvantages related to the Business Intelligence in IoT will be cover in the section. The conclusion to the entire assignment will be provided.

2 Background Literature Review

As per the author Gubbi, the internet of things is a network of physically connected devices which be used in the business organization, vehicle, home appliances etc. It mainly consists of the software, sensors, connectivity and the actuators. As the author has described that the combination of these components with the internet connectivity helps in exchanging data and information. This can be used for the purpose of performing specific operation or task [1]. The vision for the evolution of the internet of things had occurred in the year 2016. Internet of things has scope for the implication of application in every sector. It can be used for the implementation of smart homes, consumer application, media application etc. IoT is being adopted in every sector of the business. According to the authors it has been analyzed that in the near future, the internet of things will become an open source network for the business. Internet of things has not at all remained a new concept and it has become a hot topic for the business organization and industry.

Suryadevara and Mukhopadhyay, as described in their article that, internet of things is the use of the devices which are intelligently connected. The connected devices are

embedded with the actuators and the sensors and various others physical devices to leverage the data [2]. The author had described in their article that the IoT has rapidly expanded during the recent years for unleashing and providing a new dimension to the organization and the business industries. IoT is very much effective for providing an effective solution to the customers for the drastic and dramatic improvement of the energy efficiency, education, health and the security of the business process human life and the living standard. IoT is considered as the giant network for connecting the people and the things. It has been stated by the author that, the main objective of using the IoT technology is to move the functionality of the device towards a smarter technology. The author has further described that in IoT, the objects, and the devices are connected with the platform of the internet of things for integrating data from several devices and to apply analytics on the integrated collected data for sharing them for the purpose of sharing the quality information. As the author has described, IoT is very much powerful indicating the exact information that will be required for the purpose of analyzing the data that will be required. From the perspective of the author, IoT is very much effective for identifying what data to be accepted and what data that needed to be ignored. As the author has stated in their article, the pinpointed information from the is very much effective for detecting the patterns, making any sort of recommendation and for the purpose of detecting problems that may occur. Further, the author has described in their article the information that has been gathered from the IoT platform is very much effective for making any kind of smart decision.

Lazarescu has described the major characteristics of the IoT platform. As per Author has described, it is possible in an IoT platform for interconnecting everything in the global communication and the information structure [3]. The author of the article further described that the IoT platform is capable enough further providing the services related to the things such as the association of the virtual and the physical things. In the article, the author has highlighted with the point that to provide a service related to the things, the information and the technologies which are related to the physical world will get change. As per author, the devices which exist and used in the IoT platform are apparently heterogeneous as it is dependent on different network and the hardware platform. The author has included in the article that dynamic changes can be obtained when a change in the state of the device will occur. In the article, the author has, described that the safety is another characteristic of the IoT platform which is very much effective in providing the data and the information security.

Datta et al. described, about the real-world application of the IoT platform. In the article, the author has described that the application scope of IoT in the real world is very huge. The IoT platform is used by the business organization, industry or by the human to get dynamic outcome or result. As it is described by the author, that IoT is making a huge revolution for making a Smart Home [4]. It has described by the author that Smart home had become a ladder of revolution to improve and to bring advancement in the automation of the residential spaces. As per author, the approach to smart home will be popular and it will be common as like smartphone in the recent upcoming years. In the article, the author has further described that wearable product which has been developed and made from the IoT platform made a huge explosion in the present demanding market. In the article, the author has stated that the wearable products are integrated and embedded with the software and the sensor which is mainly

used for the collection of information and the data of the users. The data and the information which has been collected are later on the process for extracting the insight of the user. The other application feature which has been described by the author is the connected cars. The automobile is extensively using the IoT technology advantage for optimizing the internal functionality of the vehicle. The author has highlighted that more preferences has been given by the Automobile industry for the adoption of IoT Technology to enhance the cars experience. In the article the author, the authors have further, elaborated the concept of the Connected car by stating that it is the vehicle which is used for the purpose of optimizing the performance. In the article, the author has described about the other application features and the advantages of using IoT technology. As per the article, the other application feature of the Internet of things includes the industrial internet which is also termed as the industrial internet of things. The author has described in the article that industrial internet of things is encouraging the engineers with the software, big data analysis, sensor etc. for creating and developing an advanced machine with quality features. As per the author, the other application feature of the IoT platform includes smart cities, IoT in agriculture, smart retailing for energy management, in healthcare etc.

As per the article of Leminen et al., the IoT is being extensively used for the business intelligence. In the article, the author has described that business intelligence is a process which is more likely to be technology is driven for the purpose of analyzing the data and to present the information which is actionable to assist the managers, executives and other associated of the corporates for making an informed business decision [5]. According to the author, the main objective of using the business intelligence in an organization is for the analysis of data. It is being highlighted by the author, that the implementation of the IoT platform in the business organization has provided a huge advantage to the business organization for improving the effectiveness of business intelligence. It is being described by the author, that the effective business intelligence with the implementation of the IoT platform in the business can be very fruitful for optimizing the internal business process and to increase the efficiency of the business operation. The author has stated in the article that the implementation of IoT in the business organization for the purpose of improving the Business intelligence is effective in attaining the competitive advantage over the rivalry of the business and to generate new revenues. As per author, the Business Intelligence with the Internet of Thing technology assists the organization in identifying the presence and the upcoming marketing trends of the business and it is very much efficient in evaluating the problems of the business. According to the author, the IoT for the business intelligence leverages the services and the software for the transformation of the data into the form of actionable intelligence which assists the organization in making any critical and the complex decision for the business.

As per Larson and Chang, the adoption of IoT for business intelligence helps in creating Business Intelligence Dashboard for the business organization. The author has stated in the article that there are several reasons for the incorporation of the business intelligence in the organization [6]. According to the author, the adoption of IoT for the purpose of Business intelligence assist the organization in tracking and monitoring the real-time data and it can also be used for the purpose of sharing the real-time data to make a better decision related to the business. In the article, the author has further

described that the implementation of IoT platform for the Business intelligence increases the censurability of the business operation. As per author, the recent advancement in the technology assists an organization to collect tons of data in the world of data-driven. The collected information or data can be converted easily into a readable format with the help of business intelligence tools. Apart from that, Business Intelligence also very efficient in analyzing and maintaining updated information. According to the author, it is effective in making a quick data collection and the real-time processing and the collection of the data. According to the author, Business intelligence is related to the concept of the Big Data. In the article, the author has described that, the big data mainly focus on the complicated information and the data analytics. Therefore, in the article, the author has mentioned that the business intelligence also focusses on the analyzing the set of data over the areas such as sophisticated tools, software application, and the infrastructure. As per Author IoT is extensively important for the organization for establishing communication between different types of sources and the product such as a sensor, wearable technology and the devices. In the addition to that, the author has further described that the implementation of IoT in the business organization makes the Business Intelligence model.

3 Different Types of Attacks of IoT in Business Intelligence

The attacks can be of several types. Due to the security and the privacy issue in the IoT platform the attacks are performed. The attacks can be performed by the hacker, intruders and the unauthorized agents.

As per Larson and Chang, the attacks can be of different types, it can be physical cyber-attacks, Network cyber-attacks, Software attacks, and the encryption attacks. The author has described in the article that the main reason for the physical attack is because of the sensor's present in the IoT devices. In the context of the physical attack, the author has further described that in this type of attack the hacker generally tries to access the system of the user which are located near to the close proximity [6]. As per author, tampering will assist the hackers or the intruders to extract the data which is infused with the malicious code. In the article, the author has also mentioned about the Cyber-attacks on the network system. In the network cyber-attack, the author has described, the hackers or the intruder's tries to access the network of the user in order to identify and to check what data is exactly flowing in the network. The most common example of Network cyber-attack is a man-in-the-middle attack. Apart from the network cyber-attack, the other attack which has been described by the author is software attack. In the context of software attack, the author has described that malicious files or the malware are injected into the users to track and monitor the data of user system and also to track the data flow. The author has further described that software attack can corrupt the data or the files by introducing a virus. The last type of attack which has described by the author is the encryption attack. In the context of encryption attack, the intruders or the hackers deduce the key of the encryption for creating their own code and algorithm for unlocking the encryption key. As per author, once the hacker able to

unlock the key, they introduce their own code into the user system for monitoring the system. As per author based on the type of attack, it can be DDoS (Distributed Denial of services), Bonnets, Man in the Middle attack etc.

- **Botnet Attacks**

Bertino and Islam has described that Botnet is the connection of one or more device. According to the author, in the context of Botnet attack, it is generally being practiced with the intention to disrupt the normal operation work or it can be used for the degradation of the overall services of the target system [7]. For its creation, a huge number of Botnets are required before initializing the attack. As per the author Fig. 1 shows, once the attack has been initialized, the botnet6s are sent to the network in order to target the system at large scale. The request for the attack in IoT for the business intelligence comes in the form of messages or in the form of emails. The types of attack can create an adverse impact on the business intelligence by slowing down the network server by making the network busy for the user to access the network by temporally freezing the server.



Fig. 1. The most common type of attacks of Botnets attack is distributed denial of service attack (DDoS).

- **Identity Theft**

Vidalis and Angelopoulou has described the impact of identity theft of the Business Intelligence. As per author, the identity theft is basically a crime for accessing the information related to the finance or any personal data [8]. Despite this, the author has described that the impact of identity theft on the business intelligence is very high. In the context of Identity theft, the, by creating stealing the identity of the user or the credential information, the data related to the business intelligence or the data analytics. As per the author, the identity can be of two types. The author has mentioned in the article that account takes over and the true name is the two types of identity theft. As per the author, the attackers or the hackers use the personal information of the user in true name identity theft to collect data regarding the business analytics. In the context

of an account take over identity theft, the personal information of the users existing account is used which is been modified by the hackers or the attackers to access information related to the business data analytics of the organization.

- **Denial of Service**

Joshi and Tipper has stated that Denial of service attack is another issue of concern in the aspect of business intelligence in the Internet of things platform. From the article, the author has clearly stated that denial of service occurs when there is the unavailability of the service happens during the usual work process [9]. As the author has highlighted in the article that unavailability of service may happen due to various reasons. In the context of distributed denial of service attack, the author has described that very large number of computer systems which is used for the business intelligence in the may get affected by introducing malicious files or the malware in the system of the organization. As per author Fig. 2 shows, injecting the malicious code of files in the system of the organization will create an adverse impact where it will make the data inaccessible to the employees or the business organization for the proper analysis and make a strategic decision. As per the author in terms of denial of service attack, the network, the network or the servers are unable to find the returning address of the hackers or the attackers while sending the approval for the authentication. According to the author, it makes the business organization for making a proper analysis of the data as because the effect of denial of service makes the server to run slow.



Fig. 2. Denial of service.

- **Man in the Middle Attack**

Zhao and Ge have described the man in the middle attack Fig. 3 shows, the hackers or the attackers try to the intercept the communication occurring between two systems [10]. The author has highlighted in the article that such kind of attack has a very adverse impact on the business intelligence of the IoT Platform.

According to the author, this type of attack in the business intelligence can result into a very dangerous attacker has the complete control over the user system to manipulate and to change the analytical data of the business organization.

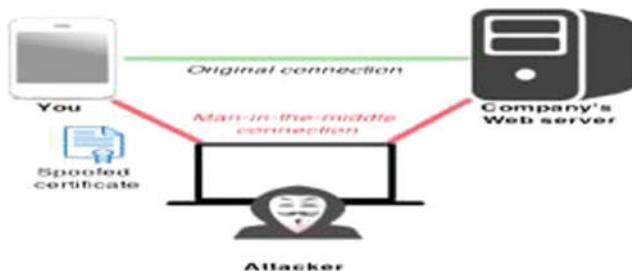


Fig. 3. The hackers or the attackers try to intercept the communication occurring between two systems.

4 Issues and the Solution

Depending on different types of attack there may be several types of issue that may occur. The issues with respect to the IoT in Business Intelligence are as follows.

Understanding the IoT: The concept of the internet of things is very much complex and the issues with the understanding the concept of the internet of things may occur. The main issue related to IoT in the sector of business intelligence is all about enhancing the understanding capability for the changes and the implication made in the IoT platform to improve the business intelligence a, activity and the data analytics survey [11]. Improper understanding of the internet of things, make the business intelligence more complex for making any proper decision and to process the actual and the real-time data.

The Issue with the Data Connectivity: The internet of things is dependent on several technologies and the internet connectivity. Problem with the data connectivity issue may occur if the network and the internet connectivity will be unavailable or slow. In the context of business intelligence, a huge amount of data is gathered and collected to process the data for the analysis. In the context of business intelligence, instant updating is required and IoT is totally dependent on the internet connectivity. Therefore, slow internet connection and the unavailable internet connectivity will create an issue while processing the data.

The Issue with the Compatibility of the Hardware: The process of data capturing occurs through different types of sensors. These sensors are connected to the gateways of the IoT for the collection and the transmission of the data on the cloud [9]. The adoption of IoT for the purpose of business intelligence could be a critical situation of an issue in the business organization if the hardware of the IoT devices are too compatible enough to support the Business Intelligence system.

Issues Related to the Analytics: Business intelligence is entirely based on the collection of the data which will be processed for actionable insights. Depending on the demand for the high-performance data analytics platform, which is capable enough to handle huge amount data for adding the data at the later point of time? Managing a large amount of data for the purpose of analysis can be a major problem too during the critical situation.

Data Security Issue: Data security is a major issue for the adoption of any type of technology. Due to the occurrence of the ransomware attacks and other malware attacks, there are several types of security issues that may arise in the IoT platform. The security issue in the IoT platform may lead to creating data loss, monitoring or the tracking of the organizational data which can be used for the purpose of Business Intelligence and the other types of data and hardware related issue may occur with the data related security issue.

Cloud Attacks: IoT uses the cloud computing and the cloud server for the storage and the transmission of the data. In the context of cloud attack, the data saved in the server by the organization for purpose of data analysis may get affected. The data related to the business intelligence are saved in the in the cloud server. The attack in the cloud on the server will lead to the change or the manipulation of the data in a cloud server. Such issue can create an adverse effect on the data related to the Business intelligence.

Security Issues in AI Built: Artificial Intelligence is taking a new shape in the world of business and the business organization uses the AI built-in technology to enhance the business intelligence. The introduction to the AI feature assists the organization in better analysis of the data. Integrating AI features into the IoT platform makes for business intelligence makes identification and analysis of the data more appropriately with respect to the market trend [11]. The rise in different types of cyber-attacks such as ransomware attack, malware attack through virus Trojan etc. is creating a serious issue in the security features of the AI built, thereby making an adverse impact on the IoT platform and the business intelligence too.

Information Leakage and the Information Tracks: Data and the information is one of the main components of every business organization. In the context of business intelligence, the growth of the organization and the future trends of the business are analyzed by collecting data related to the business. A large amount of data is collected and process for the purpose of data analysis. Lack in the security features of the IoT platform will allow the attacker or the hackers to have access on the business data. It will be easy for the hacker or the intruders to track and the monitor the data related to the business. This will assist the unauthorized agent, or the hackers top manipulate the data related to the business and can bring financial loss to the organization. Lack of security in IoT platform leads to such type of issue.

Selection of the Right IoT Platform: Selection of the Right IoT platform for the business organization is very much crucial. The implementation of the IoT in business intelligence is very much complex. Therefore, a change in technology in future leads to change an entire business intelligence model. Selection of improper platform for the business intelligence may also create security issue in the near future.

Data Encryption Issue: There are several types of IoT supported tools which are not capable enough to encrypt the business. There high rate of data flow occurs in the business intelligence approach. Even after the use of the internet, there are several IoT platforms for the business intelligence which is unable to encrypt the data properly in terms of business intelligence.

Solution

Due to several types of attacks, there are major types of issues that may occur in the IoT platform which creates an adverse effect on the Business intelligence. Preventive action and maintaining the security measures will assist in mitigating such issues.

Proper Security Requirements and the Verification of the Function: Security is one of the major aspects of IoT which is extremely important to make the function properly. Implementation of security features in the IoT platform for the Business Intelligence will make the system secure in order to prevent the system from any type of cyber-attacks and the threats.

Secure Review of Good: It is mandatory for the business organization to maintain a secure review of the code. The implementation of IoT platform for the business intelligence requires secure review of the code in order to minimize the level of threat.

Implementation of Penetration Technique: Penetration technique is crucial to identify the vulnerabilities or the threats over the network. End to end penetration technique will help in finding bugs and the errors in the network for the effective implementation of the Business intelligence in the IoT platform.

Encryption of Data: As it has been discussed, Implementation of Business Intelligence on the IoT platform is entirely dependent on the data. The data which has been gathered is later on the process for the analysis. Encryption of data will enhance the security features of the data during the collection or the transmission over the Network. Encryption of data will make the data more secure for the analysis of the data. It can be implemented with the cryptographic technique by using the public key and the private key encryption.

Use of Secure Socket Layer: The IoT is entirely dependent on the network connectivity and the web interface. In most of the business intelligence model, the exchange of information occurs through the web interface. Therefore, the user of secure socket layer will mitigate protect the data from malware or ransomware attacks.

Authorization and the Authentication of the Gateway: Gateway is a bridge which exists between the application server and the LAN which is connected to the internet. In the context of the Business Intelligence the analysis of the data occurs over the internet. There proper authentication of the server and the gateway would be effective for the proper implementation of the Business Intelligence in the organization.

5 Advantages and the Disadvantages

The advantages of IoT in Business Intelligence are as follows.

Optimizing the Performance of the Asset: The implementation and the adoption of the IoT platform for the business intelligence are very much effective for the identification of the potential problems. The performance of the asset can be improved by enhancing the capacity, reliability and the capacity of the asset.

Improving the Operational Efficiency: The operational efficiency of the organization can be improved with an effective interaction process by analyzing the operational data. The adoption of the IoT platform helps in monitoring the real-time data by providing a clear access to the historical data.

Load Management and the Dynamic Forecasting: Successful management of the demand and the supply operation can be established to reduce the load. The future business and the marketing trends of the organization can be attained through business intelligence with the implementation of IoT.

Fraud Management and Prevention on Utility Loss: The anomalies in the business activity can be determined to prevent the utility loss and the fraud management can be achieved with the implementation IoT for the Business Intelligence in the organization.

The disadvantages of IoT in Business Intelligence are as follows.

Complexity: IoT is very much complex and to integrate the IoT platform with the business intelligence will be very complex as it is dependent on several technologies.

Security and Privacy: Security and the privacy are the major concern in the IoT as it is totally dependent on the internet. Therefore, there are high chances of data attack.

Compatibility: Not all the components of the IoT are compatible enough to comply with other technology.

Safety: Lack of safety features will allow hacker on the network to hack upon the data of the organization.

6 Conclusions

As per the paper, the several issues and the challenges related to the business intelligence in the IoT platform has been discussed. A background literature review has been provided by emphasizing the different types of attacks on IoT platform. The paper has covered on aspects of several issues and the solution of business intelligence in the IoT platform followed by the future research. Lastly, the major advantage and the disadvantages of the IoT for business intelligence has been covered in the report.

References

1. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of Things (IoT): a vision, architectural elements, and future directions. Future Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
2. Kelly, S.D.T., Suryadevara, N.K., Mukhopadhyay, S.C.: Towards the implementation of IoT for environmental condition monitoring in homes. IEEE Sens. J. **13**(10), 3846–3853 (2013)
3. Lazarescu, M.T.: Design of a WSN platform for long-term environmental monitoring for IoT applications. IEEE J. Emerg. Sel. Top. Circuits Syst. **3**(1), 45–54 (2013)
4. Datta, S.K., Bonnet, C., Nikaein, N.: An IoT gateway centric architecture to provide novel M2M services. In: 2014 IEEE World Forum on Internet of Things (WF-IoT), pp. 514–519, Mar 2014. IEEE (2014)
5. Leminen, S., Westerlund, M., Rajahonka, M., Siuruainen, R.: Towards IoT ecosystems and business models. In: Internet of Things, Smart Spaces, and Next Generation Networking, pp. 15–26. Springer, Berlin, Heidelberg (2012)
6. Larson, D., Chang, V.: A review and future direction of agile, business intelligence, analytics and data science. Int. J. Inf. Manage. **36**(5), 700–710 (2016)
7. Bertino, E., Islam, N.: Botnets and internet of things security. Computer **50**(2), 76–79 (2017)

8. Vidalis, S., Angelopoulou, O.: Assessing identity theft in the Internet of Things. *J. IT Gov. Pract.* (2014)
9. Zargar, S.T., Joshi, J., Tipper, D.: A survey of defense mechanisms against distributed denial of service (DDoS) flooding attacks. *IEEE Commun. Surv. Tutor.* **15**(4), 2046–2206 (2013)
10. Zhao, K., Ge, L.: A survey on the internet of things security. In: 2013 9th International Conference on Computational Intelligence and Security (CIS), pp. 663–667, Dec 2013. IEEE (2013)
11. Davenport, T.H.: Business intelligence and organizational decisions. In: *Organizational Applications of Business Intelligence Management: Emerging Trends* (2012)



Securing a Network: How Effective Using Firewalls and VPNs Are?

Sun Jingyao, Sonali Chandel^(✉), Yu Yunnan, Zang Jingji,
and Zhang Zhipeng

New York Institute of Technology, Nanjing, China
{jsun19,schandel,yyul8,jzang,Zzhang36}@nyit.edu

Abstract. With the tremendous amount of increase in cyber threats on the Internet, the security of data traveling over a network has become a significant concern for all the netizens. As a result, a large number of Internet users have started using firewalls and VPN (Virtual Private Network) to ensure more protection for their data on the go. Though mostly considered as defenders of our network security, sometimes firewalls and VPNs can also pose some serious threats to its users. Our research focuses on addressing these security flaws by providing a specific illustration of the working principles and performance of the firewalls and VPNs, including the technologies behind them and their benefits, significant potential risks it may bring due to some considerable loopholes in their architecture, and the possible solutions to those security issues. We hope that our research will bring a better understanding of these security issues and their solution to help users and organizations to deal with these security threats and risks in a better way.

Keywords: Firewall · VPN · Network security

1 Introduction

The importance of the Internet and its security was never as great of a concern as it is now because of the amount of data that is exchanged through it 24×7 . Plenty of threats and risks exist in the cyber world that can affect the network security in a big way. Hacking is one of the most common cyber threats to a network as it allows the hackers aka malicious users to manipulate and attack the loopholes of vulnerable networks and take control of it. Besides the external damage, which mainly reflects the destruction caused by hackers by mostly using malware, virus or DDoS attacks, the threat of resource openness because of using the computer network in a shared environment, cannot be ignored at all. The exponential growth of security risks and dangers that exists outside of a network in the present times can strictly confirm the necessity for people to use and study the firewalls and VPNs so that the attacks can be prevented and detected, and the network can be protected from getting damaged [1].

Firewall is a technology that is used to control the degree of interconnection between different networks. It can prevent the external network from accessing the internal network equipment and network resources using unauthorized ways. This means a firewall can protect internal network and system from potential threats of a

network attack. This technology fully combines the potential of hardware and software in a computer network and realizes active filtering and screening of potential threats and risks to a network. A firewall usually is the first step to intercept an external attack to accomplish the adequate protection for computer network security [2]. Figure 1 [3] introduce the connection schematic of Firewall, Intranet, and the Internet. The essential features and primary functions of Firewalls are shown in Table 1 [3].

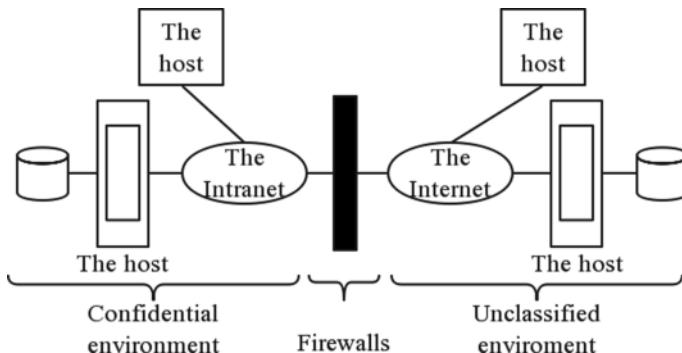


Fig. 1. The connection schematic of a Firewall, Intranet, and Internet

Table 1. The basic features and main functions of a firewall

Basic features	Main functions
All network data between the external and internal networks must pass through the firewall	Firewall is a barrier to network security
Only data flows that conform to security policies can pass through a firewall	It can strengthen network security strategy
The firewall itself should have powerful immunity against attacks	It can monitor and audit network access It prevents leakage of internal information

A VPN is a virtual encrypted tunnel between the user and a remote server operated by a VPN service. All external Internet traffic is routed through this tunnel, so our data becomes safe from the data hunters. On the other hand, the IP address of the VPN server becomes the user's IP address, enabling them to hide their actual identity [4].

Firewalls are the gateways to ensure the security of the internal network and VPNs are ways to access the internal network. There are always firewalls in the place where there are VPNs. VPNs can be used with or without firewalls, but they are not recommended to be implemented without firewalls as their primary purpose is to secure the network traffic. A VPN without a firewall makes VPN's encryption function useless. Using VPN with firewall further enhances the security of the Internet and network in general.

This paper aims to provide a detailed study of the security issues and its solutions that the users of a Firewall and a VPN should know. In this paper, we have proposed some suggestions for the safety and security in using the Firewalls and VPNs based on the literature survey that we did for our research. This paper is structured as follows: Related work is mentioned in Sect. 2. In Sect. 3, an overview of the firewall and VPN technologies has been discussed. Section 4 introduces the security issues related to using firewalls and VPNs. The most common threats and attacks in using firewalls and VPNs are discussed in Sect. 5. Solutions for the most common security issues in using firewalls and VPNs has been discussed in Sect. 6. In Sect. 7, the conclusions are drawn, and future work has been mentioned.

2 Related Work

In the past, the authors of [1, 3, 26, 30] focused on the security of firewalls and VPNs, but they did not discuss anything regarding the attacks or threats on firewalls and VPNs. [2] does not provide some specific methods about how to improve the security issues of firewalls. The authors of [9] and [12] present just one case study and focus on one aspect. [9] focuses on an example of a system using both Firewall and VPN. It presents a case of two firewalls with the same configuration in the same network node, which communicate with each other through a direct connection. [12] focuses only on the deep packet inspection technology of firewall. Furthermore, paper [10] is an old study that does not cover the recent problems and various attacks that happens in the cyber world presently. Paper [13] also talks about one model without its realization. In addition, some papers like [17, 19, 20, 24–26] only concentrate on one attack rather than its relationship with the firewalls. The work done by us will not only analyze firewalls and VPNs individually, but we will also compare them together to enhance their abilities in providing more security. It also means that the security loopholes and the solution will be related to both of them. We will analyze the better structure or system concerning the latest products like Web Application Firewall, Secure Web Gateway, and Next Generation Firewall to find the reasons behind them for being considered as safer than traditional firewalls and VPN setups.

3 The Working Principles of Firewalls and VPNs

3.1 The Working Principle of Firewalls

Firewall technology has been developing continuously since its birth. Various kinds of firewalls with different structures and functions are built into a network for getting more defense. Traditional firewall technology falls into three categories, and no matter how sophisticated the implementation of a firewall is, it is ultimately based on the following three technologies.

3.1.1 Packet Filtering

The working principle behind the Packet filtering firewall can be called as a network firewall because it works in the network layer. It usually decides whether to let the data packets pass, by examining the address, protocols, and ports of each packet. Figure 2 shows the principles of Packet filtering. The packet filter can be divided into a static packet filter and dynamic packet filter [5].

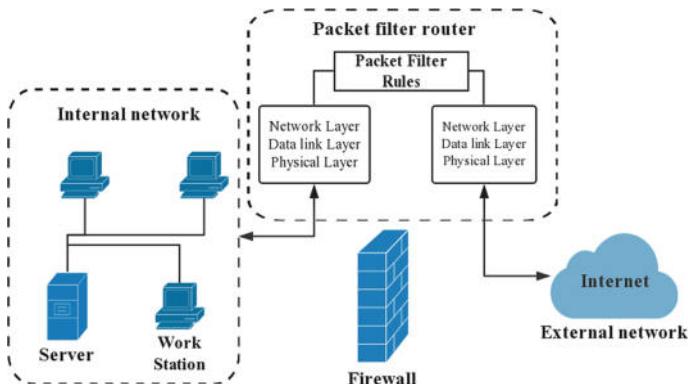


Fig. 2. The principle of packet filtering

- **Static packet filter.** Static packet filter technology is a traditional packet filter. It decides whether these data packets can be passed according to the IP addresses of these data packets. If attackers set their mainframe IP addresses as legal addresses, they can quickly pass the static packet filtering firewall. Therefore, this kind of firewall is not secure enough.
- **Dynamic packet filter.** It can automatically apply to create or delete packet filter rules according to dynamic practice application without administrators' intervention. However, the dynamic packet filter technology can only filter against data's IP address instead of the legality of users. It also does not have log records to check which brings enormous difficulties to daily network security management. Therefore, it has been replaced by a new technology called the Adaptive Proxy Protection firewall.

3.1.2 Application Proxy

Application Proxy firewall is also called as an Application Gateway firewall. This firewall participates in the entire process of a TCP connection through Proxy technology. The data packets sent from the inside are processed by the firewall to hide the Intranet structure. Network security experts recognize this type of firewall as the most secure firewall. Its core technology is the proxy server technology. The proxy server refers to the program that represents the client's connection request on the server. When the proxy server gets connection intentions from a client, they will verify the client's request and then handle connection requests through the specific secure proxy

applications. The request is transferred to the real server, which then accepts the server response. After further processing, the proxy will reply to the final client who makes a request. The proxy server plays the role of interconnecting the application of the external network to the internal network [5].

Adaptive Proxy Firewall

Adaptive proxy is a revolutionary technology implemented recently in commercial application firewalls. It combines the advantages of the safety of the Application Proxy firewall and the high speed of packet filtering firewall and improves the performance of the proxy firewall by ten times without losing the security.

3.1.3 Stateful Inspection

Stateful Inspection is an extension of the packet-by-packet filtering process, which tracks individual flows, enabling policy checks that extend across a series of packets [6]. It checks the handshakes in a communication network by exploiting detailed information of the communication protocol. It detects malicious activities by monitoring packet-by-packet connection and predicting the next move based on what happened. This makes it a more advanced tool than other firewalls [7]. These firewalls maintain a table of open connections, inspecting the payload of some packets and intelligently associating new connection requests with existing legitimate connections [8]. With modern firewalls, network administrators can control the network traffic in a more fine-grained fashion.

3.2 The Working Principle of VPNs

The mainstream applications that claim to provide VPN services are using one of the following three techniques [4]:

- **Proxy Server.** A proxy server is like a courier service that is responsible for only transcending the messages. The work of proxy servers is conducted in the HTTP layer and the Socket layer in the Open System Interconnection (OSI) model under most circumstances. Figure 3 explains how proxy server functions.

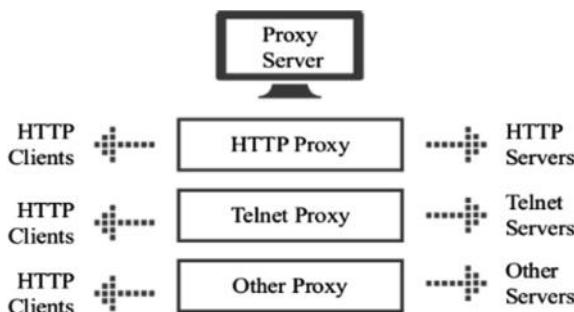


Fig. 3. The working architecture of a proxy server

- **IPSec.** IP Security (IPSec) is the most common method used by the VPN applications. It works in the third layer called the Network layer of the OSI model.
- **SSH.** An encrypted channel needs to be combined with the proxy server to overcome the blocked network. The tool that is used to scale the blocked network called SSH is, in fact, an SSH agent. In the TCP/IP 5-tier model, SSH is the security protocol that applies to the application layer and the transport layer. SSH is a remote shell, an application based on SSL. Although many people use SSH to transmit data, they merely use the SSL proxy function of SSHD software to get this job done.

3.3 Architecture of a System Using a Combination of Firewall and VPN

Figure 4 shows a typical network security architecture based on the combination of both firewall and VPN technology [9]. The system has a master node under which there are large nodes and links between them. Under the big node, there are intermediate nodes, and under the intermediate nodes, there are small nodes. The intermediate node and the small node is only connected with its own upper and lower levels. Between nodes, individual wire connections can also support other ways of connecting people, such as wireless connections. Security between nodes ensures the safe transmission of data through the virtual encryption channel between VPN array devices and VPN receiver devices.

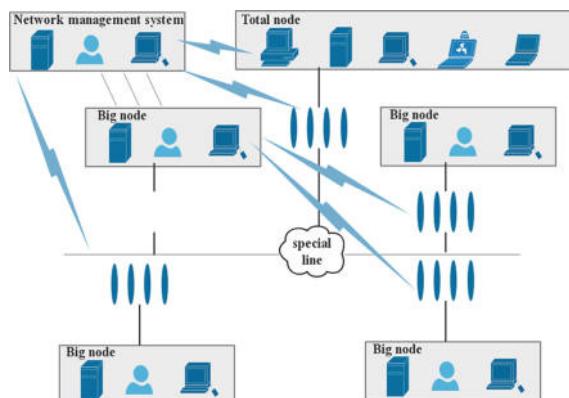


Fig. 4. A typical network security architecture based on firewall and VPN technology

The entire network adopts a network management system to manage, control and report congestion for various security devices, routers, switches and servers in the network, as well as fault management. The network management system will monitor equipment utilization, bandwidth utilization, packet loss rate, etc. Two firewalls with the same configuration are used in the same network node, and they communicate with each other through a direct connection. Under normal circumstances, one is working, which is the primary, and the other is in the backup state.

High availability is achieved through “heartbeat” mode. A direct connection between two firewalls of the same type creates the heartbeat line that uses fixed interval, master-slave equipment to exchange information. When the host accidentally crashes, or network fails, hardware failure happens. Master-slave firewall switch is a working state from the machine instead of the standard work of the host, to guarantee the regular use of the network. Switching process does not require a human operator. They also do not need the participation of other systems. The primary firewall’s restore function will automatically return the control to the firewall, to assure the safety of the network. By deploying this network security system with firewall technology as the core, a VPN can realize the secure exchange of confidential information between nodes.

4 Security Issues in Using Firewalls and VPN

Even though firewalls and VPNs are used for protecting or mitigating the external attacks on a network or a system, but they are not full proof. In this section, we will discuss some of the shortcomings in both of these technologies.

4.1 Security Issues of Firewalls

With the increasing severity of network security issues and the continuous development of security defense, shortcomings of firewalls in the security aspect have gradually attracted much attention from security researchers and organizations. Due to the existing loopholes in the firewall architecture, a series of attacks can quickly destroy a network. Analyzing firewall’s vulnerabilities and the attacks against it is of great significance to the development and improvement of firewall technology for complete network security.

The most common security flaws of firewalls are as follows:

- **Firewalls may sacrifice some useful network services.** Firewall’s “either in or out” feature is bound to shut down some valuable ports due to security problems, sacrificing some helpful network services as well [10]. After the firewall receives network packets at the network layer (including the following link layer), it matches them one by one based on the rules. It then performs prearranged actions if they are consistent, such as allowing or denying packet access. This creates a lot of discomfort and challenges for the organizations and users alike to make sure that the good packets are not misunderstood and blocked as bad ones.
- **Firewalls cannot protect against attacks from internal network users.** Firewalls are the outlet for information between different networks or network security domains. It can control the flow of information in and out of the network according to the set security policy. This prevents the illegal information from flowing into the protected network without affecting the regular access of the protected network to the Internet. This feature of firewall determines that it can only filter packets between internal and external networks but cannot process packets from within the internal networks [11]. This makes it impossible to protect against attacks from internal network users commonly known as insider threats.

- **Firewalls are not secure against software or files with the virus.** Firewall for encrypted SSL stream of data is not visible. This means that the firewall cannot quickly seize the SSL data flow and do the decryption. Therefore, it cannot stop the attack of the application as well as cannot see the application firewall's encrypted data [12]. The firewall can recognize and intercept attack data only when the attacking behavior of the application layer matches the existing attacking behavior of the database in the firewall.
- **A firewall cannot extend depth detection.** It is impossible for a general firewall to extend depth detection that is based on the data package without increasing network performance accordingly. Profound detection capabilities for all network and application traffic require unprecedented processing power to accomplish a large number of computing tasks; including (1) SSL encryption/decryption function (2) Complete two-way payload detection (3) Ensure the normalization of all legitimate traffic (4) Extensive protocol performance. These tasks cannot run efficiently on standard PC hardware [13].

The weaknesses of five general kinds of firewalls are shown in Table 2 [13].

Table 2. The weaknesses of five general kinds of firewalls

Firewalls	Weaknesses in the security system
Packet filtering firewall	Hard to configure
Status/dynamic detection firewall	Delay in the network connection
Web application firewall	Limited range of user system
Network address translation firewall	Unable to mitigate the internal attacks and threats
Personal firewall	Unable to monitor and control multiple communication

4.2 Security Issues of VPNs

Any company or organization that implements a VPN to ensure their network security still cannot ignore the risks and threats that can destroy or sabotage their network without them even knowing or realizing it. The most common risks can be seen in Table 3.

Table 3. The risks involved in using a VPN

1	Securing against lateral network movement
2	Securing and connecting to cloud-based infrastructure
3	Blocking malicious insiders, over-privileged users, and compromised third-party access
4	Preventing malware from proliferating across the network
5	Efficiently integrating with business processes and identity management systems

IPSec VPN: This VPN has a robust communication protocol and encryption algorithm, so its security issues mainly come from its client's attacks [14].

- **The local security configuration is not perfect:** The users control the local security configuration of the VPN's client-side themselves. It means there might be some security risks caused by human factors. For example, some clients may keep the license certificate on the local device. Once the attacker gets the control of these devices, they can open a VPN channel without even needing a login name and the password and bypass the authentication process altogether.
- **Stealing VPN security information:** Attackers can steal VPN security information by using social engineering methods such as phishing. This security information includes the IP address of the VPN client, configuration parameters, user license certificates, etc. The attacker can forge a communication identity and pose a threat to the security of a VPN by using this security information.
- **The internal security of VPN is weak:** The security protection requirements for trusted users are relatively low after VPN is successfully connected. Because there is no attack prevention strategy in the tunnel that decreases the security risk within the VPN.

SSL VPN: This VPN does not require specialized client software, but they use web browsers for its implementation. Therefore, the security threats of SSL VPNs are mainly focused on browsers and servers.

- **Security problems caused by incorrect system operation:** If the user does not close the SSL VPN by logging out at the end of the browser and the server process, it may keep the SSL VPN server process open. The attacker can use this situation to bypass authentication and access the VPN, which brings a high-security threat to the VPN system.
- **Malicious attacks on identity authentication:** SSL VPNs allow users to log on to the VPN system from any location through a browser. This increases the risk of leaking security information such as login id and passwords, especially when they log in from public places.
- **The virus infects the internal network through the tunnel:** SSL VPN remote users can use any location of any client remote login within the enterprise network. Once the viruses at client-side connect to the internal network, the infected file will be able to use the SSL VPN tunnel to invade the internal network. At the same time, due to the limitations of the internal network boundary of the firewall, it cannot prevent the transmission of infected software or files effectively. As a result, the virus can infect the internal network through the tunnel.
- **The security risks of the Web server itself:** Most of the SSL VPN system use Web server as its underlying platform. Therefore, the potential safety hazard of the Web servers, such as the back door or unauthorized leaks will also bring serious security problems to the SSL VPN system [14].

MPLS VPN: This VPN has adopted a strict routing information isolation mechanism. The security of user information transmission is guaranteed by using

MPLS VPN. However, as a technology based on IP communication, its transmitted information is not encrypted and authenticated, so there are still some security problems that exist in MPLS VPN [14].

- **Attacks against VPN routing devices.** This attack usually occurs during the routing information release phase. The attacker disguised as an edge device establishes a session with the server equipment to connect and exchange routing information. This will cause the disclosure of the VPN's internal routing information. The attacker can also forge or tamper with the routing information to spread the user's data in the wrong direction to eavesdrop and steal the user's personal information.
- **Security threats from the Internet.** In the case of users accessing the Internet through MPLS VPN, the attackers can attack the network by traditional attack means such as IP source address deception, session hijacking, and planting Trojan horse in the network. The user's data flow will be viewed, modified, forged and deleted by the hackers without their knowledge.

4.3 The Loopholes in Using Firewalls with VPNs

The firewalls with VPNs can provide a virtual private network on the unsafe Internet through the VPN function. Therefore, it can guarantee the security of the confidential data of the enterprise when the remote access happens. However, at the same time, there will be many loopholes in using this arrangement of firewalls with VPNs. The loopholes in using firewalls with VPNs are shown in Table 4 [15].

5 Using Firewalls and VPNs: The Most Common Attacks and Threats

5.1 The Attacks and Threats of Using Firewalls

Generally, the most common attacks happen to the Packet Filtering Firewalls and the Status/Dynamic Detection Firewalls [16]. In Table 5, we list the attacks that could corrupt the firewall security.

The IP Spoofing Attack. It can easily make use of a legal address from the ordinary users. Attackers can avoid an authentication process provided by the firewall using this way and hide. Also, when attackers use spoofing attack, this behavior of hackers will make the log, and NAC (Network Access Control) will point to the wrong person when used to track down the attackers. This kind of MAC (Medium Access Control) attack is straightforward to create and can facilitate a variety of advanced attacks [17].

Denial of Service (DoS). Unlike many other attacks, DoS attack is purely malicious because the hackers gain nothing personal from the attack. They attack the user's system with the aim of depriving the system's working ability. To overload the victim network, the hackers send large data that floods the system. To send data, they usually need to know the IP address of the targeted network, but firewalls with VPN can hide the IP address and block the malicious data package.

Table 4. The loopholes in using firewalls with VPNs

Contents	Loopholes
Firewall rule virtual test	No work can detect the effect of the configured strategy
Intranet service permissions settings	No function
Quality of service loan allocation	In general, there is no VPN within QoS permissions
Multi-line superposition and backup of firewalls and VPNs	Only double backup
VPN maximum transmission unit	Manually modify maximum transmission unit based on the Internet environment
VPN dynamic IP addressing	We can only use dynamic domain name system and other third party's solutions. People control the use
Hardware binding authentication	No function. Only username, and password
USB key security policy storage and exchange	Only security certificates can be stored, and clients still need professional staff
Support for mobile users	Not supported or incorporated into through the PPTP protocol with little support and inadequate security
Protocol, encapsulation, and compression	Standard IPSec, using the network address translation standard. User datagram protocol encapsulation ensures that data is correct with other check fields. One by one packet encapsulation and no compression technology lead to low bandwidth utilization
VPN performance	Using low-end hardware components. Single channel connection speed is slow. The number of access support is limited, and performance is unprotected
Support for VPN channels	Unable to support a large number of branches and client access, network performance significantly decreased when there are more nodes
Software	No software VPN gateway
Convenience of implementation	It is complicated that it needs professional staffing
Support for access methods	It can only access with Internet IP and does not support new access modes such as cell broadband, WLAN, and GPRS

Table 5. The most common attacks on firewall

Types of firewall	Attack
Packet filtering	IP spoofing attack
	Denial of service
	IP fragmentation attacks
	Trojan attacks
Status/dynamic detection firewall	Protocol tunnel attack
	Passive FTP
	Rebound Trojan attack

IP Fragment Attack. In an IP fragment package, only the first fragment has the information of the TCP port. When the package is transmitted through the Packet Filtering Firewall, the firewall only checks the first fragment to decide whether to let it pass. In this case, the attacker can cheat the firewall by sending a legitimate first IP fragment, and then the rest of the malicious fragment can pass through the firewall and cause a threat to the network security [18].

Trojan Attack. It is the most effective attacking method to Packet Filtering Firewall because once the Trojan is installed inside the network, there is nothing a firewall can do to stop it. The reason is that the Packet Filtering Firewall usually only filter the packet at the lower port (1-1024) and most Trojan attacks through the higher ports [19].

Protocol Tunnel Attack. The attack of the protocol tunnel is similar to the idea of a VPN, and the attacker hides some malicious attack packets in the head of some of the protocols, so it can penetrate the firewall system and attack the internal network [20].

Passive FTP. It solves the issue of an FTP client's firewall blocking incoming connections. "PASV" is the command that is used by the FTP client to let the server know that it is in passive mode. This is a preferred mode for FTP clients behind a firewall and is often used for web-based FTP clients and computers connecting to an FTP server within a corporate network [21].

Rebound Trojan. The internal network's rebound Trojan periodically connects to a host controlled by an external attacker. Since the connection is initiated from within, the firewall considers it as a legitimate connection, causing the blind area of the firewall. A firewall cannot distinguish between a Trojan's connection and a legitimate connection. The limitation of this attack is that the Trojan must be installed inside the network first [22].

5.2 The Attacks and Threats to a VPN

Choosing a VPN is a good idea to get protection against a network, especially when it is a public Wi-Fi. However, use of a VPN can sometimes be a threat to security and bring some risks as well. VPN establishes a channel between the user and the server, so the user's trust in the VPN provider is essential because the provider can see and record all the data and can even alter the content. If a VPN is not configured correctly, a hacker might be able to access the user's local LAN directly, which is worse than being exposed to public Wi-Fi. For example, GoGo, a VPN provider was accused of using fake YouTube certificates that could leak users' passwords [23].

- **Man in the Middle Attack (MITM).** Some VPN providers adopt pre-shared key for their users, and that can lead their users to be caught up in a Man in the Middle attack (MITM). In the MITM attack, there are two endpoints of victims, and the attackers are third-party. The attackers can access the communication channel between two endpoints and manipulate the messages [24]. MITM attack aims to compromise the following three targets [25]:
 - *Confidentiality:* It can be achieved by eavesdropping on the communication.
 - *Integrity:* Attackers can intercept the communication and modify messages.
 - *Availability:* By intercepting and destroying messages or modifying messages, attackers can make one of the party to end communication.

- **Hacking or Eavesdropping.** It includes physical access or listening to the devices that support VPNs. This can happen if someone loses their laptop or mobile device, which supports VPNs. Most VPN applications are not configured for the best security model, and the local license is stored in the device itself. In this case, the hackers can access the VPN channel without entering a password.
- **Unauthorized Access to the VPN Data.** Obtaining secure information from a VPN is a third way in which VPN security may be corrupted. This security information includes IP addresses, configuration parameters and user license certificates for a VPN terminal. Access to this information may come from the insiders who know the specifics of a VPN, such as, people who have left or have been fired from the company. Most networks do not change frequently, and VPN connections remain in the same state for a long time. Therefore, people leaving the company have many opportunities to learn about specific ways to access the VPN. This security information can also be obtained through other social engineering methods, such as phishing or vishing.
- **Exploit Vulnerabilities in the System.** A possible defect in the firmware itself or some other weakness of the authentication system can be exploited, such as, malicious spoofing or redoing SSL authentication. It would even be possible for a hacker to use these well-known vulnerabilities in the VPN concentrator to crash the authentication system to invade the target system [5].

6 The Solutions for the Security Issues of Firewalls and VPN

6.1 The Solutions for the Security Issues of Firewalls

The Immune-Based Firewall System. After an intruder bypasses a firewall, they must control the firewall system or break the work of the firewall system. To achieve this goal, they must destroy the vital information of the firewall. Therefore, the immune-based firewall system security model centers on the critical information files of the firewall and uses the change of these files as a means to determine whether there is an intrusion. Because the intrusion is the “differences” in the firewall system.

The critical information file is the body of the firewall system. If an intruder destroys the body of the firewall system, the firewall will find and resist it to protect the critical files that are on record. It will also record the destruction and control of the network communication. The basic structure of the system security model is shown in Fig. 5 [26], centered on the critical information file of a firewall and file information database. It also uses the immune subsystem as the core to build a relatively perfect firewall system. However, when it comes to the actual application effect, the filtering mechanism of the firewall against these attacks is still not perfect yet, and there is no effective strategy to solve this problem. An ideal firewall filtering mechanism and the security policy model is shown in Fig. 6 [26].

Multi-Stage Filter. The Multi-stage filter uses multilevel filtering in the firewall to filter out all source routing packets and the fake IP sources at the level of packet filtering. The multi-stage filter is a technology that is now widely used by firewalls as

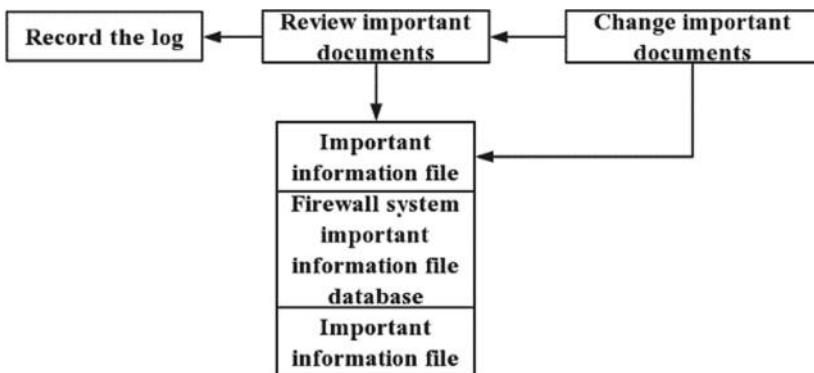


Fig. 5. The model of immune-based security firewall

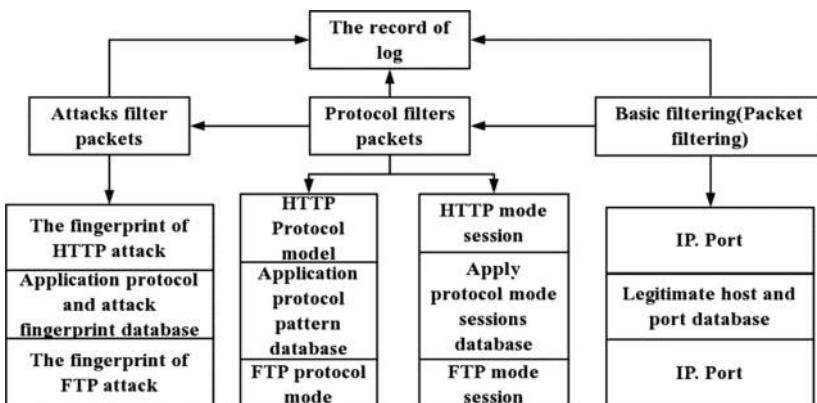


Fig. 6. Protocol-based firewall security policy model

packet filtering to efficiently help the protection of firewalls. This method is evident in the layer and can expand many new contents from this concept.

Next Generation Firewall. Next-Generation Firewall (NGFW) is the latest buzz in the firewall market at present. Through in-depth insight into users, applications, and content in network traffic, and with the help of a new high-performance single-path heterogeneous parallel processing engine, NGFW can provide users with active application layer integrated security protection. It can help users to conduct business safely and simplify their network security architecture. Application recognition is the most critical technology in the route. The technical route of NGFW is shown in Fig. 7 [27].

Secure Web Gateway. Secure Web Gateway (SWG) is a kind of product solution for Internet exploitation. It has the functions like URL filtering, malicious code protection, control functions, and the application control functions including the Web functions. This means it can enforce the enterprise's Internet access strategy while protecting it from the security threats. Most of the mainstream SWG products also

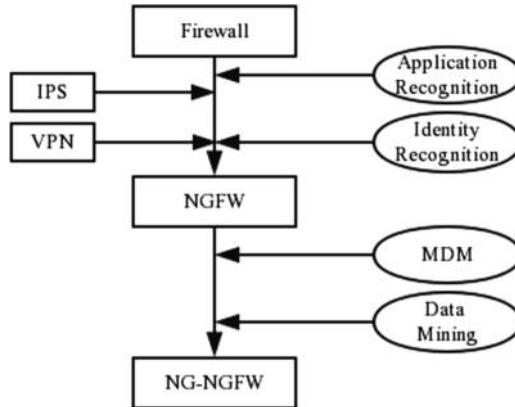


Fig. 7. Technical route of NGFW

provide the user identification and the DLP (Data Leakage Protection) function on this basis. Some company such as Intel uses SWG to protect their company's security. Figures 8 and 9 shows the structure of SWG firewall [28].

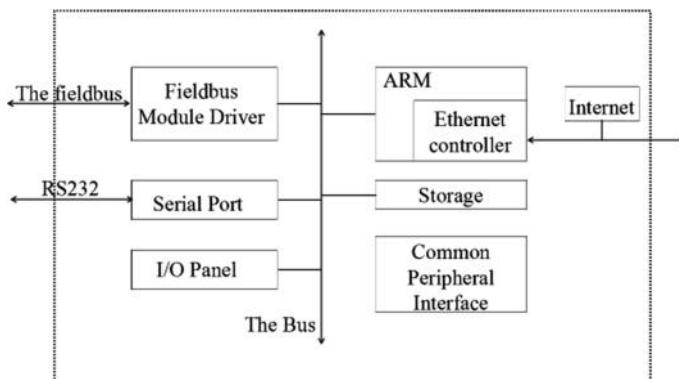


Fig. 8. The hardware structure of embedded gateway

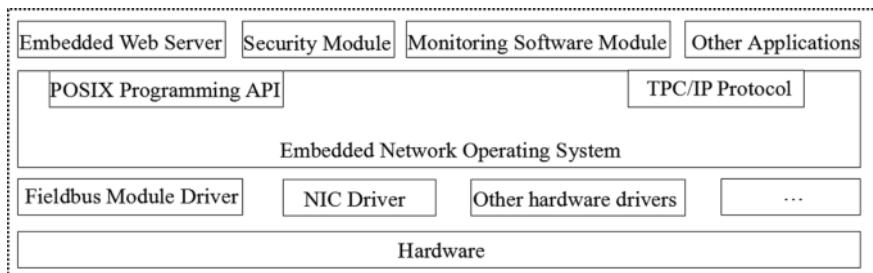


Fig. 9. The overall structure of gateway software

Web Application Firewall. Web Application Firewall (WAF) is mainly used to strengthen protection against web-specific intrusion methods such as DDoS attacks, SQL injection, XML injection, XSS, etc. WAF can be divided into front-end capture, rule setting and monitoring (brain), regulation action (monitoring or blocking), log storage/monitoring display, and corresponding processing unit as shown in Fig. 10 [29]. Currently, there are three types of WAFs in the market, namely: Hardware Web firewall, Web protection software and Cloud WAF.

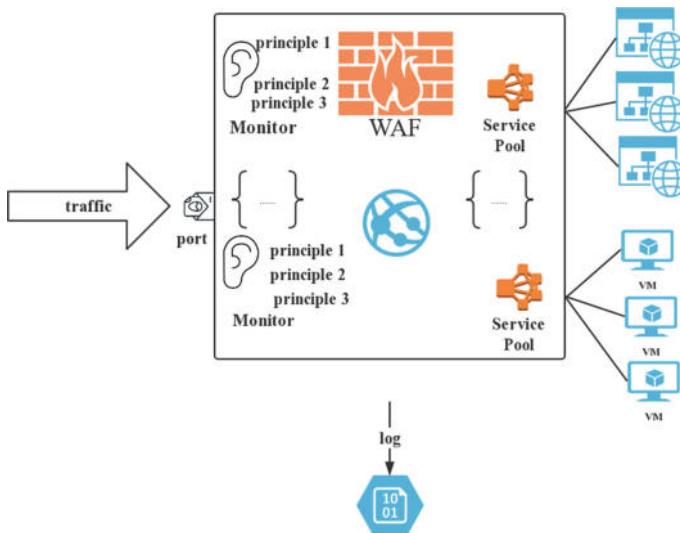


Fig. 10. WAF working principle

6.2 The Solutions for the Security Issues of VPNs

Wi-Fi wall. It is a useful technology to protect the VPNs when the user connects to the Wi-Fi. This technology can monitor the Wi-Fi traffic, and it can constantly check if there are attacks. The Wi-Fi wall will disconnect the Wi-Fi once an attack is detected.

Authentication service. The VPN service providers need the authentication service to help them to protect the identity information about the potential end-users [30].

Access control. This service can maintain the security and prevent the use of unauthorized VPN service features and access to the unauthorized resources [30].

Data integrity and confidentiality. This service can keep the integrity of information, prevent the leakage of information, and counter threats. The cryptographic hardware can protect the integrity and confidentiality of management data [30].

VPN audit requirement. The suitable auditing system is necessary to detect potential abuse, since the present security service and security mechanisms may be compromised or bypassed. Therefore, the hackers may gain the unauthorized access

and damage the VPN protected by them. Enumerating and understanding the VPN service behavior is necessary for providing enough information for studying the VPN auditing requirements [30].

VPN firewall. VPN Firewall is a kind of firewall that is installed at the server end or the front of a VPN server. It is configured with the filters only to let the VPN specific packets to access the network when installed at the server end of the VPN. However, when it is installed at the front of a VPN, it will only allow the tunnel data on its Internet interface to access the server [31].

HAIPE security gateway. In this model, VPN client edge device is intended to use network hardware encryption device HAIPE as a security gateway to protect the communication between VPN client sites. It is shown in Fig. 11 [32] that the VPN user network consists of A and B stations. The edge of station A is deployed with the HAIPE_A security gateway, and the edge of station B is deployed with the HAIPE_B security gateway. HAIPE_A and HAIPE_B are equal and establish ESP (Encapsulating Security Payload) encrypted tunnel between them, which let any information stream between station A and station B be protected by the ESP encrypted tunnel. Only the flow of the network to peer protection or the flow of the network from the peer protection network will pass through the gateway, and the rest of the traffic is stopped by the secure gateway of ESP encrypted tunnel.

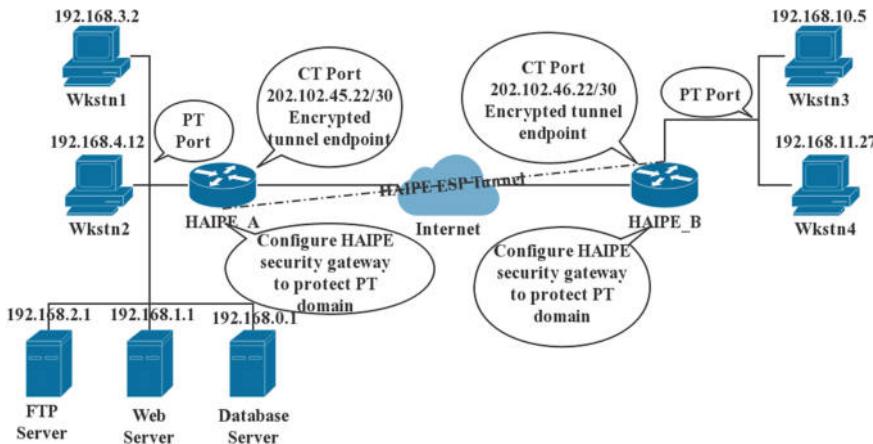


Fig. 11. Diagram of HAIPE security gateway protecting plain text (PT) domain communication

7 Conclusion and Future Work

As a result of this study, we have concluded that the potential threats and risks to the Internet and Intranet will keep growing and so does the development of firewalls and VPNs. With the increased implementation of technology in almost all possible domains of cyberspace, the security and protection of a network will need more attention with time. We also realized that the gap between the potential threats and the risks to these

systems are growing exponentially and the present-day firewalls and VPNs are not full proof yet. This results in defeating their purpose sometimes.

We also concluded that it is not easy to judge firewall or VPN against each other because there are many types of firewalls and VPNs available in the market and each one of them has its advantages and disadvantages. The kind of users and their demands regarding how much safety they want in their network and how much budget do they have to get what they need in a firewall or VPN is also the key in its implementation. A combination of firewalls and VPNs can always provide more security to a network than using them individually. The users also need to keep in mind that both VPNs and firewalls have quite a few security flaws and there are various solutions available to avoid these loopholes. Our advice is to consider the application environment and the user's expectation of performance carefully when choosing a firewall or a VPN or both for creating a secure network system. However, there are still some parts that we did not cover in this paper. Our following plan for this topic is to study some methods that can deal with the security issues both related to firewalls and VPNs since we have already researched the security threats, risks, and issues that both firewalls and VPNs have.

In all, we hope that our work can be used as a reference by the organizations and individuals when it comes to the solutions for the loopholes, threats, and risks related to a firewall and a VPN. Further study needs to be done to make a unified model of a secure firewall and VPN to fight the attacks and the attackers and save the data and the network from being destroyed or breached.

References

1. Ma, L., Liang, H.: Application of firewall technology in computer network security. *Comput. Knowl. Technol.* **10**, 3743–3745 (2014). Print
2. Jiang, C.: Research on computer network security technology and firewall technology. *Ability Wisdom* 235 (2017). Print
3. Su, J., Yuan, J.: Firewall technology and its development. *Comput. Eng. Appl.* 147–149 (2004)
4. Zhang, Z., et al.: VPN: boon or trap? A comparative study of MPLS, IPsec, and SSL virtual private network. In: *ICCMC 2018* (In Press)
5. Sun, J., Wei, J.: Computer Network Technology and Application. Xi'an University of Electronic Science and Technology, Xi'an, China (2010)
6. Stateful-inspection firewall: the Netscreen way. <http://www.netscreen.com/products/firewallwpaper.html>
7. Li, S., Tørresen, J., Sorensen, O.: Exploiting Stateful Inspection of Network Security in Re-Configurable Hardware
8. Suehring, S.: *Linux Firewalls: Enhancing Security with Nftables and Beyond*, Illustrated edn, p. 25. Pearson Education (2015). ISBN 0134000021
9. Successful case: Hanbo firewall to build a safe and efficient enterprise Intranet. *Network World* 2011-12-12 (021) (2011)
10. Qing, Y.: Shortcomings and improvements in firewall security. *Sci. Technol. Eng.* **14**, 1009–1012 (2005)

11. Sun, K.: Concrete application of firewall technology in computer network security. *Sci. Technol. Econ. Guide* **17**, 38 (2017)
12. Wang, D.: Research on Deep Packet Inspection Technology of Firewall. Xi'an University of Electronic Science and Technology (2005)
13. Zhang, T.: Design and Implementation of Firewall Based on Content Filtering Software. University of Electronic Science and Technology (2012)
14. Security analysis of VPN technology, 19 Sept 2014. Web. <http://sec.chinabyte.com/368/13082868.shtml>
15. Comparison between professional VPN and firewall with VPN. Web. <https://wenku.baidu.com/view/6e89ab1055270722192ef791.html>
16. Network security: a simple guide to firewalls. In: Network Security, pp. 2–3 (2000)
17. Yadav, A.S.S., et al.: Prevention of spoofing attacks in wireless networks. In: International Conference on Computing Communication Control and Automation, pp. 164–171. IEEE (2015)
18. Hollis, K.: The Rose Attack Explained. Retrieved on 2013-11-25
19. Sakurai, S., Ushirozawa, S.: Input method against Trojan horse and replay attack. In: IEEE International Conference on Information Theory and Information Security, pp. 384–389. IEEE (2010)
20. Chen, D., Wang, P.: Study and implementation of tunnel attack in MANET. *Comput. Eng.* **33**(9), pp. 140–141 (2007)
21. What is PASV FTP (passive FTP)? 13 Jun 2018. Web. <https://www.lifewire.com/definition-of-passive-mode-ftp-816441>
22. Zhao, T.F., et al.: Detecting rebound port Trojan based on network behavior analysis. *Netinfo Secur.* (2011)
23. VPN helps you scale the wall? Let's put our privacy first, 6 Jun 2016. Web. <https://www.jb51.net/hack/471867.html>
24. Conti, M., Dragoni, N., Lesyk, V.: A survey of man in the middle attacks. *IEEE Commun. Surv. Tutor.* **18**, 2027–2051 (2016)
25. Khan, M.M., Bakhtiari, M., Bakhtiari, S.: An HTTPS approach to resist man in the middle attack in secure SMS using ECC and RSA. In: 2013 13th International Conference on Intelligent Systems Design and Applications, pp. 115–120, Dec 2013
26. Yang, Z., Cheng, Q.: Research of immune-based technology for the firewall system security. *Microcomput. Inf.* **21**, 9-3
27. Network behavior management and the next generation of firewalls, SWG relations. Web. <https://jingyan.baidu.com/article/cbf0e50095f63a2eaa2893cc.html>
28. Zhao, Y., Du, Y.: Design and implementation of embedded secure web gateway. *Comput. Eng. Des.* **27**(4) (2006)
29. Sahin, M., Sogukpinar, I.: An efficient firewall for web applications (EFWA). In: 2017 International Conference on Computer Science and Engineering (UBMK), pp. 1150–1155 (2017)
30. Boukari, N., Aljane, A.: Security and auditing of VPN. In: Proceedings of Third International Workshop on Services in Distributed and Networked Environments, pp. 132–138, 6 Aug 1996
31. VPN firewall. Techopedia. 28 Jun 2017. Web. <https://www.techopedia.com/definition/30753/vpn-firewall>
32. Dian, A.: Security research and improvement of mainstream VPN technology (2009)



A Software Engineering Methodology for Developing Secure Obfuscated Software

Carlos Gonzalez^(✉) and Ernesto Liñan

Universidad Autónoma de Coahuila, Arteaga, Mexico
gonzalezc757@gmail.com, ernesto_linan_garcia@uadec.edu.mx

Abstract. We propose a methodology to conciliate two apparently contradictory processes in the development of secure obfuscated software and good software engineered software. Our methodology consists first in the system designers defining the type of security level required for the software. There are four types of attackers: casual attackers, hackers, institution attack, and government attack. Depending on the level of threat, the methodology we propose uses five or six teams to accomplish this task. One Software Engineer Team and one or two Software Obfuscation Teams, and Compiler Team. These four teams will develop and compile the secure obfuscated software. A Code Breakers Team will test the results of the previous teams to see if the software is not broken at the required security level, and an Intrusion Analysis Team will analyze the results of the Code Breakers Team and propose solutions to the development teams to prevent the detected intrusions. We present also an analytical model to prove that our methodology is no only easier to use, but generates an economical way of producing secure obfuscated software.

Keywords: Secure software development · Software engineering · Development methodology

1 Introduction

1.1 Assumptions

When one is developing software that needs protection from ill-intentioned users, that is, the user wants to access your software and reverse engineer your algorithms and methodology, or simply have access to your software without any of your restrictions imbedded in the software (i.e. need for a password, maximum activation time). The software developer has to be aware of the level/kind of threat he wants to avoid. The decision of which level of security to have in the developing software should be made by a group of persons, which should include:

1. Senior Software Engineer
2. Senior Obfuscation/Security Expert
3. Financial Executive
4. Administrator

We assume that a decision was made to use obfuscation as a methodology to implement the security in the developing software.

1.2 Threat Levels

For the case the attacker has access to the source code or both. From the security point of view, this case is the worst scenario. Having the source code available to the attacker facilitates such attacker a substantial advantage.

- Level-1: A casual attacker. The attacker has the software and he/she is not technically knowledgeable to retrieve data or algorithms from the machine code software. The attacker is probably capable of finding on the source code the algorithms after some work.
- Level-2: A hacker attack. This attacker has the knowledge to retrieve data or algorithms from both kinds of sources of the software.
- Level-3: An institution attack. This attack is done by an institution with all the resources of such institution.
- Level-4: A government attack. This attack is done by a government agency with all the resources (technical and legal) available for such agency.

For the case of the attacker having access only to the machine code. This should be the most common case for any secure software developed.

- Level-1 attacks to the machine code are the easier to handle. Probably no obfuscation is needed. Only if the source code is available then some obfuscation will be needed.
- Level-2 attacks need to have obfuscation used for the development of both code sources.
- Level-3 attacks need a higher level of obfuscation. It is recommended if the software is expected to be attacked at this level, that the source code will never be available to the attacker. That is, never release source code, and should be treated at the secret level in the business or government agency.
- Level-4 attacks need the highest level of obfuscation, beginning with the obfuscation of the source, and then the obfuscation of the machine code. In addition, the source code should never be released, and should be treated the secret or higher level.

1.3 Conflicts

Once we have the type and level of attack defined and defined that obfuscation is the security process to use, the question to be answered is: How do we reconcile the development of software using Software Engineering [17, 24, 28–30] techniques like, name variable with meaningful names, place comments on every instruction, etc., versus the need to obfuscate our software? [4, 5, 9–11, 16, 20, 23, 32]. We would like to point out that even if we finally have a cryptography breakthrough that could make software un-hackable with the indistinguishability obfuscation model [2, 3]; we still need a methodology to incorporate such model into our software development plan.

One should never develop obfuscated software from scratch. If one does not have the developed software that is well documented and easy to manipulate, the modifications and maintenance of such software is extremely costly. You may end up redoing all the software rather than reusing a bad documented and difficult to read software. On the other hand, this is exactly what you need to have a secure obfuscated software.

Therefore, what we propose with this methodology is to split the development of secure software into two phases: The Software Engineering development of the software, and a second phase to do the obfuscation. The aim of doing this split is to get the best-combined results out of the two developing processes.

Once software applications are implemented and start being used, several changes will occur over time [8].

- As bugs and defects are found, they need to be fixed.
- As the business evolves, new features and requirements will be needed.
- New government mandates.
- The natural aging of the software.

Therefore, another important implication in using this methodology is that any changes and modifications of the original developed software must go through the two phases again (i.e. normal software development and obfuscated development).

1.4 Software Teams

Our methodology consists in the division of work between several teams of software engineers.

- The Software Engineering Developing team (SED), developing the software following all the guidelines and best practices described in Software Engineering [8, 19].

An example of such best practice is to develop iteratively because critical risks are resolved before increasing the costs. This best practice also calls for continuous testing and integration. With our methodology, we propose a testing phase, which is not a contradiction since it can be the accumulation of all the tests done in the iteration. In other words at the end of all design iterations we will have a set of tests for the software. Since we are proposing to do the obfuscation separate from the normal development, we need to test the results of the obfuscation, and this set of tests will be used for these purposes.

Our methodology proposes iteration between the obfuscation development and the testing of such development.

The software development methodology used such as Agile [1, 6], or Waterfall [26, 27], or TSP/PSP [18, 31] by this team is usually selected by the senior Software Engineers in charge of the project, following the given guidelines of budget and time.

- A Compiler Team (CT). This team is in charge of compiling all the software generated by the developing teams. We put it here only for completeness. Most of the time the same team that does the development is the one that performs the compile.

- A team of Obfuscation High Level Code experts (OBC) which will obfuscate the code produced by the SED. This team will be in charge of doing the obfuscation of the high level code produced by the SED, using either automated tools [12–15], semi-automated tools like Crypto Obfuscator [25] which uses Incremental Obfuscation, or by hand. The main focus of this team will be to satisfy any security requirements of the software including the resistance to the required level of threat.
- A team of Obfuscation Machine Code experts (OBM) which will obfuscate the product generated by the OBC [7, 21, 22]. This second team of obfuscation experts will be needed if the threat level to resist is level-3 or level-4. This level usually is not necessary for hacker at level-2. The obfuscation of the high level code should be enough. If there is any doubt then assume threat level-3.
- A Specification and Requirements Testing Team (SRT), which will test the software against the requirements and specifications of the project. This team will be testing the requirements of the system. If the methodology is iterative, then the set of test performed is saved to be used in its entirety after the obfuscation teams are done. A Code Breakers Team (CBT), which will try to break the obfuscated code generated by the OBM. This team will have the responsibility of making sure or as close as possible that the secure obfuscated software can resist the required level of threat. Since the idea with the Code Breakers Team is to simulate the real-world threat as much as possible, then for the testing of level-2 threats the Code Breaking Team should be of single individuals acting as hackers (some teams may be of two or three people). The hacker is usually alone or with two-three accomplices. The results of each individual team will be reviewed, evaluated, and delivered to the Intrusion Analysis Team.

For level-3 and level-4 threats, the whole team of code breakers should work together. A time line with deadlines should be set so the code breaking process will not be extended indefinitely. Unfortunately, in real life the time line may not exist or be much longer for industry and government threats.

For threats of level-3, the goal is to make breaking the code cost-ineffective for the industry trying to do it. This goal makes sense since in most cases the main reason behind level-3 threats is economic gains.

- An Intrusion Analysis Team (IAT), which will analyze the results of the CBT and generate new obfuscation proposals for the obfuscation team OBC. This team has an important and difficult role, since it has to analyze the problems, and most of the time the solution will be an innovation of new techniques to solve the security problem. Therefore, the members of this team should not only be experts in obfuscation, but also highly innovative.

2 Methodology Process

Figure 1 shows the software methodology process proposed in this paper.

Once a piece of software is liberated by the SED team through the Compiler Team, it goes to the SRT to test all the requirements and specifications of the project. If it fails, the results go back to the SED to correct the problems with the failed

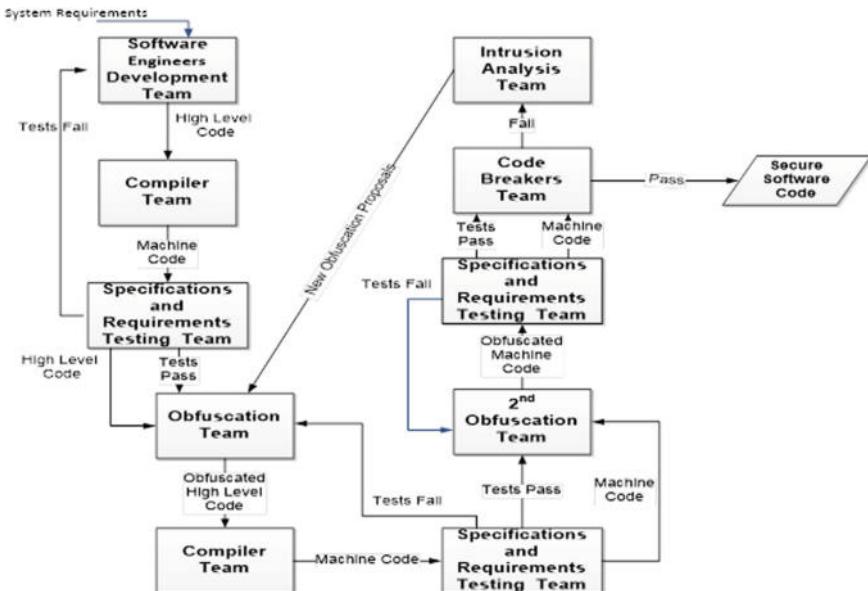


Fig. 1. The software methodology process

requirements/specifications. If the SRT passes all the tests, then the high level code is passed to the OBC team to work on it and obfuscate it.

Again, it is important to be careful with this process, since any changes/modifies to the source code by the SED team after delivery, means the SRT and OBC will have to be processed again. If this cycle is repeated very frequently by any piece of software, the cost of development could be highly increased. As has been pointed out [8] defect removal low efficiency is a major harmful practice that has been a drain on the software industry.

Once the obfuscation is done by the OBC, the software is compiled (using the Compiler Team). This software has to be tested again to check that modifications done by the OBC process did not break any of the requirements/specifications test. Thus, the obfuscated code generated by the OBC and compiled is passed to the SRT to again re-test all the requirements/specifications.

If the SRT tests fail, the results go back to the OBC team to modify the necessary code to pass the failed requirements/specifications. If on the other hand all the SRT.

If a second obfuscation is required because of the level of protection needed, then the code generated by the compiled OBC code is passed to the OBM team and such team gets to work on the machine code for the second obfuscation.

If no second obfuscation is required, and all the requirement tests done on the OBC coded are passed, this code is sent directly to the Code Breakers Team.

Again, if at this point there are any changes in the source code, means doing the two obfuscations again. Therefore, the changes should be minimized.

Once the second obfuscation is done (if needed), the obfuscated code is passed to the SRT to again re-test all the requirements/specifications. If the SRT tests fail, the

results go back to the OBM team to modify the necessary code to pass the failed requirements/specifications. If on the other hand all the SRT tests are passed, then the code is ready for the code breaking tests.

A Code Breakers Team will test and try to break the supplied obfuscated code. If the Code Breakers determines that the provided obfuscated code passes the tests, then the obfuscated code is ready for delivery as secure software. If on the other hand the Code Breakers are able to break the obfuscated code, then the test has failed, and the code and the results of the failed test are passed to an Intrusion Analysis Team, which will analyze the reasons for the failure and propose new obfuscations algorithms or modifications to the old ones obfuscation transformations to be applied to a software system. to stop the failures. This information is passed to the first Obfuscation Team (OBC) to work on, and a new cycle of obfuscated code begins.

3 The Analytical Model

Let S_0, \dots, S_n be the different $n + 1$ versions of software S through its lifecycle, which is starting with the first version S_1 to the last one S_n , and S_0 empty. Let a function named G , which $G(S_i) = S_{i+1}$. This function generates the next version of S_i that is S_{i+1} . The $EF(S_i)$ is equal to effort, in man-hours, needed to generate next version from software version S_i to S_{i+1} (i.e. generating version S_{i+1}). Then $EF(S_0)$ is equal to total effort to generate the first version of software S , so, $TE = \sum_{k=0}^{n-1} EF(S_k)$, which is the total effort used to generate the last version S_n of software S .

Let a function named $E(S_i, RE, I)$, be the effort needed for intruder I to Reverse Engineer (RE) software S_i . $R(S_i) = \text{set of } m \text{ requirements } (r_{i1}, r_{i2}, \dots, r_{im})$ needed to produce software S_i and $1 \leq m$.

$T(S_i, r_{ij}) = \text{set of tests } (t_{i1}, \dots, t_{ik})$ to be performed on software S_i to satisfy the requirement r_{ij} , where $r_{ij} \in R(S_i)$ and $k > 0$

$$P(S_i, r_{ij}, T) = \begin{cases} \text{true} & \text{if the result of applying the set of tests } T \text{ for} \\ & \text{requirement } r_{ij} \text{ of software } S_i \text{ are all true} \\ \text{false} & \text{if at least one test } \in T \text{ is false} \end{cases}$$

Let O be a set of obfuscation transformations to be applied to a software system.

$F(S_i, O) = Y_i$ transformation of software S_i by obfuscating it using obfuscation transformations O , producing software Y_i such that

$P(S_i, r_{ij}, T) = P(Y_i, r_{ij}, T)$ for $j = 1$ to m , (the requirement test are all the same) and $E(S_i, RE, I) < E(Y_i, RE, I)$ (is harder to reverse engineer)

$EO(S_i)$ is the effort to obfuscate software version S_i and get obfuscated software Y_i

$H(Y_i) = Y_{i+1}$ function H transforms obfuscated software Y_i into Y_{i+1}

$EY(Y_i) = \text{effort to generate obfuscated } Y_{i+1} \text{ starting from obfuscated version } Y_i$

In general for a developer D to reverse engineer a software version requires more effort than creating such version.

$$E(S_i, RE, D) \geq EF(S_{i-1})$$

We also claim that the effort to generate version $i + 1$ of software S is going to be easier starting from the previous non-obfuscated version of the software S_i rather than starting from obfuscated version Y_i . By definition an obfuscated software (like Y_i) is difficult to understand and follow its logic, therefore making changes is going to take more effort than using a well documented and clear software like S_i .

Therefore,

$EF(S_i) < EY(Y_i)$ —Effort for the new obfuscated code for Y_{i+1}

If we add all the inequalities for $i = 0$ to $n-1$ we have:

$$\begin{aligned} \sum_{k=0}^{n-1} EF(S_k) &< \sum_{i=0}^{n-1} EY(Y_i) \\ - \sum_{i=0}^{n-2} \text{Effort for the new obfuscated code for } Y_{i+1} \end{aligned}$$

Since $EO(S_i)$ for $i = 0$ to $n-1$ is a positive number then

$\sum_{i=0}^n EO(S_i)$ is also a positive number then we can add to the above inequality, getting:

$$\begin{aligned} \sum_{k=0}^{n-1} EF(S_k) + \sum_{i=0}^n EO(S_i) &< \sum_{i=0}^{n-1} EY(Y_i) \\ - \sum_{i=0}^{n-2} \text{Effort for the new obfuscated code for } Y_{i+1} + \sum_{i=0}^n EO(S_i) \end{aligned}$$

All algorithms and procedures for obfuscating software start with a non-obfuscated source. Then the effort of the new obfuscated software for Y_i starting from Y_{i-1} is going to be larger than the effort to generate the new obfuscated software starting from S_i , therefore,

$$- \sum_{i=0}^{n-2} \text{Effort for the new obfuscated code for } Y_{i+1} + \sum_{i=0}^n EO(S_i) < 0$$

Since this factor is always going to give us a negative number we could remove it and the inequality still holds.

Thus,

$$\sum_{k=0}^{n-1} EF(S_k) + \sum_{i=0}^n EO(S_i) < \sum_{i=0}^{n-1} EY(Y_i)$$

Or,

$$TE + \sum_{i=1}^n EO(S_i) < \sum_{i=0}^{n-1} EY(Y_i)$$

This means that the total effort to create a series of obfuscated software versions is going to be more economical using our model than starting from obfuscated versions to create the next one. Figure 2 shows the mapping between the analytical model and the software development methodology process explained in Sect. 2.

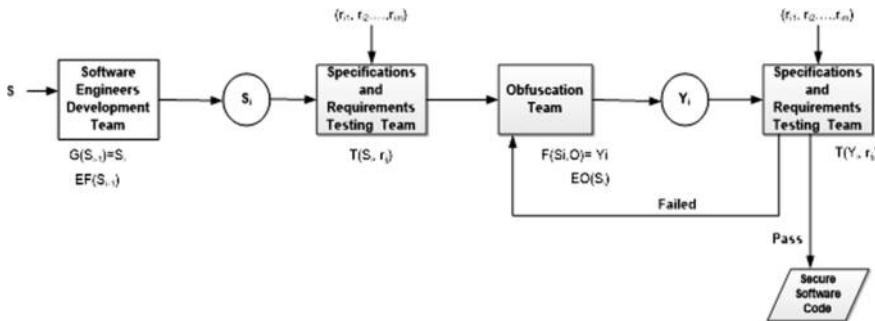


Fig. 2. The mapping between our analytical model and our software development methodology

4 Conclusions, Recommendations and Future Work

We have not tested our methodology in a real-life environment because it is very complex and expensive for our limited resources, but we have proved with an analytical model (something that we know is intuitive obvious) that our methodology is a Software Engineering best practice when developing obfuscated secure software.

The main contribution of this paper is the description of a new software engineering methodology to develop secure obfuscated software and carry out the development following software engineering concepts. The two techniques by themselves use contradictory techniques, and our methodology solves this contradiction and shows how to use both techniques for the development of the desired secure software. The methodology process is done in a manner that is easy to understand, robust, feasible and much more friendly to modifications (compared with doing the modifications on the obfuscated software) during development of the secure software, and in future when modifications need to be made to such software.

In this paper, we have also proven using analytical methods that the methodology proposed here results in a simpler and more economical effort in the development of obfuscated secure software over the lifecycle of the secure obfuscated software.

The methodology describes the types of threats one may encounter, and how to deal with them.

Another important contribution of this methodology is the use of Code Breakers Team in conjunction with Intruder Analysis Teams to make sure the software is secure for the required level of threat.

If this methodology is used for secure obfuscated software development, we recommend the following general guidelines:

- Isolate the various teams (mainly the SED and OB teams). They should be in different buildings or better in different cities.
- Make sure that the team leader for each team is an expert in the field, and knows the whole methodology approach.
- Make sure that the customer as well as all the members of the SED team understands the cost of late changes.
- For the Code Breakers Teams is recommended to include at least one of each of the SED and the two OB teams. The idea behind is that we should test that even if an intruder has some knowledge of the contents of the software, the intruder should not be able to break the obfuscation.
- Protect locally the source code and its documentation.
- Protect locally the algorithms and procedures used in both obfuscations. If an intruder gets to know what algorithms you used for obfuscation, then the reverse engineering for the intruder could become much simpler.

Finally, we are planning as future work for this topic the design of a more generalized method that includes not only obfuscation but also other security techniques to design secure software. That is, a methodology using Software Engineering practices to design and implement standalone secure software.

References

1. Alliance, A.: What is Agile Software Development? (June 2013)
2. Sahai, A., Waters, B.: How to use indistinguishability obfuscation: deniable encryption, and more (2013). <http://eprint.iacr.org/2013/454.pdf>
3. Sahai, A., et al.: Candidate indistinguishability obfuscation and functional encryption for all circuits (2013). <http://eprint.iacr.org/2013/451.pdf>
4. Aucsmith, D.: Tamper resistant software: an implementation. In: Proceedings of the 1st International Information Hiding Workshop (IIHW), Cambridge, U.K., pp. 317–333. Springer LNCS 1174 (1996)
5. Barak, B., Goldreich, O., Impagliazzo, R., Rudich, S., Sahai, A., Vadhan, S., Yang, K.: On the impossibility of obfuscating programs. In: Advances in Cryptology–Crypto 2001, pp. 1–18. Springer LNCS 2139 (2001)
6. Beck, K., et al.: Manifesto for Agile Software Development. Agile Alliance. Retrieved 14 June 2010 (2001)
7. Bernat, A.R., Roundy, K.A., Miller, B.P.: Efficient, sensitivity resistant binary instrumentation. In: International Symposium on Software Testing and Analysis (ISSTA), Toronto, Canada (2011)
8. Jones, C.: Software Engineering Best Practices: Lessons from Successful Projects in the Top Companies. McGraw-Hill (2010)

9. Collberg, C., Thomborson, C., Low, D.: A Taxonomy of Obfuscating Transformations. Technical Report 148, Dept. Computer Science, University of Auckland (July 1997)
10. Collberg, C., Thomborson, C., Low, D.: Manufacturing cheap, resilient, and stealthy opaque constructs. In: Proceedings of the Symposium on Principles of Programming Languages (POPL '98), (Jan 1998)
11. Collberg, C.: Surreptitious software exercise, attacks, breaking on system functions. Department of Computer Science, University of Arizona, February 26 (2014)
12. Dmoz.org: Open Directory—Computers: Programming: Component Frameworks: .NET: Tools: Obfuscators, 2007-01-02. Retrieved 2013-11-25 (2007)
13. Dmoz.org: Open Directory—Computers: Programming: Languages: Java: Development Tools: Obfuscators, 2013-04-09. Retrieved 2013-11-25 (2013)
14. Dmoz.org: Open Directory—Computers: Programming: Languages: JavaScript: Tools: Obfuscators, 2013-08-03. Retrieved 2013-11-25 (2013)
15. Dmoz.org: Open Directory—Computers: Programming: Languages: PHP: Development Tools: Obfuscation and Encryption, 2013-09-19. Retrieved 2013-11-25 (2013)
16. dreamincode.net: A Simple Introduction to Obfuscated Code. <http://www.dreamincode.net/forums/topic/38102-obfuscated-code-a-simple-introduction/>. Posted 25 November 2007
17. Martin, F., Beck, K., Brant, J., Opdyke, W., Roberts, D.: Refactoring: Improving the Design of Existing Code. Boch Jacobson Rumbaugh (1999)
18. Humphrey, W.: The Team Software Process (PDF). Software Engineering Institute (Nov 2000)
19. IBM: Best practices for software development projects. http://www.ibm.com/developerworks/websphere/library/techarticles/0306_perks/perks2.html. Accessed 10 August 2006
20. Kenter, A.: Obfuscation. <http://www.kenter.demon.nl/obfuscate.html>. Visited 18 August 2015
21. Roundy, K.A., Miller, B.P.: Binary-Code Obfuscations in Prevalent Packer Tools (Sep 2011). <http://ftp.cs.wisc.edu/pub/paradyn/papers/Roundy12Packers.pdf>
22. Linn, C., Debray, S.: Obfuscation of executable code to improve resistance to static disassembly. In: Conference on Computer and Communications Security. Washington, DC (2003)
23. Mateas, M., Montfort, N.: A box, darkly: obfuscation, weird languages, and code aesthetics. In: Proceedings of the 6th Digital Arts and Culture Conference, IT University of Copenhagen, pp. 144–153, 1–3 December 2005
24. McConnell, S.: Code Complete: A Practical Handbook of Software Construction, 2nd edn, Microsoft (2004)
25. Microsoft: Crypto Obfuscator For.Net, version 2013.2, updated 7/25/2013
26. MIL-STD-498: Military Standard: Software Development And Documentation, United States Department of Defense (5 Dec 1994)
27. Oxagile.com: Waterfall software development model (Feb 2014). <http://www.oxagile.com/company/blog/the-waterfall-model/>
28. Patterson, D., Fox, A.: Engineering software as a service: an agile approach using cloud computing. Strawberry Canyon LLC (2013)
29. Pressman, R.S., Maxim, B.R.: Software Engineering: A Practitioner's Approach, 8th edn, McGraw Hill (2014)
30. Somerville, I.: Software Engineering, 9th edn, Addison-Wesley (2011)
31. Chick, T.A., et al.: Team Software Process (TSP) Coach Mentoring Program Guidebook Version 1.1. Software Engineering Institute, Report CMU/SEI-2010-SR-016 (2010)
32. Ogiso, T., Sakabe, Y., Soshi, M., Miyaji, A.: Software obfuscation on a theoretical basis and its implementation. IEEE Trans. Fundam. Electron. Commun. Comput. Sci., 176–186 (Jan 2003)



Detecting Windows Based Exploit Chains by Means of Event Correlation and Process Monitoring

Muhammad Mudassar Yamiun^(✉), Basel Katt,
and Vasileios Gkioulos

Department of Information Security and Communication Technology,
Norwegian University of Science and Technology, Gjøvik, Norway
{muhammad.m.yamin, basel.katt, vasileios.gkioulos}
@ntnu.no

Abstract. This article presents a novel algorithm for the detection of exploit chains in a Windows based environment. An exploit chain is a group of exploits that executes synchronously, in order to achieve the system exploitation. Unlike high-risk vulnerabilities that allow system exploitation using only one execution step, an exploit chain takes advantage of multiple medium and low risk vulnerabilities. These are grouped, in order to form a chain of exploits that when executed achieve the exploitation of the system. Experiments were performed to check the effectiveness of developed algorithm against multiple anti-virus/anti-malware solutions available in the market.

Keywords: Exploit chain · Event correlation · Process monitoring · Windows · Process correlation

1 Introduction

Recently, the Pwn2Own 2018 researchers introduced multiple Zero Day exploits, which were primarily based on a chain of multiple exploits for the exploitation of systems and services [1]. Traditional anti-virus and anti-malware software uses process monitoring and process isolation techniques for detection, according to suspicious process behavior pattern [2]. Yet, as we see in the Pwn2Own 2018 results, the researchers were able to break such process isolation and sandboxing process protection techniques. Examples of exploits that cannot be detected using traditional techniques are the guest-to-host exploits [3], and the macro-less DDE (dynamic data execution) in an MS office application [4]. In this article, we present a novel technique for the detection of such exploits using process execution monitoring my means of event correlation. The technique performs detection in a signature free and fully autonomous manner, using only the process names for monitoring and detection of exploitations. We use event correlation with respect to events extracted from process monitoring logs to create a chain of suspicious processes generated by the application to identify a detection. This article is organized in six sections. The first section introduces the problem, while in the second section we discuss related work and provide additional information about the

problem background. The following sections present the proposed algorithm and initial experimentation results, while in the last sections we provide a discussion, future work and conclude the article.

2 Related Work

The authors were not able to identify in the literature any viable existing technique for the detection of complex exploit chains such as guest-to-host exploits, while most cloud security vendors use defense in depth architectures to avoid security incidents involving guest-to-host exploits [5]. Existing techniques for securing a host from guest-to-host exploits use a multistep approach. Initially, an external process hook or agent is added in each virtual machine, which is then updated for malware and virus definitions from an external source [6]. Another technique used for securing virtual machines, relies on VMI (Virtual Machine Introspection) based process monitoring, for malware detection on a virtual machine from an external source [7]. Graph based event correlation (on the virtual machine) for anomaly detection using machine learning techniques [8]. The problem faced by existing techniques, is that they mostly focus on the protection of the virtual machine, without taking into account the new guest-to-host exploits, which exploit guest isolation using an exploit chain and allow the guest virtual machine to access the host operating system. Furthermore, in respect to the macro-less DDE in MS office applications [4], the research focus is on using malicious PowerShell commands for exploiting the system. For the detection of malicious PowerShell commands, researchers are currently using machine-learning techniques [9]. Yet, such existing detection techniques are vulnerable to the use of command line obfuscation for avoiding detection [10].

3 Problem Background

To further explain the problem a brief technical background is given.

3.1 Exploit Chains

In a normal IT security environment one vulnerability is enough to compromise the security of a system. However, due to continue system security improvements finding such vulnerabilities is becoming harder day by day. On the other hand low impact vulnerabilities are usually easy to find, researcher demonstrated multiple exploits which use these low impact vulnerabilities [4, 5], and chain them together to compromise system security. To further explain the flow of a single exploit and an exploit chain we created a simple flow chart for easy understating. Flow chart showing comparison of traditional exploits and an exploit chain flow is given in Fig. 1.

In comparison to a vulnerability that is exploited by a single exploit, in an exploit chain multiple vulnerabilities are involved. Each exploit uses the output of the previous to accomplish the objectives. A flow chart representation of exploit chain is seen in the Fig. 2.



Fig. 1. Example of traditional exploit with a single vulnerability

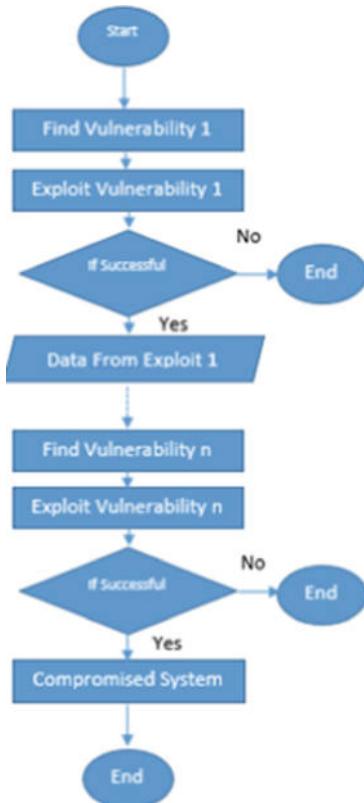


Fig. 2. Example of exploit chain with multiple vulnerability

Similar concepts exists in literature such as attack chain or attack paths which is set of possible steps that an attacker could take to compromise a system, involving multiple nodes on which exploitation is performed. In contrast, exploit chaining is the process of linking multiple vulnerabilities of one node which are present in a system and executing them in a specific order to compromise security.

3.2 Window Event Logging Mechanism

The Microsoft Security Event logging mechanism is present in every new release of Windows since Windows XP. This event logging mechanism allows the identification of the type of computer events happening in Windows based systems when an exploit is executed. Researchers at JPCERT [11] provided details of such security events in their technical report. In this research paper we focus on Event ID 4688 [12]. Which is a Windows new process creation event. Each 4688 event contains the following fields

- **SubjectUserId:** Security id of account from where the process is executed
- **SubjectUserName:** Account name from where the process is executed
- **SubjectDomainName:** Domain Name
- **SubjectLogonId:** Logon id of account from where the process is executed
- **NewProcessId:** Unique hexadecimal new process identifier
- **NewProcessName:** New process name executed by parent process
- **ProcessId:** Unique hexadecimal process identifier
- **CommandLine:** Command which is executed
- **TargetUserId:** Security id of account on which process executed
- **TargetUserName:** User name
- **TargetDomainName:** Computer name
- **TargetLogonId:** Login id of account on which process executed
- **ParentProcessName:** Name of process which executes new process
- **MandatoryLabel:** Secure object control integrity label assigned to new process.

From the information present in the fields of 4688 event we used NewProcessId, ProcessId, TargetDomainName in our detection algorithm. The ProcessId is a unique identifier issued by computer operating system to a running process. NewProcessId is a unique identifier issued by computer operating system to a process that is executed by another running process. TargetDomainName is the unique name of the computer on the domain.

3.3 Guest-to-Host Exploit

A recent report from SpiceWork [13] shows that server virtualization adoption reached 85% in comparison to 15% of physical IT infrastructure in 2017, as seen in Fig. 3.

This trend leads security researchers to develop exploits that can break guest isolation and compromise the host machine. A list of few vulnerabilities is given below

- CVE-2017-4924: An out of bound memory corruption vulnerability in Vmware 12. x to 12.5.7 Implementation of SVGA (Virtual graphic card) allows attackers to execute code host system

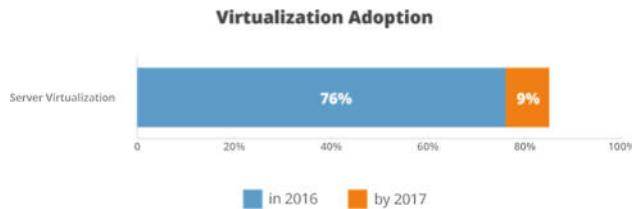


Fig. 3. [13] Server virtualization trends

- CVE-2017-4934: An heap buffer overflow vulnerability in Vmware 12.x to 12.5.8 Implementation of VMNET (virtual machine network) allows attackers to execute code host system
- CVE-2017-4936: An out-of-bounds read vulnerability in Vmware 12.x to 12.5.8 JPEG2000 parser in the TPView.dll allows guest to execute code or perform a DOS (Denial of Service) on the Windows OS.

A detailed list of guest-to-host escape vulnerabilities can be found online [14]. An example of these vulnerabilities is CVE-2017-4924 in which an out of bound memory corruption in vmwar-vmx.exe with incorrect memory mapping exists. This allows Data Execution Prevention bypass which leads to code execution on host from virtual machine.

Exploit writers were able to exploit this vulnerability and they created a POC (Proof of Concept) [15] for its exploitation. In the POC first the guest isolation is escaped by out of bound memory corruption and then CMD is executed by exploiting host Windows task registry. From CMD, PowerShell is executed to achieve remote shell level access on host. Schematically the exploit chain presented in Fig. 4.

Ideally the virtualization provide isolation between Guest OS and Host OS, where only the relevant services are shared as seen in Fig. 5.

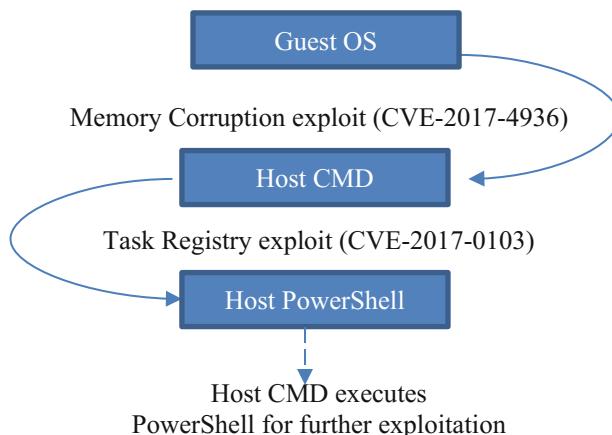


Fig. 4. Guest to host escape exploit chain

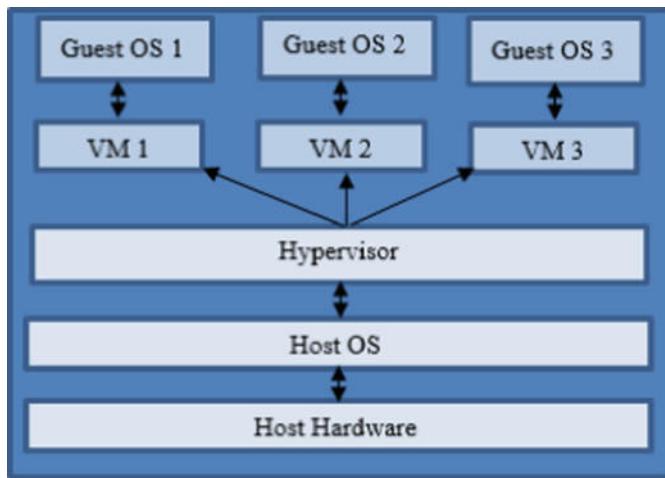


Fig. 5. Isolated guest and host in virtualized environment

But when a guest to host exploit is executed the isolation between Guest OS and Host OS is bypassed as seen in Fig. 6.

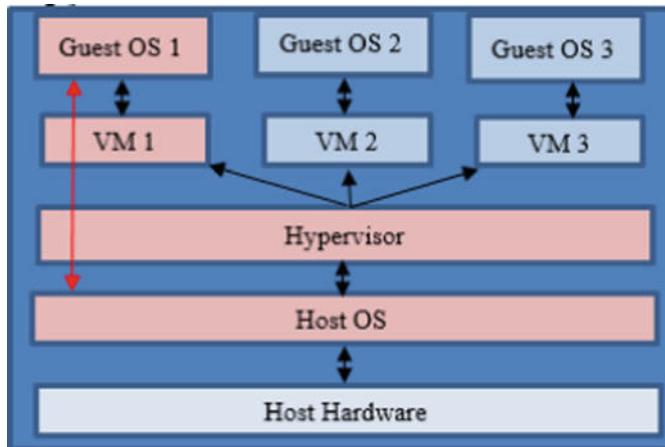


Fig. 6. Broken isolation between guest and host

3.4 Macro-Less DDE Attacks

To transfer data between different applications Windows provides the functionality of Dynamic Data Execution. The communication or COM Objects of Microsoft word and Microsoft excel, have public access to this DDE functionality. The functionality allows Microsoft Word and Excel to execute system commands legitimately. Exploit writers misused this functionality and were able to develop complex exploits such as macro-

less DDE code execution [4]. It is also very difficult to detect with traditional detection techniques since the functionality is legitimate feature and is not blocked and patched by Microsoft [4]. Anti-virus and anti-malware solutions are using signature-based detection mechanism for the detection of macro-less DDE but the signature-based detection was also easily bypass able using command obfuscation techniques [9]. The exploit execution of macro-less DDE is similar to guest-to-host escape but in this case Microsoft Word or Excel is used to create exploit chain. First DDE on Microsoft Word or Excel is exploited which allows the exploited process to use COM object in Windows and pass data related to secondary logon elevation vulnerability in windows through which CMD is executed. Now when the CMD process is started a command line argument containing malicious PowerShell script is passed to obtain a remote shell of the host a schematic repression of the explication chain is seen in Fig. 7.

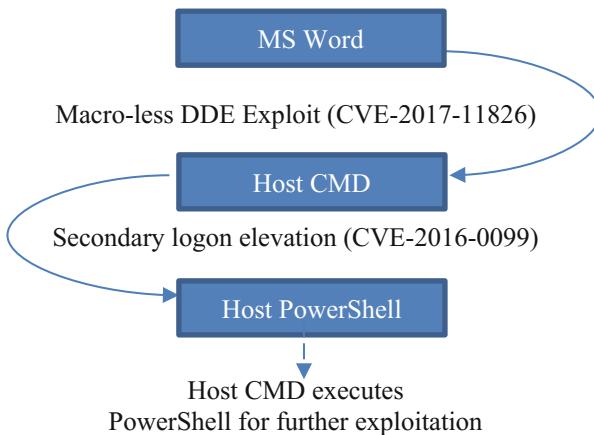


Fig. 7. Guest to host escape exploit chain

4 Detection Methodology

The detection algorithm is developed by analyzing Windows security logs. Consider the following Windows security logs of Vmware guest-to-host Escape exploit. It breaks the Guest isolation, executes a CMD command on the host to run a PowerShell Exploit. The logs generated by the exploit can be seen in Fig. 8.

After analyzing the logs a clear link is established between the processes generated by the exploit, as the ProcessID of a process is the NewProcessID of previous process involved in the exploit chain. We identified that by co-relating multiple events based upon the relation of ProcessID and NewProcessID we can create a process execution chain of the exploit. Accordingly a detection algorithm has been developed based on this finding. The proposed algorithm works in the following manner:

```

+ System
- EventData
  SubjectUserId S-1-5-21-703565726-1159332285-768548448-1001
  SubjectUserName Mudassar
  SubjectDomainName IPHONE
  SubjectLogonId 0x8abbe
  NewProcessId 0xf18
  NewProcessName C:\Windows\SysWOW64\WindowsPowerShell\v1.0\powershell.exe
  TokenElevationType %%1938
  ProcessId 0xa10
  CommandLine powershell
  TargetUserId S-1-0-0
  TargetUserName -
+ System
- EventData
  SubjectUserId S-1-5-21-703565726-1159332285-768548448-1001
  SubjectUserName Mudassar
  SubjectDomainName IPHONE
  SubjectLogonId 0x8abbe
  NewProcessId 0xa10
  NewProcessName C:\Windows\SysWOW64\cmd.exe
  TokenElevationType %%1938
  ProcessId 0x270
  CommandLine "C:\Windows\System32\cmd.exe"
  TargetUserId S-1-0-0
  TargetUserName -
+ System
- EventData
  SubjectUserId S-1-5-21-703565726-1159332285-768548448-1001
  SubjectUserName Mudassar
  SubjectDomainName IPHONE
  SubjectLogonId 0x8abbe
  NewProcessId 0x270
  NewProcessName C:\Program Files (x86)\VMware\VMware Workstation\vmware.exe
  TokenElevationType %%1938
  ProcessId 0x1b9c
  CommandLine "C:\Program Files (x86)\VMware\VMware Workstation\vmware.exe"
  TargetUserId S-1-0-0
  TargetUserName -

```

The diagram shows three Windows event logs represented as text blocks. Arrows point from the second event log to the third, and from the third back to the first, forming a cycle that represents an exploit chain.

Fig. 8. Windows event logs generated from a guest to host exploit

Exploit Chain Detector (ECD) Algorithm

Input: a list of ordered Windows event logs A; a list of process names to be monitored B
/* an event logs has the following attributes: NewProcessId, ProcessId, ProcessName, TargetDomainName */
/* B contains a list of process names that are executed after a vulnerability is exploited retrieved from report! [11] */
Output: a list of string stacks D, a Boolean represents if exploit chains are detected c
/* D will contain all exploit chains detected by the algorithm, and c is true if one chain is found*/
Initialization: create an empty event log a ; initialize c with the value false ; create integer m with initial value 0

```

1  for (i=0; i<Size(A); i++) do
2    if (A[i].ProcessId ∈ B) then
3      a=Ai
4      for (j=i; j<Size(A); j++) do
5        if (a.ProcessId == Aj.NewProcessId && a.TargetDomainName == Aj.TargetDomainName) then
6          Dn.Push(a.ProcessName)
7          a=Aj
8          if(A(j+m).NewProcessId==Null) then
9            c=true
10           m=m+1
11         end if
12       end if
13     end for
14   end if
15 end for

```

The ECD (Exploit Chain Detector) algorithm requires only two inputs for execution. First is the security monitoring logs on which detection is performed A, and second is the list of process names that need to be monitored B. A is directly retrieved from host which contains individual events with multiple fields like ProcessId, NewProcessId, ProcessName, TargetDomainName etc. B is the list of process names

given by JPCERT [11] that are executed after a vulnerability is exploited.¹ In output the algorithm returns whether an exploit chain is detected or not in a boolean variable c. If detected then it also shows the exploit chains in a stack D. For initializing the algorithm we need an empty event log a, an integer m with value 0 and c will be initialized with the value false.

When the algorithm starts processing it reads all the event logs available in A, then it starts checking one by one if the ProcessName of an event in A is present in B. If a match is found the single event of A is stored in a and ProcessName is pushed to the stack D. Now a second comparison is performed on those events of A which are present after a, in the comparison ProcessId of a and ComputerName of a is compared with the ProcessId of the next event of A and ComputerName of that event in the coming logs. If a match is found ProcessName is pushed to a stack D and value of a is updated with current value of event at A. This process is performed until there is no NewProcessID in A. When this happens true value is assigned to c while the stack D contains the whole exploit chain. We calculated the algorithm complexity and it was:

$$O(n \log n)$$

The algorithm complexity is good for detection of exploit chain in environment with small or medium amount of security logs data but in an environment with large amount of event log data the algorithm will take considerable amount of time for detection of exploit chains.

5 Implementation

We developed our proposed algorithm on a simple python based Windows logging mechanism. It is based on the standard pywin32 library presented at python library blog post [16], while the detection algorithm is built around this logging mechanism. The logs come in a recursive manner, as post exploitation is done after the initial exploitation with respect to time. We developed our detection algorithm POC on Microsoft Visual Studio 2017.² on python environment 3.6. Our implementation contains the following primary functions.

1. Get-All-System-Events

This function takes all event logs from system which include application events, security events, setup events, system events and forwarder events and write them to separate files on disk.

¹ The list is created according to the JPCERT report Detecting Lateral Movement through Tracking Event Logs, which suggest the following processes for active tracking: cmd, powershell, regsvr32, rundll32, mshta. https://www.jpcert.or.jp/english/pub/sr/20170612ac-ir_research_en.pdf

² <https://www.visualstudio.com/>

2. Event-Parser

Event log parser read the event from the disk and parse them to individual readable events and forward it to next function Get All Event Logs.

(1) Get-All-Event Logs

Get-All-Event-Logs is the core function of our detection algorithm. It takes parsed events from Event-parser, then performs event process comparison and event correlation for the detection of exploit chains.

To test the algorithm we run the developed tool on a Core i5 3320M 2.60 GHz system with 16 GB of RAM against 17,098 Windows security events and two executed exploits guest-to-host, macro-less DDE. The Execution took 7.3 s for the detection of the exploit chains, which can be seen in Fig. 9.



Fig. 9. Implemented algorithm process execution time and function calls

The implementation works without any malware, virus or malicious command signature for the detection of the exploit chain. We performed detailed experimentation on the developed algorithm to check the effeteness of our algorithm. The following section elaborates the experimental details and results.

6 Experimentation and Results

Two experiments were performed to check the effectiveness of the developed algorithm one is a guest-to-host exploit the other is a macro-less DDE exploit details of which are given below.

6.1 Guest-to-Host Exploit

1. Experimental Setup

We created our experimental setup on a 64-bit Windows 10 machine running on a VMware Workstation 12.5.5. A Guest Windows 10 operating system is installed on the

Vmware. For detection comparison analysis we installed Bit Defender Home, Avira Home, Kasper Sky Home, Avast Home and Panda Security Suite on the Host OS and deactivated them.

2. Controlled Exploit Execution

We executed a guest-to-host proof of concept for CVE-2017-4924 [17, 15] on Guest windows 10 operating system. The exploit breaks the Guest isolation and executed an CMD on host machine and then executed PowerShell. Execution of exploit on Process Hacker can be seen in the Fig. 10.

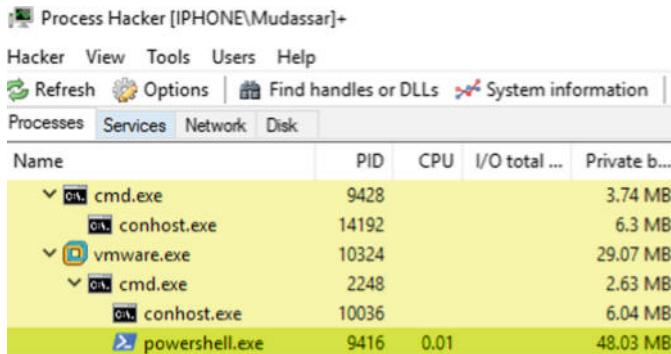


Fig. 10. Guest to host exploit execution

3. Scenarios

We created two scenarios for comparison of our developed algorithm with different anti-virus/anti-malware solution available in the market. In the first scenario we executed the exploit only against our detection algorithm. In the second scenario we executed the exploit against different anti-virus/anti-malware solution available in the market. Details of which is given below:

(a) Detection against developed algorithm

We executed our detection algorithm on the Windows 10 Host OS and executed the exploit on Windows 10 Guest OS and we were able to detect the exploit chain in the first run successfully.

Vmware → CMD → PowerShell

The detection of the exploit chain by the developed algorithm can be seen in Fig. 11. The exploit chain detected by our algorithm is according to the process execution tree shown at process hacker. However, due to the event correlation capabilities of the developed algorithm with respect to malicious process monitoring, we are able to mark the chain as being malicious.



```
C:\Program Files (x86)\Microsoft Visual Studio\Shared\Python36_64\python.exe
Logging Security events
Total events in Security = 17727
Malicious Exploit Process Launch On Host:iphone, Time:2018-05-27 13:14:26
-->Exploit Chain:vmware.exe; cmd.exe; powershell.exe;
---->Exploit Injection Tracer:
```

Fig. 11. Guest-to-host exploit detection**(b) Detection against anti-virus/anti-malware solutions**

We ran Bit Defender Home, Avira Home, Kasper Sky Home, Avast Home and Panda Security Suite on the Windows 10 Host OS one by one while executing the guest-to-host exploit on a Window 10 Guest OS for the possible detection of exploit chain we weren't able to identify any malicious activity.

4. Experimental Results

We ran a comparative analysis of our detection techniques with different anti-virus and anti-malware solution available in the market. Table 1 shows the result of detection by different security software.

Table 1. Result of comparative detection analysis of developed algorithm and different software security software

Solution	Detection Yes/No
Proposed algorithm	Yes
Windows defender	No
Bit defender home	No
Avira home	No
Kasper sky home	No
Avast home	No
Panda security suite	No

6.2 Macro-less DDE Exploit

1. Experimental Setup

We used Microsoft Office 2013 running on 64-bit Window 10 for the experimentation purpose.

2. Controlled Exploit Execution

For macro-less DDE Exploit we developed an obfuscated DDE Exploit for CVE-2017-11826. The exploit first executes CMD from MS Word then from CMD it executes PowerShell for further exploitation. The exploit execution on Process Hacker can be seen in Fig. 12.

Processes	Services	Network	Disk	
Name		PID	CPU	I/O
explorer.exe		3116	0.62	
WINWORD.EXE		10024	0.34	
cmd.exe		5404		
conhost.exe		5220		
powershell.exe		8516		
powershell.		9740	0.01	

Fig. 12. Macro-less DDE exploit execution

3. Scenarios

We created two scenarios for the evaluation of our developed algorithm in the first scenario we executed the exploit on the Experimental setup to check the detection against our developed algorithm. In the second scenario we used online service Vlrustototal³ which performed detection analysis against 59 anti-virus/anti-malware solution details of the scenarios is given below:

(a) Detection against developed algorithm

We executed our detection algorithm on the Windows 10 running MS Word and we were able to detect the exploit chain in the first run successfully.

Word → CMD → PowerShell

The detection of the exploit chain by the developed algorithm can be seen in Fig. 13.

```
Malicious Exploit Process Launch On Host:iphone, Time:2018-06-06 18:52:48
-->Exploit Chain:excel.exe; cmd.exe; calc.exe;
-->Exploit Injection Tracer:
```

Fig. 13. Macro-less DDE exploit detection

The exploit chain detected by our algorithm is according to the process execution tree shown at Process Hacker. However, due to the event correlation capabilities of the developed algorithm with respect to malicious process monitoring, we are able to mark the chain as being malicious.

³ <https://www.virustotal.com/#/file/27c058180a47a5f73ac013e908dde0ec823a28a561408749872e54e6944a4c3f/detection>.

(b) Detection against anti-virus/anti-malware solutions

As stated earlier we developed an obfuscated macro-less DDE exploit which have zero detection signature against 59 anti-virus and anti-malware solution on Virus Total⁴ as seen in the Fig. 14.

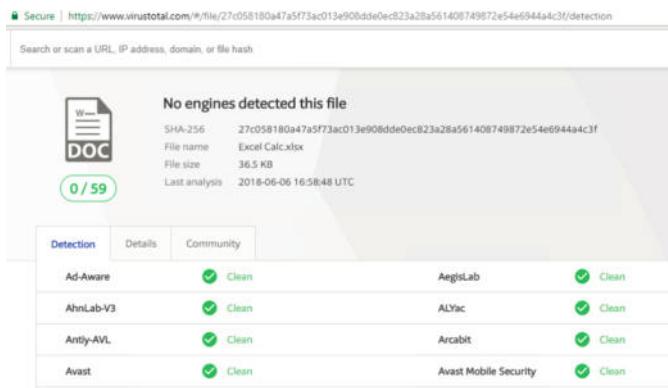


Fig. 14. Obfuscated macro-less DDE exploit

4. Experimental Results

Our analysis is being performed on 59 anti-virus and anti-malware solution for saving space few results are omitted but details of analysis can be found online (See footnote 2,3). Table 2 presenting the detection result compare to different anti-virus and anti-malware solution.

Table 2. Result of comparative detection analysis of developed algorithm and different software security software

Solution	Detection Yes/No
Proposed algorithm	Yes
Ad-Aware	No
AegisLab	No
AhnLab-V3	No
ALYac	No
Antiy-AVL	No
Arcabit	No
Avast	No
Avast mobile security	No
AVG	No

⁴ <https://www.virustotal.com/#/file/27c058180a47a5f73ac013e908dde0ec823a28a561408749872e54e6944a4c3f/detection>.

7 Discussion

The key factor of failure of other detection techniques compare to our techniques is that other detection techniques focus on signature and illegitimate behavior of processes that are being executed. As shown above legitimate behavior of an application can be used for malicious purposes. Similarly signatures of malicious code can be obfuscated as well to avoid detection. Our detection technique works completely different in comparison to other techniques it tries to identify the chain of processes that are being executed by a process and then co-relate them for the identification of malicious processes in the chain. Therefore it has the capability to detect those exploits which are not detected by other available solutions.

We believe that the algorithm complexity is not ideal and there is a lot of room for improvement. But the approach which the algorithm use is quite unique for detection of malicious exploit chains. We intend to further refine the technique for other detection like the detection malicious activates user by means of event correlation.

8 Conclusion and Future Work

With the proposed detection technique, we are able to identify complex exploit chains. We assume that some complex user administration automation scripts may cause false positives due to their complex execution nature, but overall the detection technique is satisfactory in detecting complex exploit chains. A significant benefit of this technique is that it works completely blindly, without any signature and behavior metrics. In the future, we intent to further refine our technique to trace the exploit chain when the process migrates to another process. Furthermore, we will perform our experiments in a large network for identification of false positives in our detection algorithm.

References

1. Pwn2own 2018: Day two results and master of Pwn. <https://www.zerodayinitiative.com/blog/2018/3/15/pwn2own-2018-day-two-results-and-master-of-pwn>. Accessed 17 May 2018
2. Srinivasan, D., Wang, Z., Jiang, X., Xu, D.: Process out-grafting: an efficient out-of-vm approach for fine-grained process execution monitoring. In: Proceedings of the 18th ACM Conference on Computer and Communications Security, pp. 363–374. ACM (2011)
3. Mandal, D., Zhang, Y.: The Great Escapes of VMware: A Retrospective Case Study of VMware Guest-to-Host Escape Vulnerabilities. Blackhat, London (2017)
4. Sensepost. <https://sensepost.com/blog/2017/macro-less-code-exec-in-msword/>. Accessed 17 May 2018
5. Neumann, W.C., Corby, T.E., Epps, G.A.: System for secure computing using defense-in-depth architecture. U.S. Patent 7,428,754. Issued 23 Sept 2008
6. Win, T.Y., Tianfield, H., Mair, Q.: Big data based security analytics for protecting virtualized infrastructures in cloud computing. IEEE Trans. Big Data **4**(1), 11–25 (2018)
7. Wang, X., Qi, Y., Wang, Z., Chen, Y., Zhou, Y.: Design and implementation of SecPod, a framework for virtualization-based security systems. IEEE Trans. Dependable Secure Comput. (2017)

8. Ucci, D., Aniello, L., Baldoni, R.: Survey on the usage of machine learning techniques for malware analysis. arXiv preprint [arXiv:1710.08189](https://arxiv.org/abs/1710.08189) (2017)
9. Hendlar, D., Kels, S., Rubin, A.: Detecting malicious PowerShell commands using deep neural networks. arXiv preprint [arXiv:1804.04177](https://arxiv.org/abs/1804.04177) (2018)
10. Dosfuscation: Exploring the depths of Cmd.exe obfuscation and detection techniques
11. Research Report Released: Detecting lateral movement through tracking event logs (version 2). <https://blog.jpcert.or.jp/2017/12/research-report-released-detecting-lateral-movement-through-tracking-event-logs-version-2.htm>. Accessed 17 May 2018
12. Bohannon, D.: <https://www.fireeye.com/blog/threat-research/2018/03/dosfuscation-exploring-obfuscation-and-detection-techniques.html>. Accessed 19 May 2018
13. Server Virtualization and Os Trends. Spiceworks, Inc. <https://community.spiceworks.com/networking/articles/2462-server-virtualization-and-os-trends>. Accessed 24 May 2018
14. Virtual Machine Escape. https://en.wikipedia.org/wiki/Virtual_machine_escape. Accessed 17 May 2018
15. patch Blog Luka Treiber. <http://blog.0patch.com/2017/10/micropatching-hypervisor-with-running.html>. Accessed 18 May 2018
16. 4688(s): A new process has been created. (windows 10). Mir0sh. <https://docs.microsoft.com/en-us/windows/security/threat-protection/auditing/event-4688>. Accessed 19 May 2018. URL: <https://www.blog.pythonlibrary.org/2010/07/27/pywin32-getting-windows-event-logs/>. Website Title: The Mouse Vs The Python. Date Accessed 27 May 2018
17. Comsecuris/vgpu_shader_pocs Comsecuris. https://github.com/Comsecuris/vgpu_shader_pocs. Accessed 18 May 2018



Analysing Security Threats for Cyber-Physical Systems

Shafiq ur Rehman^(✉), Manuel De Ceglia, and Volker Gruhn

Institute of Software Technology, University of Duisburg-Essen,
45127 Essen, Germany

shafiq.rehman@paluno.uni-due.de

Abstract. Cyber-physical systems have established as an essential part of the modern world and will grow in amount and complexity in the future. In the beginning, this paper provides an overview of CPS architecture, where we describe basic components at three different OSI reference model layers that includes: the human machine interface (application layer), supervisory control and data acquisition (network layer) and programmable logic controllers (physical layer). Since the physical layer enables direct contact to human beings, security is an important factor in the development process. Nonetheless, cyber-physical systems become more and more complicated and offer a wide surface of vulnerabilities which can be exploited through external threats. Since attacks can be launched at every single layer, a wide area of different threats can be inherited and need to be avoided by appropriate security measures. The main contribution of this paper is to provide a security threat tool, where we determine threats and vulnerabilities in cyber-physical systems at the application, the network and the physical layer. Furthermore, the tool is able to suggest solutions which can prevent attacks against those identified threats.

Keywords: Security · Threat · Vulnerability · Asset · PLC · SCADA · HMI · Architecture · Cyber-physical systems (CPS)

1 Introduction

Cyber-Physical Systems (CPS) have been growing in recent years and will affect the modern world increasingly in the future. Cyber-physical systems can be defined through connecting the global technology (cyber) and the physical environment. Essentially, they can be summed up by gathering information from the physical environment through sensors or changing the real environment through actuators [1–3]. Furthermore, they need to run daily and almost autonomously. Their connection of the cyber- and the physical world provides many advantages and makes them attractive for direct interactions with human. Since human beings cannot be harmed and requirements like continuous or autonomous running are essential for CPS. Therefore, security becomes an even more important factor in the development process.

Security can be defined as the approach to prevent a system against any malicious threats, which try to exploit vulnerabilities in the system [4]. These threats can be found in different layers of abstractness and in several forms of attacks. This paper aims at

illuminating the reader about the widely-spread area in which threats and vulnerabilities appear in CPS that explains their effect in the application layer, the network layer and the physical layer. The application layer is the highest one and supports the user with a graphical user interfaces. Some popular threats like adware, spyware or phishing is defined on our proposed CPS tool. The network layer on the other hand is more abstract and transparent for the user. It prepares information to be sent throughout the network and guarantees that they will be guided to the correct receiver. Man-in-the middle- or denial-of-service (DoS) are probably the best-known threats at this layer. The physical layer, which is exclusive for CPS and cannot be found in common software, sums up all physical components (sensors and actuators). Natural disasters and physical damage through intruders are basic threats at this layer. Being aware of the huge amount of threats, this paper also provides solution approaches at all three layers, which should be handled in every CPS environment.

2 Related Work

Since different references will be quoted in this article, the most important ones are summarized in this following section:

For receiving a first overview about security in cyber-physical systems, this paper [5] is a good choice. It explains four basic terms concerning security, starting with security itself, software threat, software vulnerability and attack possibilities. Four commonly known examples of CPS: industrial control systems, smart grid systems, medical devices and smart cars are analyzed.

This paper [6] explains the general principles of reference model for layering and clarifying the basic idea and sense of abstraction layers of architecture and pointing out advantages of them. Later, the general function of these layers is mentioned and the way they are interconnected and how these interconnections work is clarified by some examples like multiplexing and splitting. In the end, each of the seven layers of the model is defined and separated from each other.

The authors [7] introduces the topic by separating computer threats, attacks and assets and gives an appropriate overview about security requirements for computer systems. Next, the authors navigate the reader throughout the most common computer security principles such as cryptography or authentication that explains malicious software in general, different types of them and countermeasures to defend against them. Next to security lacks in software like buffer overflows, security of human resources and physical security has been explained. At the end, the basic concepts of cryptographic algorithms and network security are briefly described.

The author [8] provides the reader with general background information about the Open System Interconnect (OSI) reference architecture, offers key term definitions like active and passive attacks and explains the four basic security goals authenticity, availability, confidentiality and integrity. After that, the main content of cryptography gets focused and symmetric ciphers has been clarified. After a short introduction about the two classical encryption techniques substitution and transposition, examples and implementations of two highly spread encryption algorithms: data encryption standard (DES) and advanced encryption standard (AES) are presented. Next to symmetric

ciphers, asymmetric ciphers, like the most popular one called “Rivest-Shamir-Adleman” (RSA) is explained. Since asymmetric encryption always troubles with a safe key exchange between sender and receiver, solutions like Diffie-Hellman key exchange are mentioned. A couple of algorithms aiming at integrity get introduced and a bit more details about network security in general and especially about network access control, transport-level security and wireless network security are presented.

The paper [9] deals with three different types of security measures, starting with firewalls. A brief explanation of firewalls is specified and three of the most usual types, packet filtering firewall, circuit level gateways and application gateways are discussed. Furthermore, intrusion detection systems are illustrated, in which the main focus on network based and host based intrusion detection systems. Anti-virus scanners are explained afterwards. Since firewalls, intrusion detection systems and anti-virus scanners seem to do similar jobs, the differences are clarified at the end of this paper.

The third-party authentication service Kerberos seems to be very complex. The paper [10] presents the function of Kerberos in non-technical language. After an initial overview, the software components are illustrated and the way it works is explained in detail, using different images for better understanding. At the end, some disadvantages and problems of the authentication service has been discussed.

The related work mainly focused on security of cyber-physical systems, where we can see the importance of security of CPS. Generally, all researchers agreed that security is a major concerns and need to focus on this area. Therefore, it is important to develop new techniques, methods and tools to handle security of cyber-physical systems.

3 Security Challenges in Cyber-Physical Systems

Software security deals with protecting a system against any malicious actions from outsiders and ensures that it runs continuously and flawlessly at any point in time [11]. Preserving the four basic security concepts confidentiality, integrity, authenticity and availability, the system assets are the main target. Confidentiality guarantees that information access is explicitly available for authorized users, integrity protects against improper modification of data, authenticity assurances that users interacting with the system are genuine and availability provides that resources are timely and reliably accessible [12]. Due to their rising complexity modern computer systems grow in their amount of failures and complete security is almost impossible to achieve. Nonetheless, the more security measures are provided [13].

3.1 Security Threat

A software threat is a potential for security abuse, which takes advantage of design failures in the software system and is the intention of what software security is trying to protect [14]. It often arises in form of malicious software (malware). Source, target system, a motive for the attacker, an attack vector and a consequence for both, the attacker and the victim are five special characteristics for every threat. To guarantee the four basic security targets confidentiality, integrity, authenticity and availability, a wide

bunch of security measures should be provided in cyber-physical system. Even though the occurrence of software threats is potential, an attention is always better than charity [15].

3.2 Security Vulnerability

A software vulnerability is a weakness in the system, which can be exploited by software threats [16]. Many reasons for software vulnerabilities can be inherited, that is why a system never free of them. It consists out of four phases, as shown in Fig. 1.

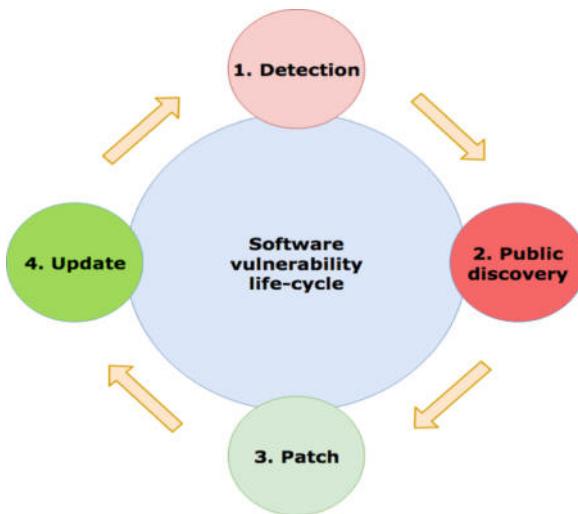


Fig. 1. Life-cycle of a software vulnerability

The initial detection of the vulnerability is the first phase. If an attacker figures out the failure then attacker has a massive advantage in contrast to the owners of the software system, since they are not aware of that weakness. Once the vulnerability is known, the second phase starts. No countermeasures are available at this point, that is why phase two is the most dangerous one in the life-cycle. When the vendor is finally able to release a patch that fixes the vulnerability, a phase three has begun and the weakness is almost eliminated. Nonetheless, the life-cycle only ends, if every user has finally installed the patch and updated the system, which could never be the case since they are generally not aware of the importance of security updates.

3.3 Importance of Cyber-Physical System Security

Since cyber-physical systems offer a new dimension of attacks in the physical layer, classic security approaches that also cover the application and the network layer. Furthermore, CPS are typically mixed up of different third-party systems, which are

often legacy systems and provide no more updates or security patches as shown in Fig. 2. Not only the fact that out-dated software is used as a lack of security, but the heterogeneity itself, too [17].

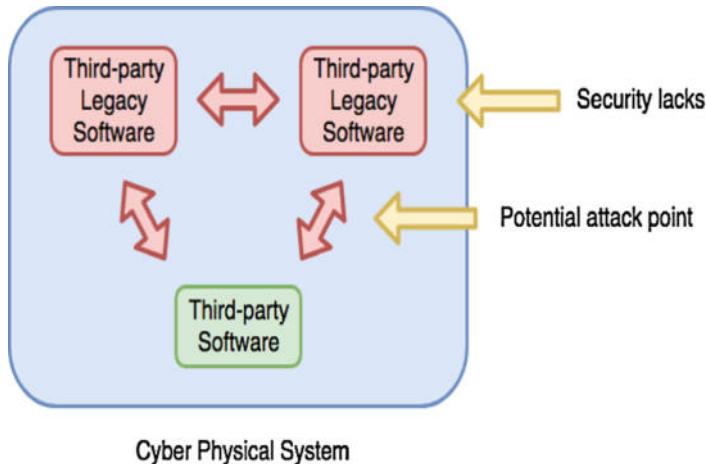


Fig. 2. Security lacks of CPS heterogeneity

Hence these critical lacks of security are rather obvious, the assumption that security measures in CPS are strictly implemented should be plausible [18]. In case, the opposite is the truth. Engineers of those systems arrogantly assumed that the technology is too complex and special for any attacker to intrude into them, that is why security has been dramatically neglected in recent years and attacks like Stuxnet [19, 20] or the Maroochy Water Breach [21] could be performed rather easily.

4 Proposed Architecture of Cyber-Physical Systems

Cyber-physical systems typically consist of three different parts. The Human Machine Interface (HMI) enables users to control the system, Programmable Logic Controllers (PLCs) change the physical environment or collect information from it and Supervisory Control and Data Acquisition (SCADA) transmits information between the HMI and the PLCs as shown in Fig. 3.

4.1 Application Layer

The application layer in a software system is the highest one. It directly serves the user by providing the distributed information in an appropriate way (applications) and abstracts from everything the user should not know (e.g. communication) [6]. The application layer of CPS comprises the human machine interface (HMI). It enables the user to interact with the system and to change the state of actuators in real-time. It

usually offers easy to understand graphical user interfaces with drag and drop, since cyber-physical systems aim at being self-learned. On the other hand, it displays information gathered from Programmable Logic Controllers (PLC) in real-time, which is the main target of the HMI [22]. Since CPS should work autonomously, human interactions are basically not wished. Nonetheless, in cases of security emergencies, where human decision making is more important in contrast to computer logic, employees should be able to interact with the system at any point in time.

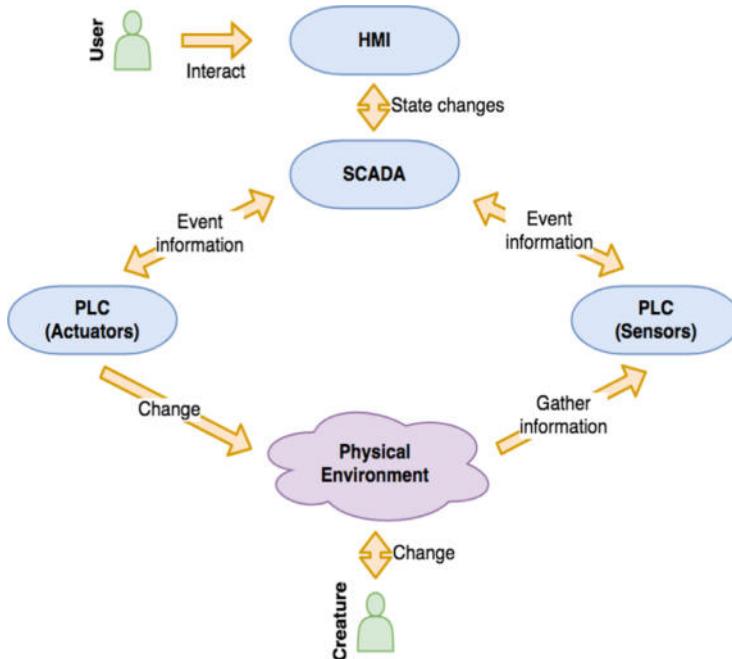


Fig. 3. Software architecture CPS

4.2 Network Layer

Since information should be transmitted in a network, the network layer provides a basis for that. Every node has a specific, unique logical address. If data should be sent from a source to a destination node, the network layer allocates each packet the specific destination address and manages the path determination through the network [23]. Supervisory control and data acquisition (SCADA) can be allocated to the network layer. Since CPS can be widely spread over large areas, centralized control is a main requirement. SCADA transmits information from PLCs to the HMI and the other way around and enables a solution for that constraint. Next to the main characteristic, SCADA is also able to acquire data. Since PLCs create lots of information, SCADA can log these down and make them persistent for any future use like maintenance or security updates.

4.3 Physical Layer

The physical layer of CPS comprises physical components in the real environment. These systems directly interact with the nature, PLCs are one of the most important parts of them and that point in which CPS differ from common software. Programmable logic controllers can be split into two categories. Actuators change the physical environment, like sprinklers in an irrigation system. They translate logical data from computers into physical commands that can be executed.

5 Analyzing Threat Tool for Cyber-Physical Systems

The analyzing threat tool offers an opportunity to create custom hazard model for CPS. Each model consists of lists of different threats, vulnerabilities and assets at the application, network and physical layer. Figure 4 shows the main interface of the program that displays the lists of application, network and physical layer threats. Furthermore, the lists of vulnerabilities and assets can also be shown, when the related buttons are clicked. A description for each item and a solution for each threat is displayed under the lists.



Fig. 4. Main interface of the program

This view also enables to create new custom lists, to delete, edit or show these custom lists and to manipulate the previously defined lists of threats, vulnerabilities and assets. If a custom list shall be created, a new view will open which is shown in Fig. 5. Security elements from the standard lists can be drag and dropped into the fields of the custom lists. If an item should be removed, it can be drag and dropped to the trash image. Once a name was entered and at least one threat, vulnerability or asset has been added, the list can be created by clicking the apply button.

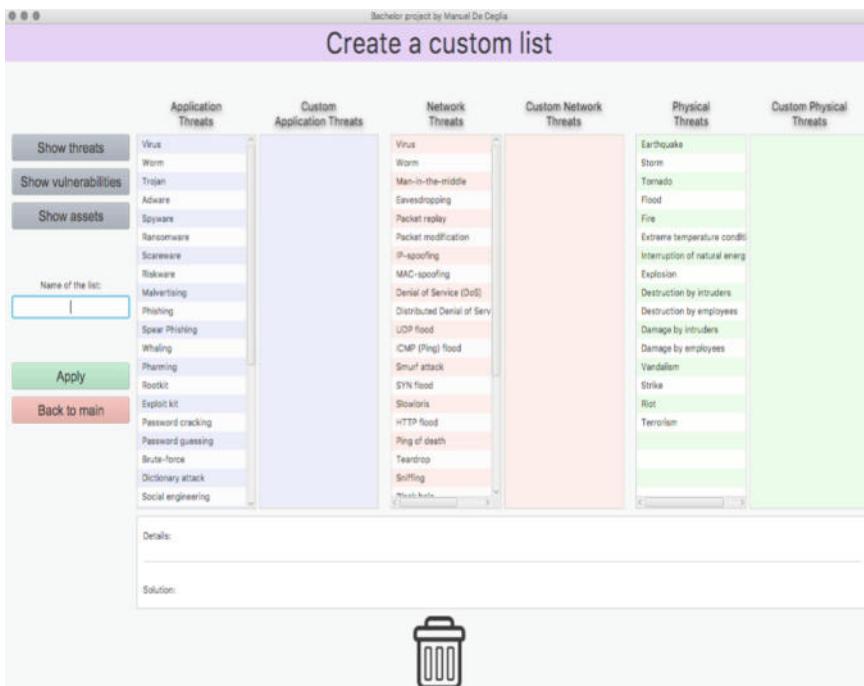


Fig. 5. View to create a custom list

If a custom list needs to be edited or deleted, a combo box pops up and the specific list should be chosen. If the list is edited, the same view as “Create a custom list” opens, but instead of empty custom lists, the threats, vulnerabilities and assets of the chosen list will be shown. If the list is deleted, the user gets directed back to the main view and the list will be removed. Since the amount of threats, vulnerabilities and assets is almost endless, the opportunity of adding or deleting item is also possible as shown in Fig. 6.

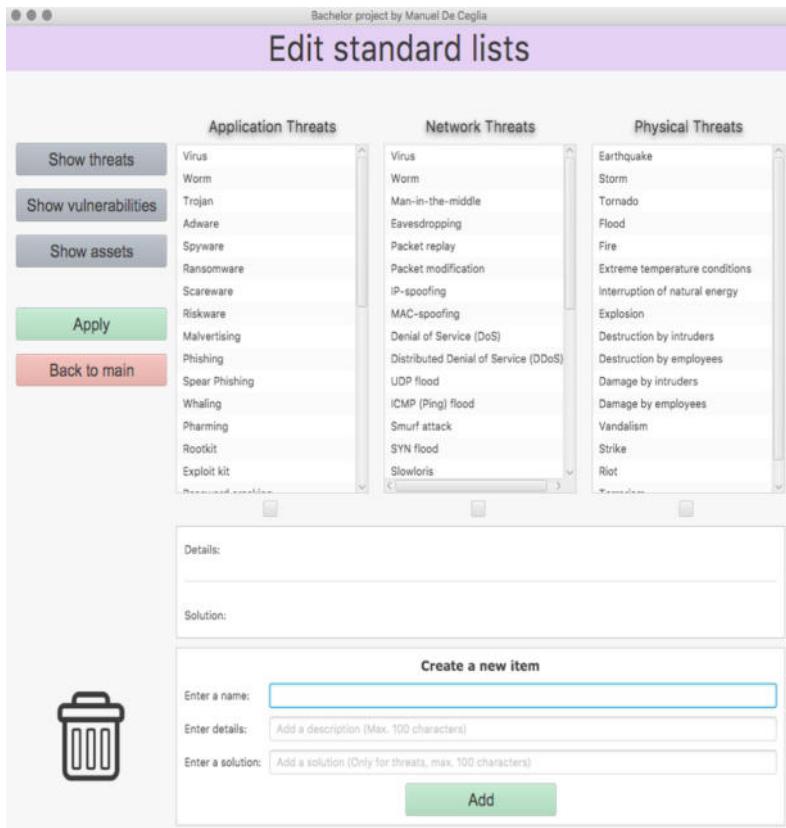


Fig. 6. View to edit the standard list items

Again, if any security element is deleted, a simple drag and drop to the trash image will remove it. If a new element is added, a name and a description should be entered. If creating a new threat, a solution should also be provided.

6 Conclusion

Security threats is always being potential problems, since software is never free of vulnerabilities. Cyber-physical systems offer a wide, eventually new area of possible vulnerabilities that can be exploited by threats. This paper has proposed the secure architecture and threat tool to identify the security threats of CPS. The tool provides the basic lists of application, network and physical threats and vulnerabilities, which every CPS should be encountered before the system is deployed. Since threats do not only appear at a single layer, a full awareness of the system and defenses in each layer should be provided. Security measures like firewalls can be implemented on several layers. Encryption is a basic requirement for integrity and confidentiality at the network

layer which guarantees continuous availability of physical components. This paper aimed at convincing the reader that threats against CPS are even more present than against classical software.

Acknowledgements. This work has been supported by the European Community through project CPS.HUB NRW, EFRE Nr. 0-4000-17.

References

1. Rehman, S., Gruhn, V.: Security requirements engineering (SRE) framework for cyber-physical systems (CPS): SRE for CPS. In: New Trends in Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 16th International Conference SoMeT_17, vol 297, p. 153. IOS Press, Sept 2017
2. Rehman, S., Gruhn, V.: Recommended architecture for car parking management system based on cyber-physical system. In: 2017 International Conference on Engineering & MIS (ICEMIS), pp. 1–6. IEEE, May 2017
3. Rehman, S., Gruhn, V.: An effective security requirements engineering framework for cyber-physical systems. *Int. J. Inf. Commun. Technol. (Special Issue Cyber-Phys. Syst. Data Process. Commun. Architect.)* **6**(3). ISSN: 2227-7080; ESCI-WoS index (2018)
4. Todd, M., Koster, S.R., Wong, P.C.M.: Hewlett-Packard Enterprise Development LP, 2016. System and method for securing a network from zero-day vulnerability exploits. U.S. Patent 9,264,441
5. Humayed, A., Lin, J., Li, F., Luo, B.: Cyber-physical systems security—a survey. *IEEE Internet Things J.* 1–1 (2017)
6. Zimmermann, J.D.: OS1 reference model—the ISO model of architecture for open systems interconnection
7. Stallings, W., Brown, L.: Computer Security: Principles and Practice, 3rd edn. Pearson, Boston (2015)
8. Stallings, W.: Cryptography and Network Security: Principles and Practice, 7th edn. Pearson, Boston (2017)
9. Greensmith, J., Aickelin, U.: Firewalls, intrusion detection systems and anti-virus scanners (2005)
10. Steiner, J.G., Neuman, B.C., Schiller, J.I.: Kerberos: an authentication service for open network systems. In: USENIX Winter (1988)
11. Omar, S.: Information system security threats and vulnerabilities: evaluating the human factor in data protection. Doctoral dissertation (2017)
12. Kalloniatis, C., Mouratidis, H., Vassilis, M., Islam, S., Gritzalis, S., Kavakli, E.: Towards the design of secure and privacy-oriented information systems in the cloud: identifying the major concepts. *Computer Stand. Interfaces* **36**(4), 759–775 (2014)
13. Wells, L.J., Camelio, J.A., Williams, C.B., White, J.: Cyber-physical security challenges in manufacturing systems. *Manuf. Lett.* **2**(2), 74–77 (2014)
14. Wang, L., Törngren, M., Onori, M.: Current status and advancement of cyber-physical systems in manufacturing. *J. Manuf. Syst.* **37**, 517–527 (2015)
15. Abomhara, M.: Cyber security and the internet of things: vulnerabilities, threats, intruders and attacks. *J. Cyber Secur. Mobil.* **4**(1), 65–88 (2015)
16. Papp, D., Ma, Z., Buttyan, L.: Embedded systems security: threats, vulnerabilities, and attack taxonomy. In: 13th Annual Conference on Privacy, Security and Trust (PST), pp. 145–152. IEEE, July 2015

17. Kornecki, A.J., Subramanian, N., Zalewski, J.: Studying interrelationships of safety and security for software assurance in cyber-physical systems: approach based on bayesian belief networks. In 2013 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 1393–1399. IEEE, Sept 2013
18. Mo, Y., Kim, T.H.J., Brancik, K., Dickinson, D., Lee, H., Perrig, A., Sinopoli, B.: Cyber-physical security of a smart grid infrastructure. Proc. IEEE **100**(1), 195–209 (2012)
19. Kushner, D.: The real story of Stuxnet. IEEE spectrum: technology, engineering, and science news, 26-Feb-2013. Available: <https://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet>. Accessed 15 Nov 2017
20. Karnouskos, S.: Stuxnet worm impact on industrial cyber-physical system security. In IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society, pp. 4490–4494. IEEE, Nov 2011
21. Slay, J., Miller, M.: Lessons learned from the Maroochy water breach. Crit. Infrastruct. Prot. 73–82 (2007)
22. Zhu, B., Joseph, A., Sastry, S.: A taxonomy of cyber-attacks on SCADA systems. In: Internet of things (iThings/CPSCom), 2011 International Conference on and 4th International Conference on Cyber, Physical and Social Computing, pp. 380–388. IEEE, Oct 2011
23. Zanella, A., Bui, N., Castellani, A., Vangelista, L., Zorzi, M.: Internet of things for smart cities. IEEE Internet of Things J **1**(1), 22–32 (2014)



Bayesian Signaling Game Based Efficient Security Model for MANETs

Rashidah Funke Olanrewaju¹, Burhan ul Islam Khan¹⁽⁾,
Farhat Anwar¹, Roohie Naaz Mir², Mashkuri Yaacob¹,
and Tehseen Mehraj³

¹ Department of ECE, Kulliyyah of Engineering, UIAM,
Gombak, Malaysia
burhan.ium@gmail.com

² Department of CSE, National Institute of Technology,
Srinagar, Kashmir, India

³ Department of ECE, Islamic University of Science and Technology,
Pulwama, Kashmir, India

Abstract. Game Theory acts as a suitable tool offering promising solutions to security-related concerns in Mobile Ad Hoc Networks (i.e., MANETs). In MANETs, security forms a prominent concern as it includes nodes which are usually portable and require significant coordination between them. Further, the absence of physical organisation makes such networks susceptible to security breaches, hindering secure routing and execution among nodes. Game Theory approach has been manipulated in the current study to achieve an analytical view while addressing the security concerns in MANETs. This paper offers a Bayesian-Signaling game model capable of analysing the behaviour associated with regular as well as malicious nodes. In the proposed model, the utility of normal nodes has been increased while reducing the utility linked to malicious nodes. Moreover, the system employs a reputation system capable of stimulating best cooperation between the nodes. The regular nodes record incessantly to examine their corresponding nodes' behaviours by using the belief system of Bayes-rules. On its comparison with existing schemes, it was revealed that the presented algorithm provides better identification of malicious nodes and attacks while delivering improved throughput and reduced false positive rate.

Keywords: Bayesian signaling model · Bayesian-Equilibrium game theory · MANETs · Secure routing protocol

1 Introduction

The use of wireless cellular systems dates back to the 1970s [1]. Since then, these have been evolving from the first, second, third, fourth to fifth generation wireless systems [2]. The wireless networks require a centralised supporting arrangement like that of an access point for their operation. Wireless users can remain connected to the wireless systems while roaming using those access points. However, these fixed supporting structures restrict the wireless systems' adaptability, i.e., this technology cannot be employed in regions with no infrastructure in place. The wireless systems of future

generations shall need quick and easy deployment of these networks which is not feasible with the standard framework of wireless systems [3–5].

As a result of the topical advancements like the introduction of Bluetooth, new wireless systems referred to as MANETs came into existence [6]. A mobile ad hoc network, also known as short-lived network works devoid of fixed infrastructures. The word ‘ad hoc’ is derived from the Latin meaning ‘for this or only for this’ [7]. MANETs are autonomous systems formed of mobile nodes interconnected via wireless channels with every node functioning as an end system as well as a router for every other node in that network [8, 9]. The mobile nodes in a MANET establish a temporary network dynamically with no centralised administration or fixed infrastructure. With the evolution of wireless networks, the ad hoc potentials are anticipated to grow in significance. Besides, the technological solutions for supporting more critical, crucial research and development in the future can be envisaged in academy and industry [5].

Nowadays, the critical, challenging issue in the research area of mobile wireless networks (i.e., MANETs) is security [10–12]. This is because the topology of a mobile ad hoc network changes dynamically. Many distributed algorithms have been put forward by researchers for managing link scheduling, network routing, and network organisation. Further, the unique characteristics of wireless networks add to security-related challenges [13].

Before underlining the prime security concerns, the common challenges existing due to the inherent nature of MANETs are listed as under [14]:

- i. Insecure wireless environment
- ii. Absence of central points
- iii. Constrained resources
- iv. Node mobility
- v. Scalability
- vi. Nature of unpredictability

Unfortunately, the low accuracy in identification of intruders in MANETs resulting due to its decentralised topology, enables intruders to perform malicious assaults on the network. Further, the growing popularity of MANETs in urban-Adhoc IoTs owing to their brisk network construction and lightweight setup, results in an urgent demand for a solution [15, 16].

Numerous assaults have been witnessed at each layer within MANETs, i.e., at the physical, data link or MAC, network, transport and application layers. Disruption of trust and reputation among nodes exists in such networks as they emerge to be vulnerable to numerous security attacks [17, 18]. The security breach in such systems mostly affect the reliable nodes hindering communication and resulting in data loss, hence reducing overall application performance. The cost associated with such breaches is intensified when considering their application in massive scale urban-Adhoc IoTs.

Further, from the review conducted, it was revealed that maximum schemes offering a solution in MANETs were based on routing protocols than dedicated security-based approaches. Hence, there exists a lack of feasible security solution in MANETs, as routing protocol based schemes overlook malicious behavior while at the same time add to network overhead. The practical implementation of routing protocol

based schemes is quite complicated, when considering node misbehavior respective strategies to deal with such behaviour. Even though a number of issues such as power, routing and QoS concerns [8, 19–22] have been identified over the past few years, security issues remain unresolved and unattended [23–25]. Although a significant amount of research has been done, a lack of efficient and feasible system ensuring proper and fail-proof security while safeguarding imminent security protocols normalization exists. In other words, network security has emerged as a prominent issue among researchers to protect against numerous adversary assaults.

The conventional protocols available for ensuring security are inappropriate for networks like MANETs mainly because of the following two reasons: (i) Wireless attacks can be launched from any direction in a wireless network, (ii) the decentralised authority and infrastructure-less characteristic of MANETs. Some of the most widely occurring attacks in MANETs comprise black-hole, Denial-of-Service (DoS), worm-hole, interference, resource consumption, etc. [26].

As per the researchers, there exist two complementary approaches to protect MANETs: (i) Prevention-based strategy (such as authentication) (ii) Detection-based strategy (such as IDS-Intrusion Detection System). Recently, Game Theory model has been presented to enhance the level of network security.

While studying some more security mechanisms in MANETs, it was observed that game theory has an unprecedented contribution in the recent past due to the accuracy of its computational efficiency and probabilistic approach. Game theory can be defined as a mathematical model that analyses interactive decisions in a particular situation that can be called a game. There are two types of game models, cooperative, and non-cooperative [27]. The former is used when the players are bonded to a specified agreement called a binding agreement, and the players will act under this agreement. In contrast, the non-cooperative model is applied when there is no binding agreement between the players; this enables the players to change their strategies at any given time. The entities following a non-cooperative model can be called self-enforcing entities. Furthermore, the applications of game theory extend beyond the realm of computers and networks as it has a central place in economic theories.

A game comprises of a collection of players, payoffs, along with certain strategies/moves. For every possible state in a game, there exists a plan of action to be undertaken by the player. During a specific state in a game, the player wins or loses as per the incentive scheme based on players' payoffs. The nodes can travel randomly since they are portable. The payoff scheme supports the players for forwarding the packets among themselves. While dealing with resource consumption, inefficient security is being offered. So, to overcome this issue, MANETs are partitioned into distinct clusters. Within each cluster, a node is chosen as a cluster head/head node which acts as an IDS for the entire cluster. A reputation model holds the responsibility of keeping the single nodes motivated. Among other game theory models, the endogenous interaction model is being offered by 'Bayesian interaction' games, defining asymptotic statistics for cooperation among players. Numerous critical issues where there occurs delay between player reaction and private information can be solved by such approach.

'Bayesian Signaling' model has been manipulated in the current study to elucidate the security concerns in MANETs. The prime concern is the identification of malicious

behaviours and activities. Further, a secure routing protocol can be designed for MANETs by utilising threshold achieved by using this approach. Moreover, the proposed scheme employs the reputation system stimulating enhanced collaboration between nodes while restricting the utility of malicious nodes.

The paper is organised in 5 sections. Section 2 is dedicated to the review of prior work conducted in the field of MANETs. Section 3 highlights the framework design of the proposed model followed by a discussion of results in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Related Work

From the review conducted, an extended amount of research was found to offer solutions based on game theory for numerous problems in MANETs like dynamic spectrum sharing, topology control, etc. The work of the researchers followed by their contributions have been summarised below:

In [28], several solutions have been presented by the authors to address the issue of intrusion in MANETs, with some of the offered solutions based on game theory. Four approaches based on game theory have been assessed in the research conducted, and manipulation of energy efficiencies have claimed the higher performance of IDS, the number of nodes with IDS abilities, clustering and identification of selfish misbehaving nodes in the form of performance metrics.

The authors in [29] have presented a unique cooperation system based on credits enabling the system to safeguard against cheating done by malicious nodes applying hash chains on messages. In the study, a lower workload is being experienced by nodes in comparison to other methods utilising digital signatures. Further, any cooperation level can be attained by the nodes as established by the game theoretic analysis, provided the mechanism is making the appropriate payments. However, the scalability of the system and coordination between malicious nodes has not been deliberated neither such nodes are considered fragile. Additionally, there exists a susceptibility of hash chains to rainbow attacks, and hence, there exists a scope for enhancement in the proposed strategies.

The authors in [30] have presented a security add-on namely “AODV-GT” founded on game theory for reactive protocol Ad hoc On-Demand Distance Vector-AODV. The prime concern of add-on was the protection of MANETs against black hole attack. On the integration of the proposed system founded on non-cooperative non-zero games with the AODV, dramatic reduction in packet drop ratio has been witnessed. However, the malicious behaviour has not been considered in the proposed system as HIDS sensors are assumed to be present in MANETs, which are anticipated to be accountable for the detection as well as the elimination of malicious nodes. Further, incompatibility of the proposed add-on to work with current MANET routing protocols reveals its incapability to mitigate other routing assaults.

A global punishment model founded on repeated game forwarding scheme has been proposed in [31], applying restrictions on selfish behaviour while administering node cooperation. The attainment of Nash Equilibrium for the cooperative state in MANETs has been possible due to emphasis by authors on diverse conditions. A stable

model has been presented in comparison to several existing schemes as the reasonability for the misbehaving node (selfish nodes) has been taken into account by the authors. However, in this study, the selfish nodes have not been considered fragile, neither any attempt was witnessed for evaluation or mitigation of malicious behaviour. Further, the system fails to offer protection against the latest bad-mouthing attack, Sybil attack, etc.

In [32], a game-theoretic system, capable of analysing the strategy profiles of regular and malicious nodes, has been presented. In this study, each node on the opposite sides is exhibited as rational regarding playing a game with every other node. The tussling has been modelled among the regular and malicious nodes as ‘Multistage Bayesian Signaling’ Game. However, the authors claim that eventually, several regular nodes might present malicious behaviour as the game continues. While designing the decision-making model for regular-malicious node game, rationality of malicious nodes towards their targets has not been considered, and hence their sole intention was to mitigate the selfish behaviour presented by regular nodes.

In [33], a framework based on game theory has been introduced using Bayesian formulation capable of analysing interactions among attacking and defending node pairs. The authors have taken into account resource and energy constraints in MANETs while contemplating Nash Equilibrium for attacker/host-based game in both the dynamic and static scenarios. As the game continues, the IDS can consistently revise its belief system regarding the malicious behaviour of the opponent, hence, offering a more realistic and dynamic game model. A unique hybrid Bayesian detection approach has been put forward by the authors, comprising of a light-weight monitoring system responsible for estimating opponent’s actions and a heavy-weight monitoring system which takes care of defence mechanism. The results revealed that the dynamic game model offers defenders’ monitoring schemes, which enhance the total detection strength of the system while at the same time proving to be an energy efficient solution.

The authors in [34] have implemented an efficient game theory-based IDS model for MANETs. The authors have emphasised that in the majority of the prior IDSs, every node runs continually a detection system resulting in subsequent overhead, especially for resource-constrained mobile devices. Game theory has been utilised to model interactions between IDS and attackers to determine whether IDS requires to run continuously without compromising its effectiveness. A model has been implemented for non-cooperative 2-layer no-zero sum game. Two game models have been constructed while taking into account imperfect as well as perfect IDS, i.e. between imperfect IDS and the attacker, and the other between perfect IDS and attacker. The solution in both the models is a mixed strategy pair of Nash Equilibrium where no players have unilateral motivation for changing their strategy. The authors have considered the significance of analysis with the establishment of optimal defence strategies deployable by network administrators.

In [35], cooperation incentive strategies offered by two distinct systems are being analysed with the support of game theory along with non-cooperation incentive scheme. The threshold values have been manipulated by the authors to determine the trustworthiness of nodes in the reputation model, and price-based system governs the

return of cooperative nodes due to the influence of wealthy nodes. The priority of integrated framework can be seen on an individual reputation system as depicted by simulation results.

The authors in [36] presented an assessment on several approaches existing in network security and privacy. The prime objective was addressing numerous current research issues present in the field of network security with the help of game theory tactics. From the research conducted, it was revealed that game theory holds the vast scope in offering solutions to various contemporary and emerging issues in network security.

In [37], the authors have presented a model based on game theory approach capable of detecting attacker/malicious nodes in MANETs. Each user is deliberated as payoff increasing strategic agent. There exist no limits on the malicious attacker strategies being adopted, and a “Fictitious Play” has been employed for carrying out genuine user action. The authors recognised the worst-case equilibrium and eventually identified the effectiveness of network topology.

Apart from these schemes, Table 1 highlights the contributions and the associated shortcomings of additional research works conducted formerly in MANETs concerning the mitigation of node misbehaviour.

Table 1. Mitigation of node misbehaviour in MANETs.

Author	Contribution	Result obtained	Limitations
Wang et al. [38]	Presented mean field game theoretic approach for enhancing MANET security	<ul style="list-style-type: none"> – Significantly improves the lifetime of MANET and reduces the compromising probability – Enables a distinct node in MANET to make shared-out security defence decisions 	<ul style="list-style-type: none"> – Scenario of multiple defenders and multiple attackers not considered
Hamdi and Abie [39]	Proposed a model based on game theory for IoT adaptive security emphasising on e-Health applications	<ul style="list-style-type: none"> – Extends the smart-things’ lifetime by 47% in comparison to the existing models – Strikes a balance between energy-efficiency and security-effectiveness 	<ul style="list-style-type: none"> – Simulated on limited threat scenarios only
Abegunde et al. [40]	Presented a dynamic game for IEEE 802.15.4 and IoT in which nodes can select and adapt their strategies of play according to the ‘state of the game’ and their energy level	<ul style="list-style-type: none"> – Better performance and security over the default IEEE 802.15.4 access mechanism – Improvement in utility, and fairness in channel sharing, as well as efficiency in energy usage 	<ul style="list-style-type: none"> – Proposed model does not account for the reality of variation in the loads level

(continued)

Table 1. (*continued*)

Author	Contribution	Result obtained	Limitations
La and Cavalli [41]	Presented a node misbehaviour detection algorithm by employing weighted-link in a hierarchical 6LoWPAN sensor network	<ul style="list-style-type: none"> – Supported by some experiments in the real platform displaying promising results with lesser false positives and no false negatives 	<ul style="list-style-type: none"> – Mobility of nodes has not been considered – Vulnerable to some complicated attacks/intrusions in application and network layer
Das et al. [42]	Put forward a new game theoretic approach for selfish node detection in MANET	<ul style="list-style-type: none"> – Guarantees the least idle time and secure low-cost data transfer 	<ul style="list-style-type: none"> – Presence of malicious nodes has not been considered
Taheri et al. [43]	Presented an approach for detecting malicious nodes using game theory	<ul style="list-style-type: none"> – Showed better efficiency in malicious node detection and lesser false positives than previous algorithms 	<ul style="list-style-type: none"> – Multiple attacker-defender scenarios have not been considered
Rajkumar and Narsimha [44]	Proposed a CA distribution and trust-based threshold revocation mechanism to enhance MANET security	<ul style="list-style-type: none"> – Eliminates misbehaving nodes – Simulation revealed better delivery ratio, resilience and packet drop 	<ul style="list-style-type: none"> – Network overhead, inaccuracy, and slow revocation issues
Sengathir and Manoharan [45]	Developed a security add-on for multicast ad-hoc on-demand distance vector protocol	<ul style="list-style-type: none"> – Effective in the detection of misbehaving nodes 	<ul style="list-style-type: none"> – No clear distinction between malicious and selfish nodes – Malicious nodes have been modelled as fragile – Cannot be applied to other routing protocols

3 System Design

In MANETs, malicious nodes are detected when such a node acts irregularly causing a subsequent decrease in network performance. In the current study, the exceptional actions exhibited by selfish nodes are revealed as well as the mitigation of malicious nodes behaviour is achieved by the adoption of Bayesian Signaling (BS) game model. The game model is offering a reliable and secure communication among nodes. A multi-stage game theory has been considered for handling the security issues in large scale MANETs.

A two-player strategy involving both the sender and receiver has been considered in the BS game model. The type of sender is depicted by node behaviour. The receiver

does not require to observe the sender's nature/behaviour. According to behaviour type of sender, forwarding of data is done from a set of possible messages $\{I = [i_1, i_2, \dots i_n]\}$, while at the receiver side, the message is being noticed without knowing sender type. Then as per the set of actions $A = \{C, D\}$, where C specifies the 'cooperate' and D specifies 'decision', the receiver selects the appropriate action. The payoff values are collected by two players that depend on the type of sender; the receiver is choosing actions while the sender selects messages/data.

The BS model includes a game plan under it, i.e. Bayesian Equilibrium (BE), which highlights the corresponding concerns; like a message $[i * (S_t)]$ is forwarded by a sender of type (S_t) in the probability distribution set I. As per the nodes probability, any action is performed by the receiver from the action set (C or D), while the sender S_t contemplates any message i from set I.

Besides, the belief system update as well as the payoff evaluation mechanism are the deciding factors in node strategy. The node strategy can be pure, mixed, or BE strategy. Node behaviour can be corrected randomly while in case of pure approach, no alteration of node behaviour is done whatever the circumstances in mixed strategy. The BE provides with the strategy profile and the belief system update depending on the nodes' type. Depending on the payoff value, pure procedure chooses a necessary action, while the belief system is revised for the rest of the nodes in case of a mixed strategy. In this study, both the relay as well as the sender nodes are deliberated as malicious nodes. The following algorithm highlights the approach adopted by game players to achieve optimal action.

Algorithm

Input: Sender node (S_n) and Receiver node (R_n)

Output: Identify malicious activity

Start.

 Init Sender node and Receiver node

 Define any profile strategy for S_n and R_n .

 Choose the type of node {Regular or Malicious}.

 Update the S_n and R_n beliefs by applying Bayes rule.

 Compute optimal payoff value for S_n and R_n updated beliefs

 Realize the rational action {C, D}.

 if action not rational then

 Convey to the related nodes as malicious node

 else

 Conclude action C and D

 endif

End

3.1 Assumptions

The primary assumptions of the presented model are specified below:

- i. Malicious nodes can act rationally towards their targets.
- ii. Proper modelling of malicious nodes should be done so that there are no indications of selfishness from them at any stage of the game.
- iii. Nodes may trace outgoing packets from neighbours at a one-hop distance via the passive observation/network monitoring mechanism.
- iv. There exists a less probability of committing an observation error.
- v. Each physical node is identified by an authentication mechanism, which binds an identity to node restricting the possibility of altering or faking this identity while in the cluster.
- vi. The detection policy is challenging as the malicious node leaves the cluster in which it launches an attack while at the same time it destroys all history of transactions being performed by it while in the cluster.
- vii. Monitoring trust of nodes in cluster externally is not possible.
- viii. In this study, a multi-stage game has been considered, where time is being classified in slots, and each slot represents the current game stage process.
- ix. During initial game stages, no attack is launched by malicious nodes to maximise their utility through decamping regular node trust factor.

3.2 Formulation of Payoff

The players' result for payoffs has been displayed numerically. The utility or performance of the player is evaluated. The overall process to formulate the payoff is presented below:

1. For regular guest nodes, the target gets a payoff value 'a' if it trusts, 'a' being greater than 1.
2. For malicious guest nodes that successfully assault the target, the damage incurred to the target equals a value 'a'.
3. In case the guest node is suspected by the target node, the cost incurred is equal to 1.
4. In case of reasonable doubt, the feedback message attained by the target node is said to be equal to 'b', with b ranging from 0 to 1.
5. In case the trust is invaluable, the target loses a payoff equivalent to 'b'.
6. In case there is a malicious guest node pretending to be normal through the game, the guest node is assumed to be the sender, and the target node is deemed as the receiver.
7. If the stranger node is malicious but pretends to be normal as the game proceeds.

The payoffs in Sect. 3.1 measure the players' strategy. In the Bayesian Signaling game mode, the strangers' strategy D dominates C based on the selection of trust by the end node where the payoff value equals 2 when C is selected and 3 when D is selected. If the end node chose doubt, the node attains 0 for C and 1 for D; thus, the better result is D. Likewise, the end nodes opt for a strategy of doubt that dominates trust.

3.3 Evaluating the Payoffs

In the BS model, specific misbehaving players are stimulated by payoffs that look for improved results of the same. It may signify ordinal, or cardinal payoffs and a payoff matrix is employed for calculating payoffs. The decision maker forecasts the best outcomes. Dual players (S_n and R_n) perform in the BS game model with the sender node choosing an action from an action space set to forward the message ‘i’ to the receiver.

This message ‘i’ is noticed by the receiver that replies by choosing an action from the set of actions. No private message is retained with the receiver; thus it has only a single type of node information. The receiver about the sender type retains a particular previous belief. After this, the receiver takes action; payoffs are assigned to each player by message type from S_n . The receiver acts after selecting a type of node for responding to the sender. The expected payoff opens up the players’ attitude towards a probable danger. All the players attain a payoff based on their action as well as the actions of their neighbours.

The payoffs of the malicious and regular nodes have been overviewed in Table 2. In the table, SM signifies ‘signalling malicious’, and PS signifies ‘prefers to send’. The anticipated payoff is computed as the product probability of node type and the payoff associated with each chosen action. In case of a high anticipated payoff, the corresponding action is chosen as the sender or receiver action. The expected utility of the sender is a combination of the payoffs and the receiver’s pure strategy.

Table 2. Payoffs for regular and malicious nodes.

Node type		R action	
Normal node	Malicious node	Co-operate	Decline
SM	SM	0, 0	0, 0
SM	PS	P, 0	p - 1, p - 1
PS	SM	P, p	P, 0
PS	PS	1, p	-1, p - 1

The sender opts for a single action, either to cooperate (C) or decline (D), based on the receiver’s type. The receiver being assumed a regular node, its actions comprise of C, D or report. Decline action signifies the node’s rejection to participate and cooperate action implies that the node makes itself available to communicate. The sender can perform a packet dropping assault just like the decline strategy of a regular node. The payoff values from malicious nodes are generated by the sender nodes whereas no results are received by regular nodes from D. In case the receiver chooses report, it allocates gain SM and in case of a malicious sender, the regular node allocates PS from an effectual C, where SM and PS are growing for report and cooperate respectively. These nodes can choose decline (D) that attains zero gain and no cost even if the

opponent chooses to attack. Although the receiver rejects the incoming message from the sender and even notifies the nearby nodes about the information coming from the sender being malicious or regular. The receiver opts the decline action based on BE model.

The sender purports to recommend transmitting the information to the receiver. Later, the receiver chooses either C or D action for the message offered by the sender. Nevertheless, the excellent reply from the receiver is the approval to take any message irrespective of the type of node sending it. Such a message is not captured in the profile of strategy (C, D) because the receiver's message set is not seized down that route.

4 Results and Analysis

The analysis of the behaviour of malicious nodes' performance in this system has been presented in this section. Besides, it also assesses the node strategies—pure, mixed and BE—for malicious as well as regular nodes. The parameter lists have been presented in Table 3. During its simulation, about 40% of nodes are assumed to be misbehaving. The simulation results exhibit the effects of different game levels by making a comparison of various regular node strategies.

Table 3. Performance parameters.

Performance parameters	Values
Area of simulation	500×500 m
Simulation period	1000 s
Total number of nodes	50, 85, 100, 150
Transmission rate	200 m
T_v (threshold value)	(0, 1)
Size of packet	512 Bytes

4.1 Average of Node Utility

The node utility shall display the real values of node payoffs. The average payoff can be computed by considering the expected payoff values that are taken from the payoff matrix. The expected payoff incorporating the behaviour of players towards danger shall examine the category of product probability, and thus every payoff action shall be chosen.

4.2 Nodes Strategies

In the present study, three diverse strategies are included viz. (i) pure, (ii) mixed and (iii) Bayesian Equilibrium, where nodes choose offering actions to all the players. The strategies chosen assess the utility of nodes. The comparison of these strategies with the utility of nodes has been shown in Figs. 1 and 2. The regular nodes' utility is maximum when BE strategy is followed in the first comparison. This is due to the presence of

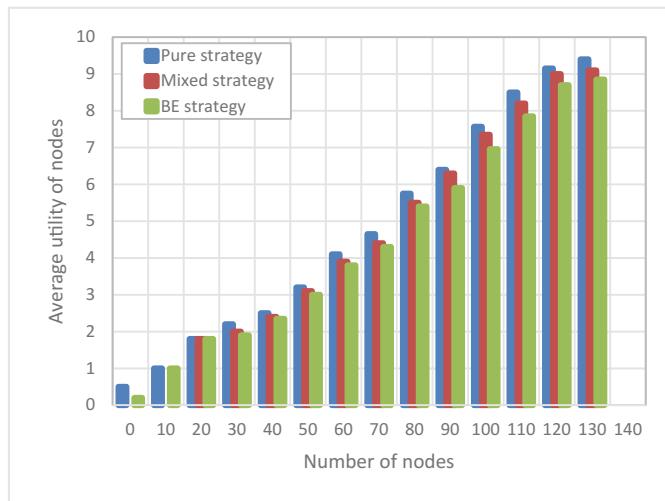


Fig. 1. Comparison of regular node utility under a malicious node strategy.

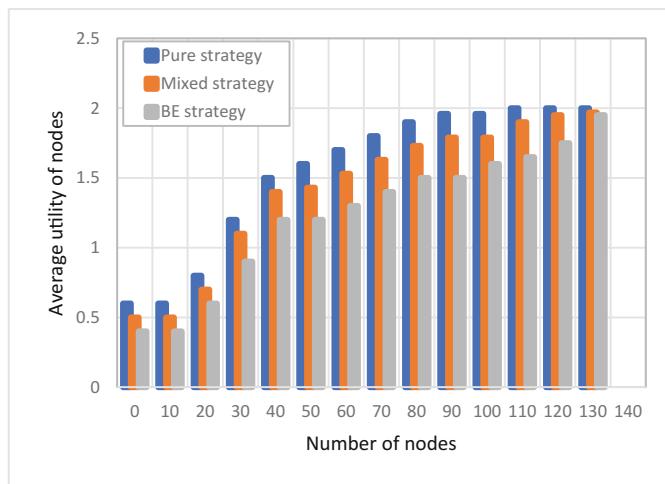


Fig. 2. Comparison of malicious node utility under malicious node strategy.

regular nodes that hold all the chances of cooperating with every regular node and with a lower proportion of malicious nodes. From Fig. 2, it is evident that the utility of malicious nodes is high. In this case, the regular nodes may choose either mixed or pure strategy; the payoff of malicious nodes is reduced, and their utility drops considerably. Also, it can be observed from Fig. 2 that BE shows efficient performance in comparison to others when malicious nodes employ a mixed or pure strategy. The outcome from the simulation reveals that the presented system using BE strategy is apt for normal nodes that lower the malicious node utility.

4.3 False Positive Rate and Detection Rate of Malicious Nodes

The malicious node detection rate and normal node misdetection rate in the proposed system have been examined after their comparison with the algorithms presented in [43, 46] when run under changing conditions. The results achieved have been displayed in Figs. 3 and 4.

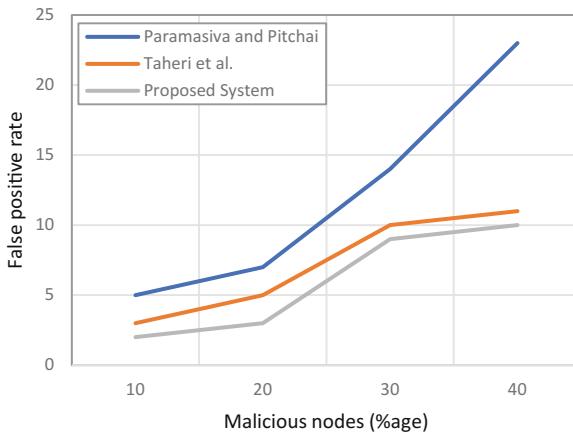


Fig. 3. False positive rate versus percentage of malicious nodes.

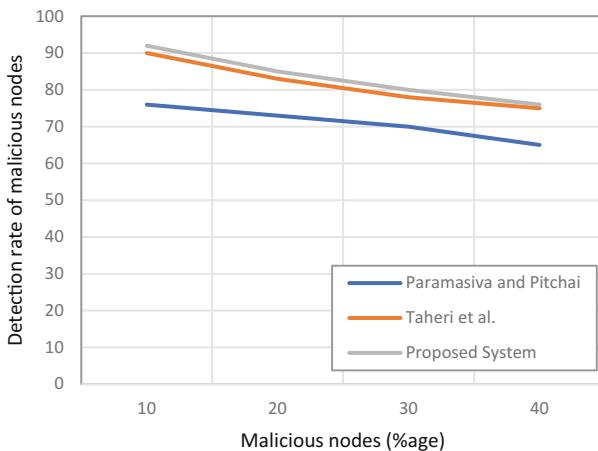


Fig. 4. Detection rate of malicious nodes vs percentage of malicious nodes.

The regular nodes' false positive rate (FPR) and the malicious nodes' detection rate have been exhibited in Figs. 3 and 4 respectively, with the percentage of malicious nodes ranging between 10 and 40. The results attained show the effective performance of the proposed system in malicious node detection in comparison to algorithms [43, 46].

4.4 Throughput and Attack Detection

Parameters such as attack detection percentage and throughput were also examined in each simulation round as the percentage of malicious nodes rose, and the outcome was contrasted with the algorithm in [46].

From Fig. 5, it is seen that the throughput drops with the growing percentage of malicious nodes in the network. This suggests the better throughput obtained in the proposed system as compared to the system in [46]. Further, Fig. 6 depicts the decline in attack detection as the malicious node percentage grows.

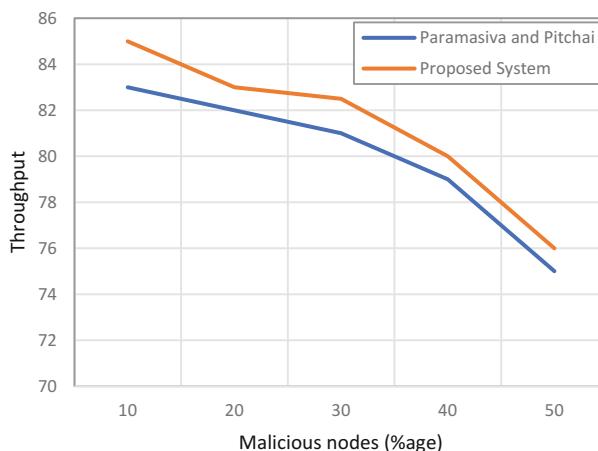


Fig. 5. Throughput versus percentage of malicious nodes.

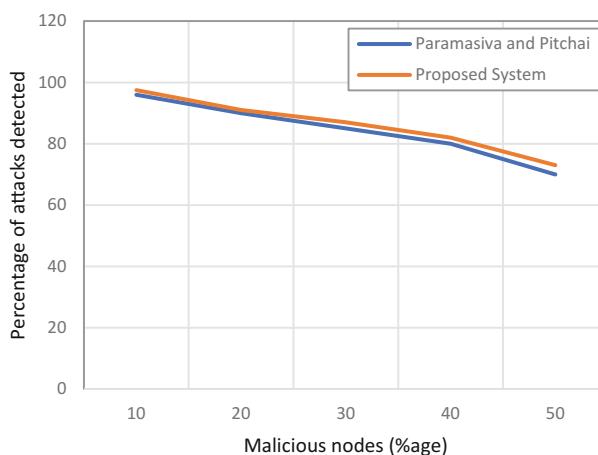


Fig. 6. Percentage of attacks detected versus percentage of malicious nodes.

5 Conclusion

This paper presents a Bayesian Signaling game model that explores malicious behaviours and actions in a MANET. A solution has been provided for the model, and threshold values are produced that can be further taken into consideration for designing MANET secure routing protocols. Malicious as well as regular nodes have been considered in this system for experimentation purposes. Furthermore, the regular and malicious node strategies have also been assessed with the goal to reduce the utility of malicious nodes and boost the regular node utility by employing the BS model. The results reveal the improvement of the proposed system over the existing systems, and therefore this system is suitable for a more secure and reliable micropayment operation in MANETs.

Acknowledgments. This work was supported by Ministry of Higher Education Malaysia (Kementerian Pendidikan Tinggi) under Research Initiative Grant Scheme number P-RIGS19-020-0020.

References

- Cheng, X., Huang, X., Du, D.Z.: *Ad Hoc Wireless Networking*. Springer Science & Business Media, United States (2013)
- Ma, Y., Jia, Z.: Evolution and trends of broadband access technologies and fibre-wireless systems. In: *Fiber-Wireless Convergence in Next-Generation Communication Networks*, pp. 43–75. Springer, Cham (2017)
- Huang, J.H., Wang, L.C., Chang, C.J.: Architectures and deployment strategies for wireless mesh networks. In: *Wireless Mesh Networks*, pp. 29–56. Springer, Boston (2008)
- Olanrewaju, R.F., Khan, B.U.I., Anwar, F., Khan, A.R., Shaikh, F.A., Mir, M.S.: MANET—A cogitation of its design and security issues. *Middle-East J. Sci. Res.* **24**(10), 3094–3107 (2016)
- Ghosekar, P., Katkar, G., Ghorpade, P.: Mobile ad hoc networking: imperatives and challenges. *IJCA (Special Issue on MANETs)* **3**, 153–158 (2010)
- Hogie, L., Bouvry, P., Guinand, F.: An overview of MANETs simulation. *Electron. Notes Theor. Comput. Sci.* **150**(1), 81–101 (2006)
- Suri, P.R., Rani, S.: Bluetooth network—the adhoc network concept. In: *SoutheastCon. Proceedings*, pp. 720–720. IEEE (2007)
- Khan, B.U.I., Olanrewaju, R.F., Ali, N.A., Shah, A.: ElePSO: energy-aware elephant swarm optimization for mobile ad-hoc network. *Pensee J.* **76**(5), 88–103 (2014)
- Khan, B.U.I., Olanrewaju, R.F., Anwar, F., Shah, A.: Manifestation and mitigation of node misbehaviour in ad-hoc networks. *Wulfenia J.* **21**(3), 462–470 (2014)
- Khan, B.U.I., Olanrewaju, R.F., Habaebi, M.H.: Malicious behaviour of node and its significant security techniques in MANET—A review. *Aust. J. Basic Appl. Sci.* **7**(12), 286–293 (2013)
- Khan, B.U.I., Olanrewaju, R.F., Mir, R.N., Baba, A., Adebayo, B.W.: Strategic profiling for behaviour visualization of malicious node in MANETs using game theory. *J. Theor. Appl. Inf. Technol.* **77**(1), 25–43 (2015)

12. Khan, B.U.I., Olanrewaju, R.F., Mir, R.N., Yusoff, S.H., Sanni, M.L.: Trust and resource-oriented communication scheme in mobile ad hoc networks. In: Proceedings of SAI Intelligent Systems Conference, pp. 414–430. Springer, Cham (2016)
13. Khan, B.U.I., Olanrewaju, R.F., Anwar, F., Najeeb, A.R., Yaacob, M.: A survey on MANETs: architecture, evolution, applications, security issues and solutions. Indonesian J. Electr. Eng. Comput. Sci. **12**(2), 832–842 (2018)
14. Olanrewaju, R.F., Khan, B.U.I., Najeeb, A.R., Zahir, K.N., Hussain, S.: Snort-based smart and swift intrusion detection system. Indian J. Sci. Technol. **11**(4), 1–9 (2018)
15. Rath, M., Panigrahi, C.R.: Prioritization of security measures at the junction of MANET and IoT. In: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, p. 127, ACM (2016)
16. Bellavista, P., Cardone, G., Corradi, A., Foschini, L.: Convergence of MANET and WSN in IoT urban scenarios. IEEE Sens. J. **13**(10), 3558–3567 (2013)
17. Olanrewaju, R.F., Khan, B.U.I., Mir, R.N., Shah, A.: Behaviour visualization for malicious-attacker node collusion in MANET based on probabilistic approach. Am. J. Comput. Sci. Eng. **2**(3), 10–19 (2015)
18. Olanrewaju, R.F., Mechraoui, A.L., Khan, B.U.I.: Game theory probabilistic application to detect misbehaving nodes in ad-hoc networks. In: Proceedings of the 2nd IEEE International Conference on Intelligent Systems Engineering (ICISE), Kuala Lumpur, Malaysia, 20–21 Mar 2018
19. Tantubay, N., Gautam, D.R., Dhariwal, M.K.: A review of power conservation in wireless mobile ad-hoc network (MANET). Int. J. Comput. Sci. Issues (IJCSI) **8**(4), 378–383 (2011)
20. Arulanandam, K., Parthasarathy, B.: A new energy level efficiency issues in MANET. Int. J. Rev. Comput. **1**(5), 104–109 (2009)
21. Singh, G., Singh, J.: MANET: issues and behavior analysis of routing protocols. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **2**(4), 219–227 (2012)
22. Parvez, J., Peer, M.A.: A comparative analysis of performance and QoS issues in MANETs. World Acad. Sci. Eng. Technol. **48**, 937–948 (2010)
23. Khan, B.U.I., Olanrewaju, R.F., Baba, A.M., Mir, R.N., Lone, S.A.: DTASR: dual threshold-based authentication for secure routing in mobile ad-hoc network. World Eng. Appl. Sci. J. **7**(2), 68–73 (2016)
24. Khan, B.U.I., Zulkurnain, N.F., Olanrewaju, R.F., Nissar, G., Baba, A.M., Lone, S.A.: JIR2TA: joint invocation of resource-based thresholding and trust-oriented authentication in mobile ad-hoc network. In: Proceedings of SAI Intelligent Systems Conference, pp. 689–701. Springer, Cham (2016)
25. Khan, B.U.I., Olanrewaju, R.F., Mattoo, M.U., Aziz, A.A., Lone, S.A.: Modeling malicious multi-attacker node collusion in MANETs via game theory. Middle-East J. Sci. Res. **25**(3), 568–579 (2017)
26. Khan, B.U.I., Olanrewaju, R.F., Baba, A.M., Zulkarnain, N.F., Lone, S.A.: STCM: secured trust-based communication method in vulnerable mobile ad-hoc network. In: 9th International Conference on Robotic, Vision, Signal Processing and Power Applications, pp. 149–161, Springer, Singapore (2017)
27. Ilavendhan, A., Saruladha, K.: Comparative study of game theoretic approaches to mitigate network layer attacks in VANETs. ICT Express **4**(1), 46–50 (2018)
28. Javidi, M.M., Aliahmadipour, L.: Game theory approaches for improving intrusion detection in MANETs. Sci. Res. Essays **6**(31), 6535–6539 (2011)
29. Janzadeh, H., Fayazbakhsh, K., Dehghan, M., Fallah, M.S.: A secure credit-based cooperation stimulating mechanism for MANETs using hash chains. Future Gener. Comput. Syst. **25**(8), 926–934 (2009)

30. Panaousis, E.A., Politis, C.: A game theoretic approach for securing AODV in emergency Mobile Ad Hoc Networks. In: IEEE 34th Conference on Local Computer Networks, LCN, pp. 985–992 (2009)
31. Wang, K., Wu, M.: Nash equilibrium of node cooperation based on metamodel for MANETs. *J. Inf. Sci. Eng.* **28**(2), 317–333 (2012)
32. Li, F., Yang, Y., Wu, J.: Attack and flee: Game-THEORY-based analysis on interactions among nodes in MANETs. *IEEE Trans. Syst. Man Cybern. Part B (Cybernetics)* **40**(3), 612–622 (2010)
33. Liu, Y., Comaniciu, C., Man, H.: A Bayesian game approach for intrusion detection in wireless ad hoc networks. In: Proceeding from the 2006 Workshop on Game Theory for Communications and Networks, p. 4, ACM (2006)
34. Marchang, N., Tripathi, R.: A game theoretical approach for efficient deployment of intrusion detection system in mobile ad hoc networks. In: International Conference on Advanced Computing and Communications, ADCOM, pp. 460–464, IEEE (2007)
35. Li, Z., Shen, H.: Game-theoretic analysis of cooperation incentive strategies in mobile ad hoc networks. *IEEE Trans. Mob. Comput.* **11**(8), 1287–1303 (2012)
36. Manshaei, M.H., Zhu, Q., Alpcan, T., Bacşar, T., Hubaux, J.P.: Game theory meets network security and privacy. *ACM Comput. Surv. (CSUR)* **45**(3), 1–35 (2013)
37. Theodorakopoulos, G., Baras, J.S.: Malicious users in unstructured networks. In: 26th IEEE International Conference on Computer Communications, pp. 884–891, IEEE (2007)
38. Wang, B., Zhu, J., Liu, K.R.: Self-learning repeated game framework for distributed primary-prioritized dynamic spectrum access. In: 4th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, pp. 631–638, IEEE (2007)
39. Hamdi, M., Abie, H.: Game-based adaptive security in the Internet of things for eHealth. In: IEEE International Conference on Communications (ICC), pp. 920–925, IEEE (2014)
40. Abegunde, J., Xiao, H., Spring, J.: A dynamic game with adaptive strategies for IEEE 802.15. 4 and IoT. In: Trustcom/BigDataSE/ISPA, pp. 473–480, IEEE (2016)
41. La, V.H., Cavalli, A.R.: A misbehavior node detection algorithm for 6LoWPAN Wireless Sensor Networks. In: 36th International Conference on Distributed Computing Systems Workshops (ICDCSW), pp. 49–54, IEEE (2016)
42. Das, D., Majumder, K., Dasgupta, A.: Selfish node detection and low-cost data transmission in MANET using game theory. *Procedia Comput. Sci.* **54**, 92–101 (2015)
43. Taheri, Y., Garakani, H.G., Mohammadzadeh, N.: A game theory approach for malicious node detection in MANETs. *J. Inf. Sc. Eng.* **32**(3), 559–573 (2016)
44. Rajkumar, B., Narsimha, G.: Trust-based certificate revocation for secure routing in MANET. *Procedia Comput. Sci.* **92**, 431–441 (2016)
45. Sengathir, J., Manoharan, R.: Security algorithms for mitigating selfish and shared root node attacks in MANETs. *Int. J. Comput. Netw. Inf. Secur.* **5**(10), 1–10 (2013)
46. Paramasiva, B., Pitchai, K.M.: Modeling intrusion detection in mobile ad hoc networks as a non-cooperative game. In: International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), pp. 300–306, IEEE (2013)



End-to-End Emotion Recognition From Speech With Deep Frame Embeddings And Neutral Speech Handling

Grigoriy Sterling and Eva Kazimirova^(✉)

Neurodata Lab, Moscow, Russia
sterling@phystech.edu,
e.kazimirova@neurodatalab.com

Abstract. In this paper we present a novel approach to improve machine learning techniques in emotion recognition from speech. The core idea is based on the fact that not all parts of the utterance convey emotion information. Thus, we propose to separate a given utterance into emotional and neutral parts and clean up the database to make it more univocal. Then we estimate short speech interval embeddings using speaker recognition convolutional neural network trained on the VoxCeleb2 dataset with the triplet loss. Sequences of these features are processed with a recurrent neural network to get an emotion label for the considered utterance. This stage consists of two sub-stages. At the first one we train a model to recognize neutral frames in a given utterance. Next we separate a corpus into emotional and neutral parts and train an improved model. Our experiments on the IEMOCAP corpus show that the final model achieves 66% of unweighted accuracy (UA) on four emotions and outperforms other known approaches like out-of-the-box Connectionist Temporal Classification (CTC) and local attention by more than 4%.

Keywords: Emotion recognition from speech · Human-computer interaction · Triplet loss · Connectionist temporal classification

1 Introduction

Despite emotion recognition being one of the most popular fields in the last decade, there is still much scope for improvement. The speech signal is challenging for researchers due to its multi-level structure where one can distinguish segments relying on phonemes, words, utterances or various features such as energy. The speech signal is modulated by many factors and conveys information about the speaker's state as well as the environment. Due to these reasons speech analysis is demanding, and even recognition of basic emotions such as Happiness, Sadness and Anger today does not give a hundred-percent accurate result. It should be noted though, that even human performance is far from perfect.

In this work we consider the problem of emotion recognition from speech. We separate this paper into 4 logical parts. The first one describes history of the problem and known techniques. Next we propose an effective method to estimate audio features for neural network and a strategy to train it properly. We also compare our results with reported ones in other papers. In conclusion we suggest some ways to improve the proposed approach.

So, for the emotion recognition purpose different acoustic features can be used as an input to the machine learning algorithm. Some researchers use utterance-level features for emotion classification such as different statistics across all duration [2]. Another way is to assume that some fragments of the utterance contain more relevant information in relation to the emotion recognition purpose. Since the utterance is considered to be emotionally non-uniform, it can be divided into emotional and non-emotional (or neutral) parts based on some criteria. Using frames with high energy as the key frames for emotion recognition and considering other frames non-emotional is one possible approach [4, 17]. Other researchers conclude that different phoneme classes are the most informative parts of the phrase and consider a phoneme-based approach to be effective [2]. The segment-based emotion recognition uses speech segments without any prior knowledge about words or phonemes segmentation [14].

To compare different approaches, it seems reasonable to use a common dataset. Among various emotion corpora, The Interactive Emotional Dyadic Motion Capture (IEMOCAP, [3]) dataset is one of the most widely used. IEMOCAP consists of annotated audiovisual data for dyadic interactions. It contains both scripted scenarios and improvisations.

In [11] the researchers assumed frames to have different significance for emotion recognition. They used only improvised session of IEMOCAP and could recognize different states (Happy, Sad, Angry and Neutral) with a 63.89% unweighted accuracy (UA).

The authors of [15] achieved on IEMOCAP an UA of 58.8% with emotion low-level descriptors (LLDs) and 56.3% with raw spectral features using an RNN with attention mechanisms. This architecture was designed to focus the network on emotionally salient parts of a sentence.

Along with LLD features, representations of data from hidden layers of neural networks turned out to be effective [12]. The representation learning on spectrogram extracted either from speech waveform or glottal flow signal was performed by [6]. The researchers found out that the learned features obtained with an autoencoder are discriminative and applicable to a discrete emotion recognition task. The best UA gained in the [6] for the overall IEMOCAP dataset was 51.86% for four basic emotions (Neutral, Angry, Sad, Happy).

Speaking of the most effective approach in emotion recognition, it should be mentioned that Satt and co-authors [13] achieved a 61.7% class accuracy on IEMOCAP using raw spectrogram as a CNN input. In this work an emotion label was assigned to each segment of the utterance. Spectrograms, that were processed and fed into the recurrent layers of a neural network, were shown to be more effective than LLDs. However, emotional corpora are usually too small

to train a model properly. Is some works the idea of using pre-trained models for features estimation was proposed.

The question of how to estimate the most useful features is related to representation learning. One of the ways to obtain discriminative feature vectors is to use the triplet loss. The main idea of the triplet loss paradigm is to create a fixed-length speaker-discriminative embedding [10] based on distance, which helps to maximize the difference of distances between points in different classes. Using the triplet loss method was shown to be effective for the task of speaker verification in short utterances [19]. Intuitively, these speaker embeddings should contain information about common properties of a given speech segment like timbre, intonation and emotionality.

2 Method

In our work we focus on the fact that a one-labeled utterance contains neutral (non-emotional) parts along with emotionally meaningful segments, and investigate what features are most relevant for the emotion recognition task.

The proposed features estimation technique is found upon the idea that we can process short spectrograms to obtain low-dimensional representations without any crucial information loss. In order to construct a proper embedding model we consider speaker verification problem and train a neural network to minimize the triplet loss as an objective function to localize discriminative embeddings in feature vector space. As a core model we use a convolutional neural network that processes spectrograms as regular images. Next we represent a given utterance as a sequence of overlapping frames and obtain a sequence of embeddings that are processed by the recurrent neural network to learn time-based emotional patterns.

Finally, we train a model with a special objective function and decode a sequence of RNN outputs into a single label. This model contains the second key feature of the proposed approach, namely a special method to recognize and handle neutral frames in a given utterance. It relies upon the fact that emotionality is not uniform over the whole utterance, so the core idea involves splitting utterances into parts that have only one emotional label. In order to do that we train a model to differentiate between the neutral and the emotional state in utterances of different lengths, varying from 1 to 30 s. Then we predict emotions for short (1s) parts of utterances to obtain time-distributed labels and split the speech in the training set into single-label segments, and train a new model. Note that the process can be performed iteratively several times. The algorithm is described more thoroughly in Sect. 2.2.

The complete emotion recognition model is presented in Fig. 1.

2.1 Short-Term Frame Embeddings

In machine learning the choice of features for the input is an important issue. Raw audio signal contains the most complete information but seems to be very

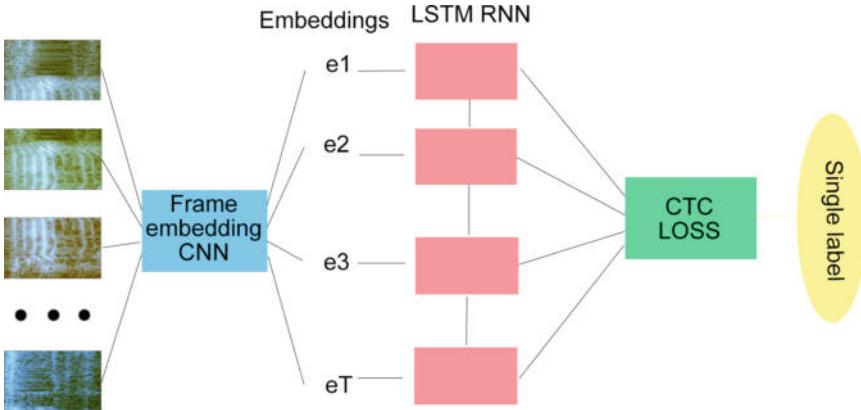


Fig. 1. Emotion recognition model architecture. First, it processes frames into embeddings that are fed into the recurrent neural network with LSTM cells, that returns time-distributed outputs. They are collapsed into a single label by a CTC decoding algorithm

surplus. On the other hand, hand-crafted audio descriptors or mel-spectrograms are more compact but some useful information can be lost during their estimation.

If we consider only information-lossless features estimation strategies, spectrograms appear to be a good alternative to a raw signal. They are computed as Fourier transform of speech signal in sliding windows of a short length (about 10–30 ms). The information in spectrograms is virtually the same as in initial audio, because it can be obtained via inverse Fourier transformation of spectrograms and some post-processing (see the Griffin-Lim algorithm [8] and its modifications).

The main advantage of using spectrograms is that they can be represented as pictures and processed via fast and effective convolutional neural networks.

In this work we suggest viewing speech as a sequence of overlapping frames. Under one voice frame we understand a short speech segment of 0.5 s length, that covers a few phonemes. An example of a frame is shown on Fig. 2

Thus, we can process each frame with a time-distributed neural network to obtain high-level frame embeddings, so-called d-vectors, which are fed into recurrent layers.

However, in relation to the emotion recognition task there are no datasets rich enough to train out-of-the-box spectrogram-based models properly. One of the possible ways to improve emotion recognition performance is to use pre-trained models for the features estimation purpose. The main advantage of high-level features as opposed to raw input processing is dimension reduction without any critical information loss. Moreover, as we show in the Experiments section, qualitative emotional features can be obtained in the course of performing other tasks.

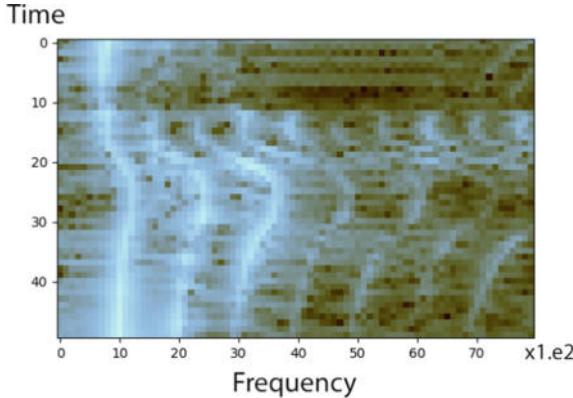


Fig. 2. Color log-spectrogram of a half-second signal. To take into account spectrogram shape and power we also estimate frame-normalized and histogram-equalized log-spectrograms [1]. We regard these three inputs as RGB channels of one color image and process it with a convolutional neural network

In order to train that model properly, after a few epochs with categorical cross-entropy loss function we finally apply the triple loss proposed in [19]. It forces points from different classes to be much farther than points inside one class in the embedding space. For a given anchor frame x^a let x^- and x^+ be negative and positive exemplars from the dataset, in other words x^- and x^+ have a different and the same class as x^a , respectively. So, let e be the function that estimates embedding for x . During the training phase we try to minimize

$$L = \sum_{i=1}^N \left[d(e(x_i^a), e(x_i^-)) - d(e(x_i^a), e(x_i^+)) + \alpha \right]_+ \quad (1)$$

where d denotes Euclidean distance measure and α is a hand-chosen threshold. This loss is small if the distance between points in different classes is more than the sum of α and the distance inside one class.

2.2 Connectionist Temporal Classification

Another challenging problem in emotion recognition from speech is that frames in a given utterance have different significance in terms of emotion recognition. In other words, some frames can be more neutral or more emotional than others. So, if we separate all frames into two emotional classes, an emotional utterance may potentially include some neutral frames, and a neutral utterance may include emotional frames as well. This fact leads to emotion database ambiguity and overfitting: the model tends to remember samples instead of finding complex patterns in them.

As can be seen from the related work, there are a few approaches to working with soft-labeled sequences. In addition to a branch of methods based on frame

weights estimation we chose the Connectionist Temporal Classification (CTC), which is aimed at classifying each element in a sequence. For this purpose a specific loss function is constructed.

Consider a classic sequence-to-label problem. Let $X = \{\mathbf{x}^j\}_{j=1}^N$ be the set of sequences $\mathbf{x}^j = \{x_i^j\}_{i=1}^{T_j}, x_i^j \in \mathbb{R}^n$, where the upper index denotes the sequence number and the lower index corresponds to the element number in a sequence, T_j is the length of the j -th sequence and n is the feature space dimension (embedding size). Each sequence has a global label $y \in \mathbb{N}$ from set $C = \{1, 2, 3, \dots, M\}$. Assume that each sequence element has its own local label from set $Null \cup C$, where $Null$ denotes a blank label. Finally, we should determine how sequence labeling corresponds to the global one. Consider a reducing function L , that processes a sequence of labels, removing repeats and blank labels from the input sequence, for example, $L(Null, A) = A$, $L(A, A, B) = AB$. Let us also denote $\mathbb{B}(C) = \{\pi : L(\pi) = C\}$ as a set of all possible labelings, that can be reduced to a single label C .

Our goal is to train a probabilistic model $p(\pi|\mathbf{x})$, that for a given sequence returns probabilities of x_i having label $\pi_i \in Null \cup C$. Note that the classifier part of the model is time-distributed, but the final label is single. Assuming independence, we get $p(\pi|\mathbf{x}) = \prod_{j=1}^T p(\pi_j|x_j)$. Since we only know the global label instead of local labeling, we need to construct a specific loss function. It is a sum of log likelihoods estimated for all possible local labelings from $\mathbb{B}(C)$. It is called the CTC loss:

$$Q(\mathbf{x}, C) = \sum_{\pi \in \mathbb{B}(C)} -\log p(\pi|\mathbf{x}) = \sum_{\pi \in \mathbb{B}(C)} -\log \prod_{j=1}^T p(\pi_j|x_j) \quad (2)$$

The sum could have a very large number of elements, so naive calculation seems to be very computationally expensive. However, Graves [7] suggests an efficient way to evaluate $Q(\mathbf{x}, C)$ gradients by means of dynamic programming methods, which enables training the CTC neural network with the usual back-propagation algorithm.

To make this approach more clear we need to elaborate on how the final label is predicted. At the evaluation stage we have no global labels, only frame-wise distributions. In accordance with CTC objective function minimization, we have to find the most probable global labeling. In order to do that we have to brute force all possible labelings, which is computationally very expensive, or to settle for an approximate solution in the hope of it being accurate. There are a few ways to get that approximate labeling, but in our case of emotion recognition we know that the final labeling consists of only one global label, which makes brute forcing cheap. Thus, as the final frame labeling we get $\pi = \{\pi_j = \text{argmax } p(c|x_j) \text{ over all } c \in [1..C]\}_{j=1}^T$, and then we just have to apply the reducing function L .

2.3 Neutral Frames Handling

The CTC approach, however, does not handle neutral parts of speech properly. To illustrate this, let us assume that ground truth labels are assigned to any frame in a given utterance. The resulting emotional pattern can have more than one label, including emotional as well as neutral states. In this case at the training stage the labeling will not represent ground truth emotions correctly. Our idea is to train the model consistently, splitting utterances into single-label parts with each iteration. The process would be as accurate as the model on the previous iteration, and the best way of choosing the model is discussible.

Assuming that we know exact labels in the training set, we propose to exclude neutral frames from emotional utterances and exclude emotional frames from neutral utterances (that have a neutral ground truth label). Since at this stage the CTC model learns to predict only one label, we need to slice given utterances into short parts to get time-distributed labelings. In addition to the global accuracy measure of this model, we have to balance between very poor and greedy cleaning. In other words, we have to control both precision and recall, so we have to take into account, say, F1 measure on neutral class only. Examples of patterns that we got are shown in Table 1.

Table 1. Examples of local labelings obtained in the dataset cleaning procedure

Utterance	True label	Obtained labeling
Ses01F_impro06_F009	Emotional	EENNEENNNEEEEEEEEEEENN
Ses01M_impro03_M021	Neutral	NNNNNNNEENNNNNNNNNNEE
Ses03F_impro08_M010	Emotional	NNEEE
Ses03F_impro07_M000	Neutral	NNNEE

Further, we separate utterances in the whole training set and train a new model to recognize emotions. To illustrate how it works let us consider training sample $\{x_i, y_i\}_{i=1}^N$. Each utterance x_i is a sequence of frames and has a ground truth label $y_i \in N$. Our goal is to train a classifier $f^0(x)$ that would predict the global label for x . Next we evaluate local labelings of x_i as a sequence of predictions for short parts of x_i . Then we split x_i into overlapping subsequences $\{z_{i,s}^0\}_{s=1}^T$ of length Q and take only emotional parts if y_i denotes an emotional label, and only neutral parts if y_i is a neutral label. For example, if an emotional utterance x_i of length 16 has a local labeling “ENNEEEENNEEEEEE”, it transforms into three new utterances with labelings “E”, “EEEEEE” and “EEEEEE” of lengths 1, 5, and 6 respectively. One can also skip very short utterances. After gathering all sequences into a new sample $\{x'_i, y'_i\}_{i=1}^{N'}$ we are ready to train a more accurate model $f^1(x)$. So, the scheme of the proposed algorithm is as follows:

```

1:  $x, y \leftarrow \text{initialize}$ 
2: repeat
3:    $f \leftarrow \text{Train}(x, y)$ 
4:    $x', y' = \emptyset, \emptyset$ 
5:   for  $i$  in  $[1..N]$  do
6:      $z \leftarrow \text{Split}(x_i, y_i, f)$ 
7:     for  $s$  in  $[1..len(z)]$  do
8:        $x' \leftarrow x' \cup \{z_s\}$ 
9:        $y' \leftarrow y' \cup \{y_i\}$ 
10:    end for
11:     $x, y \leftarrow x', y'$ 
12:  end for
13: until Convergence

```

where the *Train* procedure fits the classifier with x and y and function *Split* divides the input sequence into single-label subsequences. Note that predictions should be made on previously unseen data to prevent model overfitting. This can be achieved by means of out-of-fold cross-validation: we divide the sample into K parts. $K - 1$ folds are used for training and the last one—to make predictions, evaluate the performance, etc.

3 Experiments

In order to train a frame-embedding model we use the VoxCeleb2 speaker verification corpus [5]. It consists of more than 1.1 million utterances of 6114 celebrities, about 2000 h in total. All speakers are in different environments and conditions, therefore we expect high-level features to contain the information about the speaker, his/her emotion, background noises, etc. So, for a given half-second speech interval we train the model to recognize the speaker, and the second-to-last layer is used as an estimator of frame embeddings. We also normalize feature vectors corresponding to the Euclidean l2 norm.

The architecture of the speaker recognition model is relatively simple. We took the residual network (ResNet) with 18 layers as the convolutional part and fed its *pool_1* output [9] into two fully-connected layers with 1024 and 512 nodes. The last fully-connected layer with softmax activation has 6114 outputs and returns conditional probabilities of the given frame being produced by each speaker from the set. This architecture is shown in Fig. 3 in detail.

In terms of metrics we did not get any competitive results in speaker verification task: 51% with the categorical cross-entropy objective function and 53% with the triplet loss. The main reason of such low scores is that we considered only half-second intervals, which could not be enough for efficient speaker recognition. Moreover, as we have mentioned before, the question about model selection is open. However, we will further show that short-term frame embeddings help to improve emotion recognition accuracy by a few percent.

After features processing methodology is done, we evaluate the proposed iterative approach on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [3]. It is divided into five sessions, each consisting of emotional

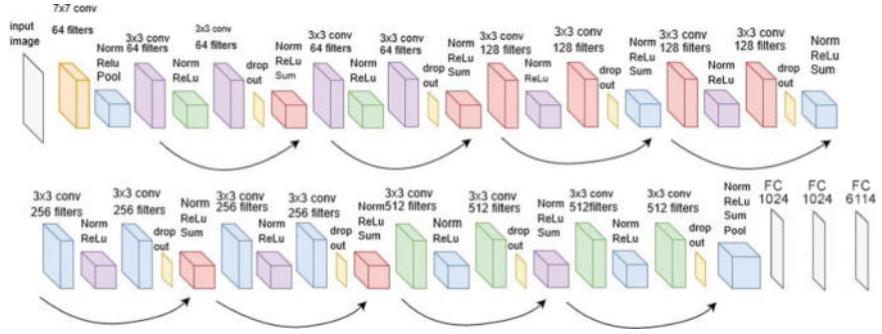


Fig. 3. Speaker verification model based on the residual network (ResNet) with 18 layers. It is a widely-used CNN architecture. It has four complex convolutional blocks with residual connections and three fully-connected layers

dialogues between two unique actors (male and female). Labeling was done at the utterance level by 3–4 annotators and involved eleven emotion categories, although frequent enough were only five emotions: anger, happiness, neutral state, sadness and excitement. In some studies different emotion sets were used. Usually it is a set of four emotions: angry, sad and neutral states are used always, and the fourth one is either excited or happy. One can also consider scripted and improvised subsessions that differ by naturalness of speech.

For evaluation we implement cross-validation over sessions, which guarantees speaker-independence of our model.

To measure performance we use both unweighted and weighted accuracies, like it was done in previous works. Weighted accuracy (WA) is the percent of points in the test set, that were classified correctly. However, since the IEMOCAP corpus is very imbalanced, unweighted accuracy (UA) is a bit more informative. It is calculated as the average recall of each class. Formulas for evaluation metrics are shown below:

$$WA = \frac{\sum_{i=1}^N f(x_i) == y_i}{N} \quad (3)$$

$$UA = \frac{1}{C} \sum_{c=1}^C \frac{\sum_{i=1}^N (f(x_i) == c) \& (y_i == c)}{\sum_{i=1}^N y_i == c} \quad (4)$$

where ‘==’ denotes the indicator operation (it equals 1 if and only if the left and the right operands have the same values, and returns 0 else) and & is *and* bool operation.

We divide the experiments into three parts. Firstly, we have to verify that short-term speaker embeddings improve emotion recognition performance, results are shown in Table 2. Next we apply cleaning procedure to the IEMOCAP corpus and obtain results both on initial and cleaned test sets to infer how cleaning process influences classification accuracy. Finally, we evaluate our approach on the IEMOCAP database with different subsessions used, to compare

our results with those reported in previous works. In all experiments we consider only *Angry*, *Sad*, *Happy* and *Neutral* states.

Table 2. Efficiency of the speaker verification model as a feature estimator. We fix the model architecture and compare three cases that differ by weight initialization strategy: random initialization, weights from speaker verification model with categorical cross-entropy loss (CE), and the same one with the triplet loss (TL). Results for the *Angry*, *Sad*, *Happy* and *Neutral* emotion set are averaged by out-of-session cross-validation. The best performing cases are given in bold

Model	UA (%)	WA (%)	Gain in UA (%)
CNN-LSTM-CTC ^a	54.58	56.62	–
CE-CNN-LSTM-CTC	60.38	61.41	5.80
TL-CNN-LSTM-CTC	61.97	63.22	7.39

Table 3. Performance of the proposed dataset cleaning procedure. We compare two iterations of cleaning process and report accuracies both on initial test sessions and on the test sessions modified in the same way

Iteration	Test set	UA (%)	WA (%)
0 ^a	Initial	61.97	63.22
1	Initial	63.11	65.84
1	Modified	65.95	67.39
2	Initial	63.40	65.77
2	Modified	65.86	67.44

^aas reported in Table 2 for TL-CNN-LSTM-CTC model

As we can see from Table 3, the results for the first and second iterations are very similar, which allows us to stop the iteration process.

Comparison between the related work and the proposed approach is illustrated by Table 4.

4 Discussion

According to Table 3, there is a 1% gap in scores between the models on the initial test set. It indicates that the proposed cleaning process makes the dataset more convenient for training. Moreover, if we clean utterances from the test set as well, we obtain much better results. However, on the cleaning stage our approach takes into account ground truth labels (we take emotional or emotionless subsequences depending on the global label), that are unavailable in the wild. So, in that case

Table 4. Results reported in previous works in comparison with the proposed approach. In fact, we follow all experimental setups used in those studies. The considered emotion set consists of Angry, Happy, Sad and Neutral emotional states

Model	Subsessions	UA (%)	WA (%)
Ours ^a	Both	61.97	63.22
Ours ^b	Both	65.95	67.39
Chernykh[4]	Both	54	54
Tripathi[16]	Both	55.65	—
Mirsamadi [15]	Both	58.8	63.5
Ours ^a	Improv.	68.90	71.21
Ours ^b	Improv.	71.14	72.44
Satt [13]	Improv.	62.0	67.3
Xia [18]	Improv.	62.4	60.9
Tripathi[16]	Improv.	62.72	—
Jinkyu Lee[11]	Improv.	63.89	62.85

^aon initial test set

^bon modified test set

we suggest finding neutral parts first, as in the sixth step of Algorithm 1, and then classifying only emotional subsequences.

However, there may be a better way to take into account emotionality distribution across utterances. We leave it for the future work and only describe the main idea. It is based on the CTC ability to work with non-single labels. At the first stage we just need to relabel utterances instead of splitting them, and get a sequence of labels for classification. In this case the recurrent part of our neural network would predict probabilities that could be decoded by CTC into sequences, that are more informative than single labels. This approach allows us to avoid the necessity of splitting training utterances as well as the influence of possible misclassified frames. On the other hand though, the information about timing of neutral and emotional frames would be lost.

To sum it up, in this work we proposed a novel approach to emotion recognition from speech. It provides solutions for two challenging problems: features estimation and dealing with neutral fragments in emotional speech.

As a feature estimator we suggest using embeddings for short frames, obtained from the speaker verification model. Although it is not guaranteed that emotional information would not be lost from embeddings, in our experiments we show that the proposed feature estimation strategy outperforms any other known approaches by 3–5%.

We also attempt to solve the neutrality problem by means of the iterative procedure of splitting utterances within the dataset. At the first iteration we train the model to recognize emotions on the given dataset. Next we find neutral and emotional parts of each utterance and get a new sample x', y' by splitting

the original one. The process is converged rapidly: in our experiments just one iteration was enough to achieve the highest performance.

In general, the proposed approach outperforms other state-of-the-art ones by 7–10%, depending on experimental setups.

Acknowledgements. All of this work is a part of Emotion Recognition Project at Neurodata Lab company.

References

1. Pitaloka, D.A., Wulandari, A., Basaruddin, T., Liliana, D.Y.: Enhancing CNN with preprocessing stage in automatic emotion recognition. *Procedia Comput. Sci.* **116**, 523–529 (2017)
2. Bitouk, Dmitri, Verma, Ragini, Nenkova, Ani: Class-level spectral features for emotion recognition. *Speech Commun.* **52**(7–8), 613–625 (2010)
3. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **42**(4), 335–359 (2008)
4. Chernykh, V., Sterling, G., Prihodko, P.: Emotion Recognition From Speech With Recurrent Neural Networks (2017)
5. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: *INTERSPEECH* (2018)
6. Ghosh, S., Laksana, E., Morency, L.P., Scherer, S.: Learning Representations of Affect from Speech, pp. 1–10 (2015)
7. Graves, A., Fernndez, S., Gomez, F.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the International Conference on Machine Learning, ICML*, vol. 2006, pp. 369–376 (2006)
8. Griffin, D., Lim, J.: Signal estimation from modified short-time fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **32**(2), 236–243 (1984)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *CoRR*, abs/1512.03385 (2015)
10. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. *Lecture Notes in Computer Science (including Subseries Lecture Notes Artificial Intelligence Lecture Notes Bioinformatics)*, vol. 9370(2010), pp. 84–92 (2015)
11. Lee, J., Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition. In: *Interspeech*, pp. 1537–1540 (2015)
12. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986) Commentary from News and Views section of Nature
13. Satt, A., Rozenberg, S., Hoory, R.: Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms, pp. 1089–1093. IBM Research-Haifa (2017)
14. Schuller, B., Rigoll, G.: Timing levels in segment-based speech emotion recognition. In: *Proceedings of INTERSPEECH 2006, Proceedings of International Conference on Spoken Language Processing ICSLP*, pp. 1818–1821 (2006)
15. Zhang, C., Mirsamadi, S., Barsoum, E.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: *Proceedings of 42nd IEEE International Conference on Acoustics Speech, and Signal Processing ICASSP 2017*,

- pp. 2227–2231. Center for Robust Speech Systems , The University of Texas at Dallas , Richardson , TX 75080, USA Microsoft Research, One Microsoft Way, Redmond , WA 98052 , USA (2017)
- 16. Tripathi, S., Beigi, H.: Multi-Modal Emotion Recognition on IEMOCAP Dataset using Deep Learning
 - 17. Wang, Z.-Q., Tashev, I.: Learning utterance-level representations for speech emotion and age / gender recognition using deep neural networks. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing—Proceedings, pp. 5150–5154 (2017)
 - 18. Xia, R., Liu, Y.: A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Trans. Affect. Comput.* **8**(1), 3–14 (2017)
 - 19. Zhang, C., Koishida, K.: End-To-end text-independent speaker verification with triplet loss on short utterances. In: Proceedings of Annual Conference Inter Speech Communication Association, INTERSPEECH, 2017-August(August), pp. 1487–1491 (2017)



Algorithm and Application Development for the Agents Group Formation in a Multi-agent System Using SPADE System

Alexey Goryashchenko^(✉)

Federal Research Center “Computer Science and Control” of RAS, Moscow,
Russia

a.goriashchenko@gmail.com

Abstract. This paper describes the group formation problem in multi-agent systems. A finite set of agents should be divided into the best set of groups. The task is to form the groups of agents of several types, which have different properties and abilities, so that they to be able to achieve the goal which cannot be achieved by any single agent on its own. To date, several group formation algorithms have been proposed: optimal algorithms are able to work with up to 50 agents; approximate algorithms can handle the group formation problem if number of agents is about hundreds or thousands. In this work SPADE multi-agent system is chosen after evaluation of its scalability and performance for several hundreds of agents. Greedy algorithm for agents group formation is proposed and implemented using SPADE multi-agent system. Characteristics of implemented algorithm are studied. Results show that performance of proposed SPADE based framework allows the implementation of more complex algorithms for agents group formation, which may provide results closer to optimal.

Keywords: Agents and multi-agent systems · SPADE multi-agent system · Agent group formation · Performance evaluation

1 Introduction

Agent-based modeling is widely used for the study of properties of complex dynamic systems. Agents are parts of the whole system and may communicate with each other. The main idea of this approach is to test how changes in individual agents' behaviors affect the system's overall behavior. Agents are autonomous, decentralized and no agent has full knowledge about the whole system. Agents may have different reactions for different events (incoming messages from other agents, from environment, etc.).

The paper is organized as follows. Section 2 provides the information about multi-agent systems. Section 3 presents different approaches to group formation problem. Section 4 presents experiments on SPADE multi-agent system performance. Section 5 describes algorithm for test problem simulation. Finally, Sect. 6 concludes the paper and outlines future directions.

2 Multi-agent Systems

Multi-agent systems are used for implementation of software agents with designated features. Several agent types with different properties and behavior can be implemented. Multi-agent system provides tools to manage, identify and search for agents, to transmit messages.

To date, many multi-agent systems have been proposed. Two of the most widely known and used are JADE [1] and SPADE [2] systems. Their functionality is mostly similar, but according to [3], performance of SPADE is better than JADE for agents number lower than 100. So, in this work SPADE multi-agent system was used.

3 Group Formation

According to [4], group formation is the process of association of agents in order to improve their collective capabilities.

Formally, group formation problem can be defined as a finite set of agents A and a characteristic function, which assigns a numeric value for each subset (i.e. group) of agents from A. The goal is to divide the set of agents so as to maximize the sum of the values given by a characteristic function of the groups [5].

3.1 Exact Approaches

Many approaches to find an optimal solution have been developed. For example, the current state-of-the-art algorithm, IDP, has been proposed. Specifically, the algorithm uses a novel representation of the search space, which partitions the space of possible solutions into sub-spaces such that it is possible to compute upper and lower bounds on the values of the best coalition structures in them. These bounds are then used to identify the sub-spaces that have no potential of containing the optimal solution so that they can be pruned. The algorithm, then, searches through the remaining sub-spaces very efficiently using a branch-and-bound technique to avoid examining all the solutions within the searched subspace(s). The algorithm is anytime, and if interrupted before it would have normally terminated, it can still provide a solution that is guaranteed to be within a bound from the optimal one [6].

Recently, IDP was improved using computer's Graphics Processing Unit (GPU). This provided significant speedups, new version is faster than IDP by two orders of magnitude [7]. However, these solutions are limited to 30 agents at most due to their large memory requirements.

Another approach to increase the number of agents was proposed by Voice et al. [8]. Authors consider coalition formation problems with an underlying synergistic graph, where edges between agents represent some vital synergistic link, such as communication, trust, or physical constraints. A coalition is infeasible if its members do not form a connected subgraph, meaning parts of the coalition are isolated from others. Thanks to great limitation of number of possible coalitions between agents, proposed algorithm can work with up to 50 agents [8].

3.2 Approximate Solutions

However, if number of agents is about hundreds or thousands, only approximate algorithms can handle the coalition formation problem.

In many cases, agents must be reactive and they should act as fast as possible, so there is a time limit for finding a solution. The approximation algorithm able to quickly find solutions that are within a specific factor of an optimal solution was proposed by Di Mauro et al. [9]. The paper focuses on the coalition structure generation when there are too many coalition structures to enumerate and evaluate because of limited computation time. Instead, agents have to select a subset of coalition structures on which to focus their search. In this work a stochastic local search procedure was adopted. The main advantage of using a stochastic local search is to avoid exploring an exponential number of coalition structures providing a near optimal solution. Proposed algorithm tries to build a coalition structure of good quality by merging coalitions. It considers other operations beyond coalition merging (such as, for example, splitting a coalition in two and exchanging a pair of agents between two coalitions) and it performs randomized search. While this helps the algorithm to escape from local minima, it does not provide guarantees about finding the global optimal solution. Experiments have shown that for 27 agents created algorithm was about 2 million times faster than IDP [9].

Recently, C-Link algorithm was proposed by Farinelli et al. [5]. It is based on the definition of the linkage function that indicates how convenient it is for two coalitions to be merged. The approach starts from the completely disjoint case: a partition where every coalition is composed of a single agent. Then, at every iteration, the most suitable pair of coalitions is computed. If there is a pair of coalitions for which the linkage function is positive, such coalitions are removed from the current partition and replaced with their union, in order to define the next level of the hierarchy. Otherwise the algorithm stops, and the current partition is returned. The algorithm is anytime [5].

Mechanism for forming groups of agents in package delivery domain was proposed by Van De Vijel et al. [4]. The process of forming groups of agents with possibly conflicting individual goals, in order to improve their collective capabilities, was investigated. Main questions for these agents are: should they form a group for fulfilling the specific task, or they prefer to operate alone; how to select the most promising group to join into fulfill the specific task; can agent trust other agents in the group, etc.

The domain involves a set of agents that receive packages to deliver to specified addresses. For each delivered package agent receives a reward. Agents are heterogeneous, they have different values in attributes named: honesty, memory, speed and trust.

Agents travel on a grid, and when one agent meets another, each of agents asks of itself:

1. Have I met this agent before? If so, do I have a favourable opinion of this agent?
2. Is this agent in a coalition that I am already in?
3. Do I think this agent is worth asking to join any coalitions I belong to?
4. Am I interested in joining any of the coalitions this agent is a member of?
5. Do I want to form a new coalition with this agent?

6. Should I ask this agent to assist me with any of my goals? (this is only asked if the agents now share a coalition).

As a result of answering these questions, each agent determine whether it will attempt to work with the other in a new or existing coalition [4].

Another approach to agents group formation is a sign world model, which is based on psychological views.

This model stores agent's internal information about its goals and experience and information about external objects and processes. Model also has the methods for knowledge acquisition and manipulation, and its use in different processes, i.e. perception, reasoning, goal-setting and behavior planning. The sign world model consists of four elements: a sign, which represents all entities of external environment and internal space for the agent; objects with their properties; processes; relations between objects and processes [10, 11]. This approach is very promising and now is under further development.

Many different approaches to the agents groups formation show that this problem is of sufficient theoretical and practical interest. An important aspect is the ability to manage substantial numbers of agents without significant performance loses. So, the aim of the present paper is to choose multi-agent system and evaluate its performance; to propose a way to form the groups of agents so that they are be able to achieve the goal which cannot be achieved by a single agent on its own; to implement and test this algorithm using multi-agent system.

4 Evaluation of a Multi-agent System Performance

Scalability and performance of multi-agent systems are important in order to provide a basis to test a group formation algorithm performance. Additional experiments were carried out to evaluate SPADE scalability and performance for several hundreds of agents.

In this work, two experiments to study scalability and performance of SPADE multi-agent system were carried out. Intel Core i5—2.6 GHz, 4 GB RAM, Win7 computer was used for experiments, all programs were written in Python.

Average round trip time (RTT) of each message [12], i.e., the time elapsed from when a sender agent sends a message until it receives the same message sent back by a receiver agent, was used for evaluation of scalability and performance.

To evaluate the scalability of SPADE, several sender agents and one receiver agent were used. Sender agents were sending messages to receiver simultaneously, and average RTT was measured for various numbers of sender agents. With the increase of number of sender agents from 2 to 250, the RTT increased from 0.13 to 0.46 s (Fig. 1). So, 125 times more agents caused only 3.5 times increase of RTT. This shows very good scalability of SPADE multi-agent system.

To evaluate the performance of SPADE, pairs of one sender agent and one receiver agent were used. Sender agent sent message to receiver agent, receiver agent performed time consuming computations consisting of brute force finding of all primitive numbers in range from 1 to 20,000, and then sent the result back to the sender agent.

Average RTT was measured for various numbers of agent pairs. The increase of agents pairs number from 2 to 250 led to the increase of RTT from 0.35 to 18 s (Fig. 2).

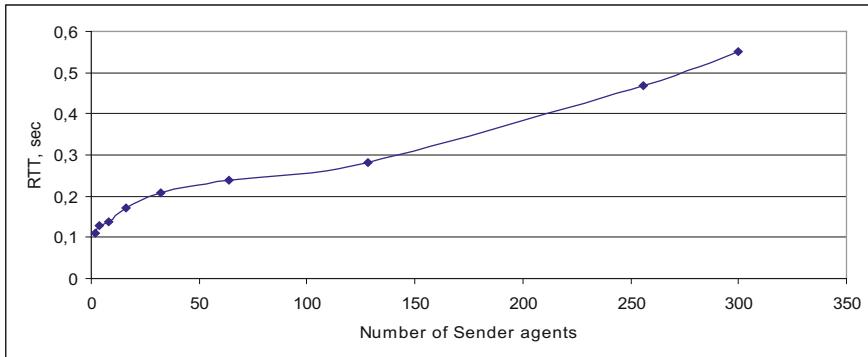


Fig. 1. Dependence of RTT on number of sender agents

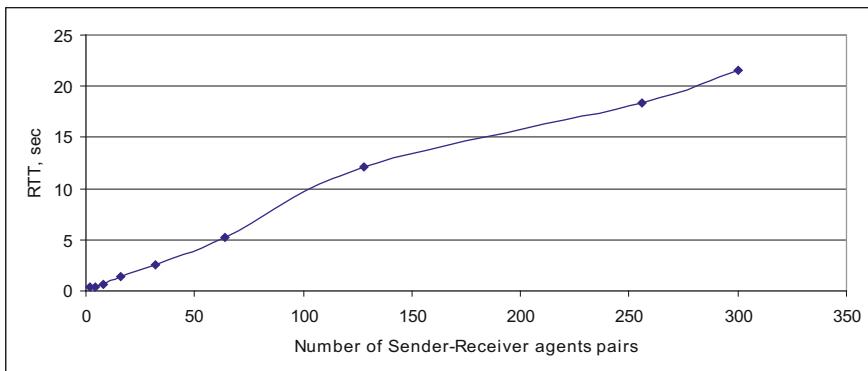


Fig. 2. Dependence of RTT on number of sender-receiver agents pairs

Such increase can be explained by the fact that experiment was performed using single computer and its CPU resources were exceeded.

These data show satisfactory results of SPADE's scalability and performance, so it is suitable for the test problem simulation used in this work.

5 Test Problem

We define the test problem for simulation as follows: there are agents of two types—factories and trucks. Each of the factories demands some amount of materials to start its operation. There are several trucks loaded with materials. Trucks can move on the road network. Capacity of a single truck is insufficient to start operation of any factory, and the speed of trucks may be different. To start the factory operation, a group trucks with

total amount of materials greater than factory demand should arrive to this factory simultaneously. The goal is to start as many factories operation as possible. The secondary goal is to minimize trucks total mileage.

For group formation algorithm implementation, two main types of agents (factories and trucks) and two auxiliary agents (Timer and Logger) are introduced.

Factory type agent has the following properties: position, materials demand and started/not started attribute. Truck type agent has the properties: position, speed, materials amount. Timer agent informs all agents about new iteration, logger agent keeps history of actions of each agent during simulation.

Sequence diagram for group formation iteration is shown on Fig. 3.

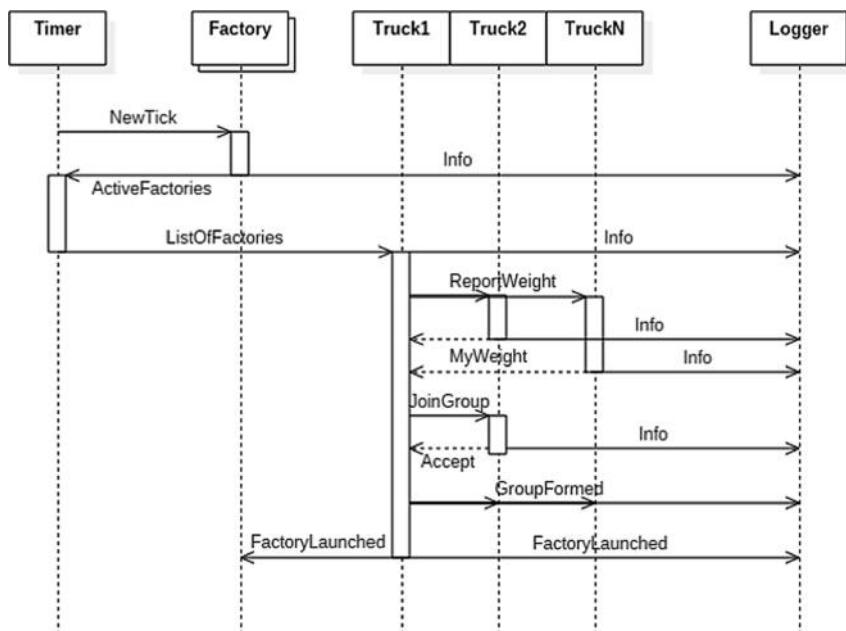


Fig. 3. Sequence diagram for group formation process

Group formation algorithm consists of several steps.

1. Timer agent informs all the agents about a new iteration. Factory agents (F_{act}) send their coordinates and amount of materials needed in reply. Timer agent resends list of all the unstarted factories to truck agents (T_{act}), which carry materials. Each of the truck agents sends to other agents from T_{act} amount of materials and number of steps to arrive to each factory agent from F_{act} .
2. Truck agent with maximum load of materials is denoted as leader. Leader agent finds factory agent that can be accessed with minimum of steps (N). Leader truck agent selects from T_{act} all truck agents T_{act}^N , which can access the chosen factory agent in N or less steps. If total load of materials in trucks from T_{act}^N is greater or

equal than selected factory agent demands, then leader truck agent sends the order to truck agents from T_{act}^N to move to selected factory.

If total load of materials in trucks from T_{act}^N is insufficient to start the selected factory operation, then leader increases the number of steps and finds a new subset of trucks, T_{act}^{N+1} . If total load of new subset of trucks T_{act}^{N+1} is sufficient, then leader agent orders them to move to selected factory. In case the total load of trucks agents from T_{act}^{N+1} is less than selected factory agent needs, the leader truck agent tries to start another factory agent, which can be accessed in more than N steps. Trucks from T_{act}^N which have not obtained orders, choose new leader agent and the process repeats.

3. When a group of trucks arrives to factory with a sufficient total load of materials, the factory agent changes its property to “started”. The possible left materials may be used to start another factory agents. The process stops when the total amount of materials is less than the smallest demand among the factory agents, or when all factory agents have started operation.

The dependence of number of messages sent and time used to form all the groups on different number of factory and truck agents are given in Table 1. Factories agents' position and materials demand and trucks agents' position and materials amount were generated randomly.

Table 1. Dependence of number of messages and time used on number of agents

Number of factories agents	Number of trucks agents	Number of messages	Time, s
5	20	25	0.2
	50	84	4.6
	100	117	105
10	20	40	1.5
	50	144	40
	100	237	230
20	20	42	5
	50	61	86
	100	378	307

The number of messages is relatively small due to SPADE's ability to send message to several recipients simultaneously. The simulation time for 20 factory agents and 100 truck agents was about 5 min. Results of experiments are satisfactory, and performance may be further improved by using distributed computations on several computers linked to local area network.

6 Conclusion

A greedy algorithm for group formation in multi-agent system was proposed. Algorithm was implemented using SPADE multi-agent system and its scalability and performance were studied: computational time for more than one hundred agents was about 5 min using personal computer with average characteristics. Results show that performance of proposed SPADE based framework allows the implementation of more complex algorithms for agents group formation, which may provide results closer to optimal.

References

1. Bellifemine, F., Poggi, A., Rimassa, G.: JADE—A FIPA-compliant agent framework. In: Proceedings of The Practical Applications on Intelligent Agents and MultiAgent Technology (PAAM), pp. 97–108 (1999)
2. Gregori, M.E., Cámara, J.P., Bada, G.A.: A jabber-based multi-agent system platform. In: Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 1282–1284, ACM (2006)
3. Amato, A., Di Martino, B., Scialdone, M., Venticinque, S.: Design and evaluation of P2P overlays for energy negotiation in smart micro-grid. *Comput. Stand. Interfaces* **44**, 159–168 (2016)
4. Van De Vijsel, M., Anderson, J.: Coalition formation in multi-agent systems under real-world conditions. In: Proceedings of Association for the Advancement of Artificial Intelligence (2004)
5. Farinelli, A., Bicego, M., Ramchurn, S.D., Zucchelli, M.: C-Link: A hierarchical clustering approach to large-scale near-optimal coalition formation. In: IJCAI, pp. 106–112 (2013)
6. Rahwan, T., Ramchurn, S.D., Jennings, N.R., Giovannucci, A.: An anytime algorithm for optimal coalition structure generation. *J. Artif. Intell. Res.* **34**, 521–567 (2009)
7. Pawłowski, K., Kurach, K., Svensson, K., Ramchurn, S.D., Michalak, T.P., Rahwan, T.: Coalition structure generation with the graphics processing unit. In: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems, pp. 293–300 (2014)
8. Voice, T., Ramchurn, S.D., Jennings, N.R.: On coalition formation with sparse synergies. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, vol. 1, pp. 223–230. International Foundation for Autonomous Agents and Multiagent Systems (2012)
9. Di Mauro, N., Basile, T.M., Ferilli, S., Esposito, F.: Coalition structure generation with GRASP. In: International Conference on Artificial Intelligence: Methodology, Systems, and Applications, pp. 111–120. Springer, Berlin (2010)
10. Osipov, G.S., Panov, A.I., Chudova, N.V.: Behavior control as a function of consciousness. I. World model and goal setting. *J. Comput. Syst. Sci. Int.* **53**, 517–529 (2014)
11. Osipov, G.S., Panov, A.I., Chudova, N.V.: Behavior control as a function of consciousness. II. Synthesis of a behavior plan. *J. Comput. Syst. Sci. Int.* **54**, 882–896 (2015)
12. Such, J.M., Alberola, J.M., Mulet, L., Espinosa, A., Garcia-Fornes, A., Botti, V.: Large-scale multiagent platform benchmarks. In: LADS, pp. 192–204 (2007)



On Compressed Sensing Based Iterative Channel Estimator for UWA OFDM Systems

Sumit Chakravarty^{1(✉)} and Ankita Pramanik²

¹ Kennesaw State University, 30060 Kennesaw, GA, USA

schakra2@kennesaw.edu

² IIEST, Shibpur, W.B, India

pramanikankital@gmail.com

Abstract. The ever-increasing demand for high-data-rate communication over a wireless multipath fading channel usually necessitates that at the receiver, prior knowledge about the channel is known. This is often achieved using known pilot signals that track the channel and produces at the receiver channel impulse response reconstruction obtained from the received signals. Recently, empirical studies have demonstrated that rich multipath channel assumption is violated in most physical systems and that the channel instead exhibits a sparse multipath behavior that is characterized by only a few dominant paths in propagation. In past decades, there has been a growing interest in the discussion and study of using underwater acoustic channel as the physical layer for communication systems. In this work, Compressed Sensing (CS)-based iterative channel estimators for Underwater Acoustic (UWA) OFDM systems are proposed where channel is assumed to be both sparse and time varying. The estimation of UWA channel is mainly based on Kalman filtered Compressed Sensing (KFCS) algorithms. CS with Kalman filter (KF) provides new idea about channel estimation for UWA OFDM communication systems, whose result outweigh traditional CS-UWA results.

Keywords: Underwater acoustic communications · Compressed sensing · Kalman filtering · Iterative channel estimation

1 Introduction

Channel State Information (CSI) is essential for coherent communication over multi-antenna Multi Input Multi Output (MIMO) channels. MIMO is a ubiquitous technology, offering large data rate transmission and is fading resilient. There are various MIMO configurations. MIMO technology improves data throughput, spectrum efficiency, coverage area and link reliability and diversity of wireless systems. Because of these, MIMO is used in various wireless radio communications standard, namely Wi-Fi, WiMAX (4G) and LTE (4G). MIMO systems do not inherently use CSI. Due to lack of CSI, the amplitude or phase of the reconstructed signal can vary more widely than the actual value. The performance of the receiver thus gets degraded. For good signal reconstruction, the CSI should be fed back to the transmitter and estimated at receiver. The CSI can be obtained from various algorithms of Channel Estimation

(CE). Almost all existing training-based channel estimation methods in the literature are based on the assumption of a rich underlying multipath environment; the numbers of degrees of freedom in the MIMO channel are assumed to scale linearly with the signal space dimensions [1]. However, in practice, the Channel Impulse Response (CIRs) of MIMO channel is actually dominated by a relatively small number of dominant resolvable paths. This is especially true with large bandwidth, long signaling duration, or large number of antennas. Traditional CE methods lead to overutilization of the key communication resources of energy and bandwidth in sparse MIMO channels. Because of this sparsity in the multi-path signals, CS can be used to improve the performance of CE in MIMO systems [2].

In order to formulate the CE problem as a CS problem, the sparse channel (Sparse Channel) model is the key ingredient and prior researches pay much attention to the channel modeling research [2, 3]. However, they rarely take the time varying nature of the wireless channel into account, which makes most of the sparse channel models do not adequately reflect the dynamic environment. For instance, the number of propagation paths and the path delays are often modeled as fixed, although the transmitter or the receiver can be moving.

Channel estimators based on such static channel model are no longer suitable in the dynamic channel environment, especially for the UWA and high-speed mobile communication systems [4]. In this work, both the sparseness and the time varying nature of the UWA channel are considered into account. The channel parameters, including the path delays and the path gains, are assumed to vary over time. In CS theory, this type of signal is called the dynamic sparse signal. Experiment results show the efficacy of this approach, which are presented in Sect. 3 along with details of our dynamic CS model. In the following section we provide brief introduction to the core concepts utilized in our work, namely, Compressed Sensing, Kalman Filtering and Underwater Acoustic communications.

2 Background Concepts

2.1 Compressed Sensing (CS)

In a traditional acquisition system, all samples of the original signal are acquired. This number of signal samples can be in the order of millions. Hence, traditional acquisition systems first acquire a huge amount of data, a significant portion of which is immediately discarded. This creates an important inefficiency in many practical applications. CS addresses this inefficiency by effectively combining the acquisition and compression processes. Traditional decoding is replaced by recovery algorithms that exploit the underlying structure of the data [5]. CS has become a very active research area in recent years due to its interesting theoretical nature and its practical utility in a wide range of applications. Figure 1 provides the framework for compressed sensing wherein matrix Φ is used to sample data x to generate output y . Considering the data to be inherently sparse, the output can be elaborated as

$$y = \Phi \Psi \alpha \\ \text{where, } x = \Psi \alpha \quad (1)$$

2.2 Kalman Filter (KF)

The KF is widely applied concept in the time series analysis used in the fields such as signal processing and econometrics. The traditional KF has also been employed for recovery of sparse signals from noise observations. Recent works utilize notions from the theory of CS such as Restricted Isometric Property (RIP) and related probabilistic recovery arguments for sequentially estimating sparse state in intrinsically low-dimensional systems. The Kalman Filter uses the state Eq. (2) and measurement Eq. (3) to provide a time varying prediction of system state, which in our case corresponds to the underwater acoustic channel state information. Here $x(t)$ and $y(t)$ correspond to state variable and measurement variable at time t respectively. The matrix A and C are state and measurement matrices respectively while $w(t)$ and $v(t)$ correspond to state and measurement noise.

$$x(t+1) = Ax(t) + w(t) \quad (2)$$

$$y(t) = Cx(t) + v(t) \quad (3)$$

2.3 Underwater Acoustic Communications (UAC)

UWA channels are considered to be “quite possibly nature’s most unforgiving wireless medium” [6]. The complexity of UWA channels is dominated by the ocean environment characteristics which include significant delay, Double-side-spreading, Doppler-spreads, frequency-selective fading, and limited bandwidth [7]. However, current acoustic communication technologies can only provide limited data rates due to the particular physical features of the channel [8]. The properties of the underwater

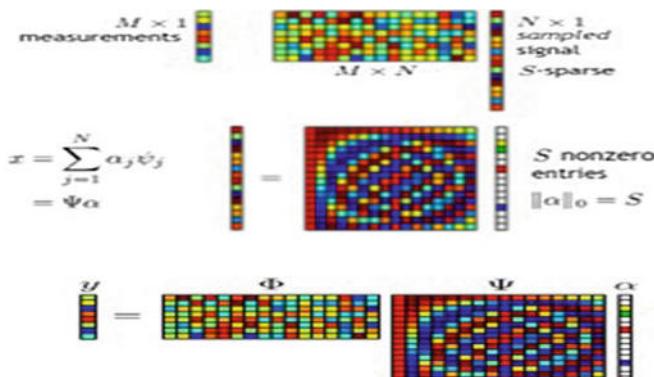


Fig. 1. CS signal acquisition and reconstruction model

medium are also extremely varied, and change in both space and time. Fluctuations due to environmental characteristics include seasonal changes, geographical variations both in temperature and salinity, seabed relief, currents, tides, internal waves, movement of the acoustic systems and their targets, etc. All this makes the UWA signal fluctuate randomly. Research has been active for over a decade on designing the methods for wireless information transmission underwater. The current trend is to use channel state aware-MIMO communications to offset the low data rate and other channel deficiencies mentioned above.

3 CS Based Iterative UWA Channel Estimation

Consider a UWA channel, whose CIR length after sampling is L,

$$\begin{aligned}
 y &= \mathbf{Ah} + n \\
 \text{where, } \mathbf{A} &= \mathbf{XF} \\
 \text{and} \\
 \mathbf{X} &= \text{diag}(X(k_1), \dots, X(k_{Np})) \\
 \mathbf{F}_{Np \times L} &\subseteq [k_1, \dots, k_{Np}] \\
 \mathbf{h} &= [h(0), \dots, h(L-1)]^T \\
 \eta &\sim CN(0, \sigma_\eta^2 I_{Np})
 \end{aligned} \tag{4}$$

The transmitted pilots and the received OFDM pilots are denoted as $X(k_1), \dots, X(k_{Np})$ and $Y(k_1), \dots, Y(k_{Np})$. Since the system sampling interval is much smaller compared to the channel delay spread, most channel coefficients are either zero or nearly zero, which means that \mathbf{h} is a sparse vector. Consider $\mathbf{h} \in \mathbb{R}^L$ has S non-zero components, with $S \ll L$. If A has more rows than columns ($Np > L$), then Eq. (4) is a standard LS problem. However, in the sparse case the number of pilots is smaller than the number of channel coefficients ($Np < L$), therefore Eq. (4) can be formulated as mathematical model of compressed sensing. Compressed sensing need only $n = O(S \log(L/S))$ measurement points to recover compressible signal with sparse degree S by linear optimization or greedy algorithm. Conventional CE algorithm usually assumes the CIR is time invariant, but the actual environment is not the same, especially for channel estimation algorithm with the application of ML, the computation of the process takes too long, the channel real information has changed. In this work the new CS-based iterative channel estimator based on KFCS is proposed for the case of time-varying UWA OFDM system. The actual channel states can be described by the first order Auto Regressive (AR) model, CIR at different interval of time, can be expressed as:

$$h_k = h_{k-1} + w_{k-1} \tag{5}$$

where, w_{k-1} is the process noise. A brief overview of KFCS is as below.

3.1 Kalman Filtering Compressed Sensing

KFCS is a new sparse signal processing method proposed by N. Vaswani in 2008 [9]. This is different from conventional compressed sensing which addresses the single spatial signal; KFCS considers the problem of reconstructing time sequences of spatially sparse signals from a limited number of linear incoherent measurements. In this method, the sparsity pattern is assumed to be unknown and change slowly with time. The overall idea of the solution is to use compressed sensing to estimate the support set (a position index collection of the nonzero elements; in channel estimation, it means the channel path delays) of the initial signals transform vector. At future times, run a reduced order KF with the currently estimated support and estimate new additions to the support set by applying CS to Kalman innovations or filtering error [10].

3.2 KFCS—Iterative UWA Channel Estimation

CS-based iterative channel estimators for time-varying UWA OFDM systems is proposed, which employs KFCS as signal processing method. Firstly, the time-varying channel is modeled as an AR process as described in Eqs. (4) and (5). Here, the CE is formulated as an iterative process. During the iteration, the path delays are estimated through a simple CS reconstruction algorithm. With the estimated path delays, the KF is performed to obtain the Minimum Mean Square Error (MMSE) estimation of the CIR. Through the iteration, the CE accuracy is improved. Hence CIR is also improved.

The KFCS iterative equation for UWA channel estimation can be expressed as:

$$\begin{aligned}\hat{\mathbf{h}}_{k|k-1} &= \hat{\mathbf{h}}_{k-1} \\ \mathbf{P}_{k|k-1} &= \mathbf{P}_{k-1} + \mathbf{Q}_1 \\ \hat{\mathbf{h}}_{k|k-1} &= \hat{\mathbf{h}}_{k-1} + \mathbf{K}_k(\mathbf{r}_k - \mathbf{C}\hat{\mathbf{h}}_{k-1}) \\ \mathbf{K}_k &= \mathbf{P}_{k|k-1}\mathbf{C}^H(\mathbf{C}\mathbf{P}_{k|k-1}\mathbf{C}^H + \mathbf{R})^{-1} \\ \mathbf{P}_k &= (\mathbf{I} - \mathbf{K}_k\mathbf{C})\mathbf{P}_{k|k-1}\end{aligned}\quad (6)$$

where, subscript $k|k-1$ denotes the one-step prediction from $k-1$ th symbol to k th symbol. \mathbf{h}_k is the state vector and $\hat{\mathbf{h}}_k$ is its prediction. \mathbf{P}_k stands for the error covariance matrix

$$\mathbf{P}_k = E[(\mathbf{h}_k - \hat{\mathbf{h}}_k)(\mathbf{h}_k - \hat{\mathbf{h}}_k)^H] \quad (7)$$

\mathbf{K}_k represents the Kalman gain matrix. KFCS includes two parts: KF and sparse bases search. First, CS algorithm is used to search the set of signal sparse bases \mathbf{h}_{k-1}^T . The KF process is applied on the \mathbf{h}_{k-1}^T support set, which mean a reduced order iterative. The filtering error is defined as

$$\tilde{\mathbf{y}}_k = \mathbf{y}_k - \mathbf{A}\hat{\mathbf{h}}_{k|k-1}^T \quad (8)$$

It is assumed that the filtering error contains the information which indicates the possible channel path delays. FEN is calculated for detecting new bases of \mathbf{h}_{k-1}^T . FEN is computed as

$$\begin{aligned} FEN &= \tilde{\mathbf{y}}_k' \sum_{fc}^{-1} \tilde{\mathbf{y}}_k \\ \text{where } \sum_{fe} &= [\mathbf{I} - \mathbf{A}_{\mathbf{h}_k^T} \mathbf{K}_k] \sum_{ie} [\mathbf{I} - \mathbf{A}_{\mathbf{h}_k^T} \mathbf{K}_k]' \\ \text{and } \sum_{ie} &= (\mathbf{C} \mathbf{P}_{k|k-1} \mathbf{C}^H + \mathbf{R}) \end{aligned} \quad (9)$$

The value of FEN is compared with threshold value. If the FEN is larger than the threshold, DS algorithm is performed on the filtering error for new set of the iterative. Another potential L path delays and the new support set $\tilde{\mathbf{h}}_k^T$ is obtained. The new support set is formed by combining \mathbf{h}_{k-1}^T and $\tilde{\mathbf{h}}_k^T$ as

$$\mathbf{h}_{k+1}^T = \mathbf{h}_{k-1}^T \cup \tilde{\mathbf{h}}_k^T \quad (10)$$

Bases which remains near-zero or nearly-constant for significant times are deleted from the support set. The KFCS algorithm is summarized in Fig. 2 and its implementation in UWA is present in Fig. 3.

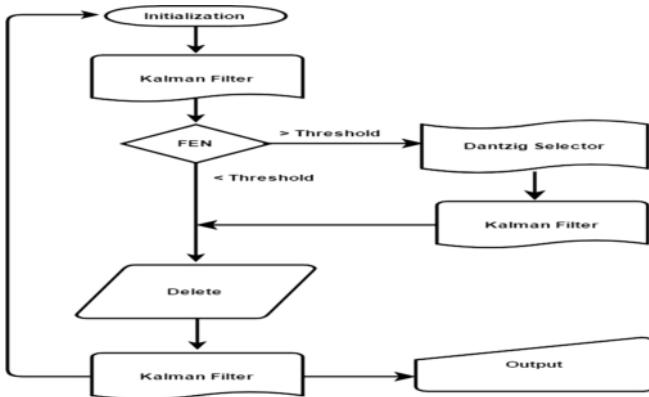


Fig. 2. Flowchart of KFCS algorithm

3.3 Results

Simulation is performed to investigate the performance of the proposed CS-based channel estimator. In order to evaluate the performance, MSE is calculated to quantize the CE errors.

In the simulation, an UWA OFDM system with total $N_d = 528$ subcarriers, among which $N_p = 20$ and $N_u = 110$ are used for pilot and null subcarriers, respectively. They are coded using a (1056, 528) binary low LDPC code. The length of guard interval is $N_g = 64$. The bandwidth $B = 4$ kHz is centered on the carrier frequency

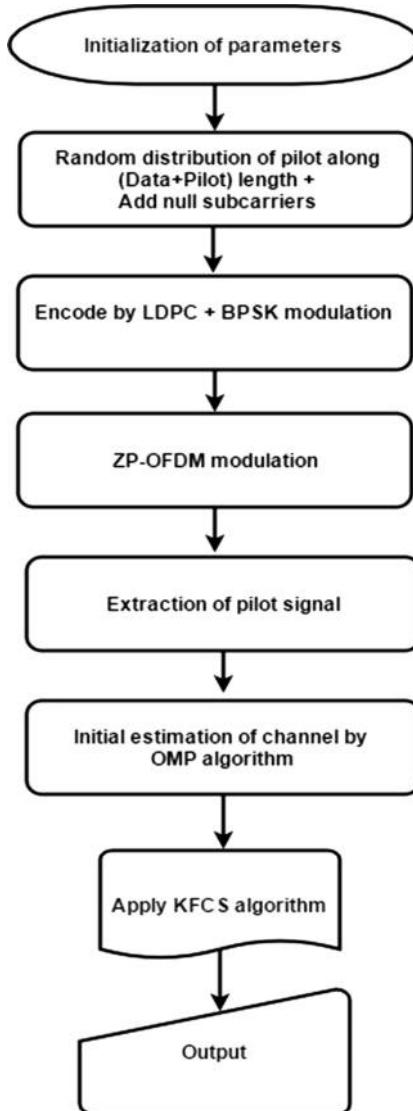


Fig. 3. Flowchart of iterative KFCS-Iterative UWA channel algorithm

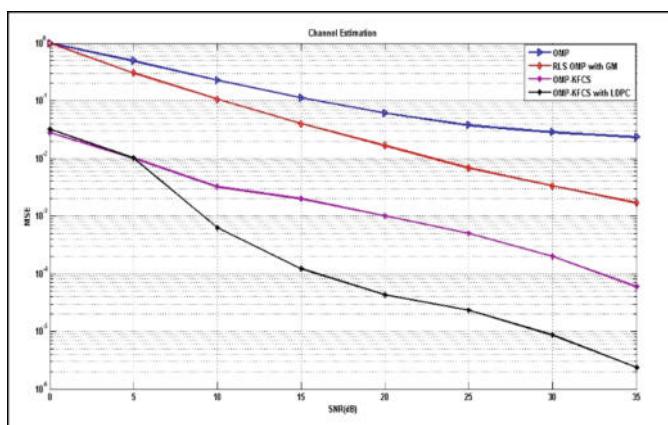
$f_c = 24$ Hz and divided into OFDM subcarriers. A five-path ($S = 5$) channel with the maximum channel delay spread $L = 50$ is considered. Pilot locations in frequency domain are randomly determined. These parameters are assumed to be constant within an OFDM symbol. Details of the simulation parameters are summarized in Table 1.

We compared the MSE performance of propose KFCS-based iterative channel estimator with and without LDPC with the conventional CS-based channel estimator Orthogonal Matching Pursuit (OMP) and Recursive Least Square (RLS) OMP with Gauss

Table 1. Simulation parameters for UWA ZP-OFDM system

Parameters	Values
Number of total subcarriers	$N_d = 528$
Number of null subcarriers	$N_u = 110$
Number of pilot subcarriers	$N_p = 20$
Pilots position	Randomly
Length of zero padding	$NG = 64$
Channel length	$L = 50$
Number of multipaths	$S = 5$
Carrier frequency	$f_c = 24 \text{ kHz}$
Signal bandwidth	$B = 4 \text{ kHz}$
Modulation	BPSK
Channel estimation methods	Conventional CS-based method, proposed method
SNR (dB)	0–35

Markov (GM) proposed in [11]. The MSE performance of these algorithms is shown in Fig. 4. Respectively, the OMP-KFCS represents the case where OMP is used for initial estimate of path delays then the standard KFCS is applied to estimate the CSI, and OMP-KFCS with LDPC represents the proposed algorithm with LDPC, where LDPC is used for further improvement in estimation performance of UWA OFDM system. From the figure we see that with the increase in SNR, the MSE of the channel estimators are all gradually reduced. On the one hand, among the all channel estimators, two kinds of KFCS-based channel estimators i.e. OMP-KFCS with and without LDPC outperform the conventional CS-based channel estimator OMP and RLS OMP with GM proposed in [11]. This is because the KFCS-based channel estimators utilize the time-domain correlation of the wireless channel. Hence, the computational complexity of proposed method using OMP algorithm with KFCS is much lower than the RLS OMP with GM method.

**Fig. 4.** MSE performance comparison for UWA ZP-OFDM system

4 Conclusion

CS recovery algorithms OMP are used here to obtain a preliminary estimate of the channel. With this channel estimation, KFCS is performed which gives much better performance than conventional CS based channel estimator. Compared with conventional CS-based channel estimators which perform CS at each time separately, the proposed method here takes the time-varying of the channel parameters into account and improves the accuracy of estimation algorithm and reduces the required sampling points and brings superior performance.

Also, LDPC code is introduced to adopt the UWA channels which suffer from severe ISI produced by long multi-path delay spreading and bandwidth extremely limited by acoustic propagation loss. This solution provides a significant improvement in the performance of iteratively estimation of channel.

The simulation results compare the CS-based iterative channel estimator with the existing conventional CS-based algorithm. Simulation results show that the proposed channel estimators outperform the conventional CS-based channel estimator and other existing algorithms in terms of MSE performance.

References

1. Bajwa, W.U., Sayeed, A., Nowak, R.: Compressed sensing of wireless channels in time, frequency, and space. In: 42nd Asilomar Conference on Signals, Systems and Computers, pp. 2048–2052 (2008)
2. Bajwa, W.U., Haupt, J., Sayeed, A.M., Nowak, R.: Compressed channel sensing: a new approach to estimating sparse multipath channels. Proc. IEEE **98**(6), 1058–1076 (2010)
3. Berger, C.R., Wang, Z., Huang, J., Zhou, S.: Application of compressive sensing to sparse channel estimation. IEEE Commun. Mag. **48**(11), 164–174 (2010)
4. Jakobsen, M.L., Laugesen, K., Manchn, C.N.: Parametric modeling and pilot-aided estimation of the wireless multipath channel in OFDM systems. In: Proceeding of IEEE International Conference on Communications (ICC), pp. 1–6 (2010)
5. Candes, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies. IEEE Trans. Inf. Theory **54**(6), 5406–5425 (2006)
6. Brady, D., Preisig, J.C.: Underwater acoustic communications. Wireless Commun. Signal Process. Perspect. **8**, 330–379 (1998)
7. Singer, A.C., Nelson, J.K., Kozat, S.S.: Signal processing for underwater acoustic communications. IEEE Commun. Mag. **47**(1), 90–96 (2009)
8. Iglesias, I., Song, A., Garcia-Frias, J., Badiey, M., Arce, G.R.: Image transmission over the underwater acoustic channel via compressive sensing. In: 45th Annual Conference on Information Sciences and Systems, pp. 1–6 (2011)
9. Candes, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(11), 4203–4215 (2005)
10. Chen, B., Cui, Q., Yang, F., Xu, J.: A novel channel estimation method based on Kalman filter compressed sensing for time-varying OFDM system. In: 2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–5. Hefei (2014)
11. Qi, C., Wu, L., Wang, X.: Underwater acoustic channel estimation via complex Homotopy. In: IEEE International Conference on Communications (ICC), pp. 3821–3825 (2012)



Development on Interactive Route Guidance Robot for the Mobility Handicapped in Railway Station

Tae-Hyung Lee^(✉), Jong-Gyu Hwang, Kyeong-Hee Kim,
and Tae-Ki Ahn

Korea Railroad Research Institute, Uiwang, Republic of Korea
thlee@krrri.re.kr

Abstract. In this paper, the functions of interactive route guidance robot were described. That robot will be guided and serviced by the mobility handicapped from the time of arriving to the exit for convenience when the mobility handicapped was moving within railway station for train boarding, transfer, and facilities using.

Keywords: Service robot · Mobility handicapped · Railway station

1 Introduction

Under the Mobility Handicapped Convenience Movement Promotion Act of Korea, the mobility handicapped has the right to use all vehicles safely and comfortably without discrimination. Through this, Korea Government is presenting a vision of implementing a transportation welfare society [1].

But, current status data for 2011 and 2016 for urban railroads and subway, general trains show that the average installation rate of the mobility handicapped facilities is 92%, but, the level of user satisfaction is less than 68%. In particular, the mobility handicapped has many difficulties in using railway station. In addition, for 2017–2021 years, government policy is a plan focused on installing facilities, and the national R&D task focuses on improving user satisfaction with the construction of facilities [2].

Accordingly, it is necessary to develop the interactive route guidance and supporting system for the mobility handicapped in railway station. In this paper, the functions of interactive route guidance robot (CoBoT, collaborative robot) were described. That robot will be guided and serviced by the mobility handicapped from the time of arriving to the exit for convenience when the mobility handicapped was moving within railway station for train boarding, transfer, and facilities using.

2 Operation Scenario of Interactive Robot Guidance Robot

2.1 Classification of the Mobility Handicapped

The mobility handicapped is classified as disabled, elderly, pregnant women, person with infant, and children. Disabled person are classified as physical disability, brain lesions, sight disability, hearing disability, etc.

Physical disability, sight disability and brain lesions move by CoBoT's handle or by CoBoT. Sight disability is provided with CoBoT's voice while hearing disability is provided with the screen.

2.2 Basic Operation Scenario

The basic operation scenarios for CoBoT are shown in Fig. 1. When the mobility handicapped arrives in railway station or requests guidance services from CoBoT, CoBoT approaches the mobility handicapped. When the mobility handicapped orders the guidance service to start, it automatically moves with the mobility handicapped in the optimal way to the point of interest. When the destination is reached, the mission is terminated and moved to the standby position. During the journey to a standby position, it can be moved to perform additional tasks.

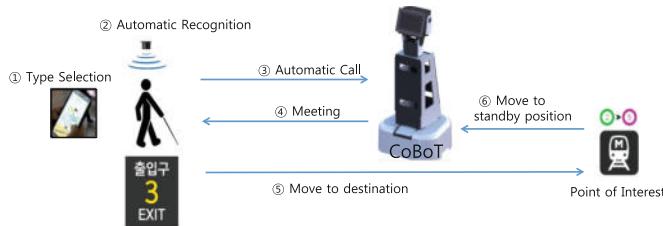


Fig. 1. Basic operation scenario

Here, the points of interest refer to boarding and transfer positions, elevators, wheelchair lifts, disabled toilets or general toilets, nursing rooms, ticket offices and automatic ticketing machines.

2.3 Emergency Scenario

The emergency scenarios for the mobility handicapped are shown in Fig. 2. In case of emergency situations, such as fire, the control center checks the location of the mobility handicapped. The control center transmits the location of the mobility handicapped to CoBoT near the mobility handicapped. The control center orders CoBoT to move, and CoBoT move to the mobility handicapped and recognize. CoBoT guides the mobility handicapped to the exit according to evacuation plan.

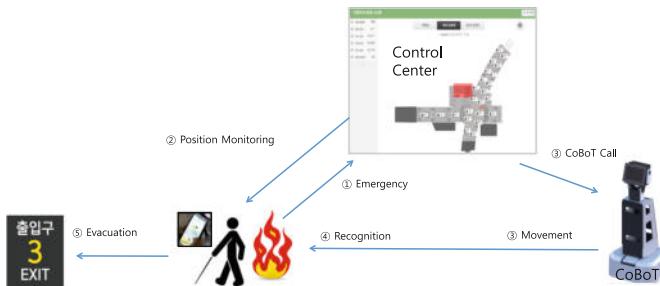


Fig. 2. Emergency scenario

3 Major Functions of Interactive Route Guidance Robot

3.1 System Structure

In order to guide the mobility handicapped through the route, the system consists of three parts as shown in Fig. 3. The first is the Control Center (CC), to control the whole system. The second is the mobile application used by the mobility handicapped, and the other is the CoBoT, which actually guides the mobility handicapped.

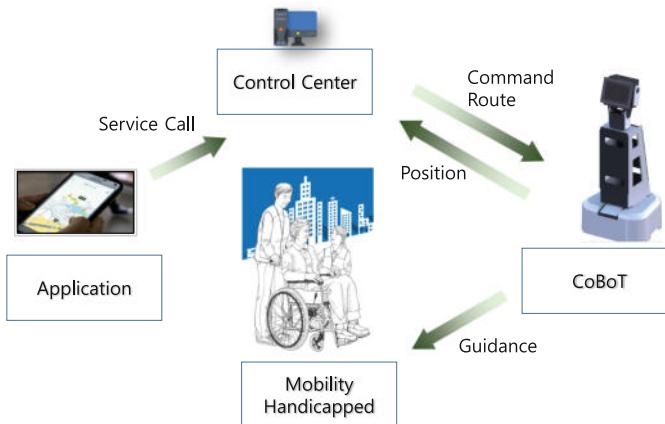


Fig. 3. System structure

When describing the subsystem, CC sends a service command to CoBoT for starting and closing the service and provides the optimal route. CC perform monitoring function by receiving position information from CoBoT. CoBoT guides the mobility handicapped from the original position to the destination position according to the instructions from CC, and sends the current position to CC. CoBoT must always keep proper distance from the mobility handicapped and have anti-crash capability.

3.2 Autonomous Driving

A map is needed for the autonomous driving of robots. A map editor was developed to generate maps. It is described in detail below.

- The map editor uses the railway station floor plan as shown in Fig. 4a.

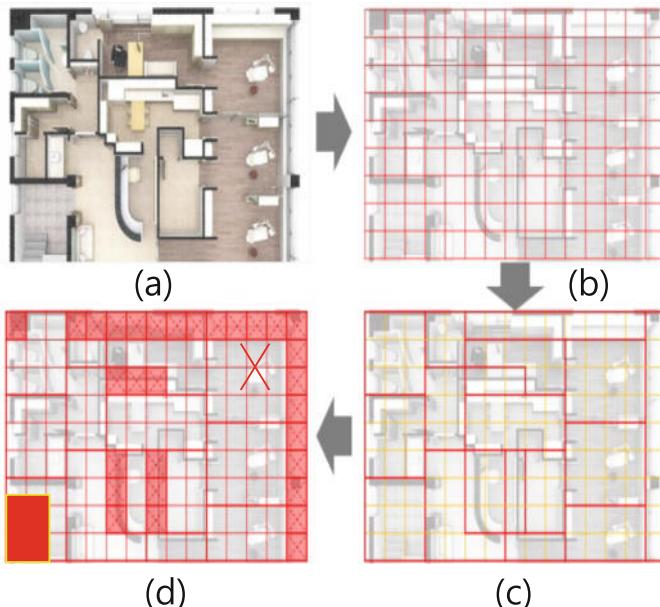


Fig. 4. Map editing procedure

- The map editor divides the railway station floor plan into horizontal and vertical grids as shown in Fig. 4b.
- Then set up an accessible area and the remainder of the space then becomes inaccessible as shown in Fig. 4c. It utilizes fixed structure information, such as walls.
- However, even in an accessible area, if a structure is installed after construction, access becomes impossible.
- The above information is collected using sensors such as LiDAR and transmitted to the control center as the robot moves through the railway station.

Combining the results completes the final map as shown in Fig. 4d. The area marked in red is inaccessible.

Even though the map is complete, the robot must know its location on the map to reach its destination. For this purpose, the map is shared as shown in Fig. 5. Maps are managed by the control center.

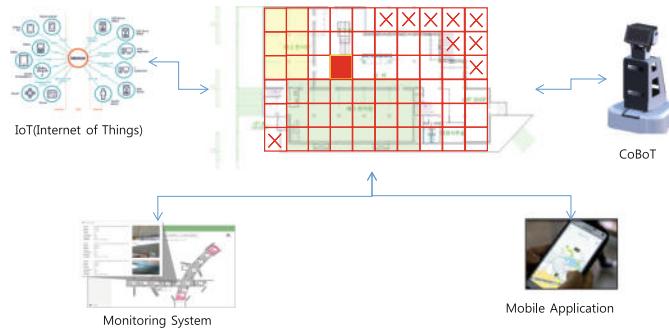


Fig. 5. Map utilization

- The control center transmits the optimal path to the CoBoT based on map information and its location. The CoBoT transmit spatial information acquired using sensors to the control center.
- Information on IoT installed in railway history is recorded and managed based on maps. In other words, IoT that is installed in front of toilet has basic location information.

Systems that manage the operation of support systems are also operating on a mapping basis. Data such as location of CoBoTs on this map, location and status of IoT sensors, and location of emergency situations will be displayed on this map.

3.3 Wireless Charging

The operation time of CoBoT is about two hours. Charging shall be performed with minimum time after completion of operation. CoBoT's guidance service has the significance of supporting the insufficient service personnel and should minimize human intervention. Therefore, it is necessary to automatically charge CoBoT. If charging is required, it should move itself to the charging position and perform charging without human assistance. To this end, it is planning to use wireless charging.

3.4 Intelligent Braille Block

In the previous section, we mentioned the application of RFID to Braille blocks to correct CoBoT position error. The RFID sensor is mounted inside the Braille block and is designed to allow CoBoT to read information in close proximity to correct its current location.

In addition, BLE (Bluetooth Low Energy) is also mounted inside the Braille block. BLE communicates with the phone's applications that the mobility handicapped carries. The phone's applications receive Braille block's information, such as location, and point of interest, and transmit to the control center.

4 Conclusion

The interactive route guidance and supporting system for the mobility handicapped in railway station is introduced. The functions of interactive route guidance robot (CoBoT) were described. CoBoT will be guided and serviced by the mobility handicapped from the time of arriving to the exit for convenience when the mobility handicapped was moving within railway station for train boarding, transfer, and facilities using.

Major functions of CoBoT were derived according to the scenario in which they operate. Major functions were allocated to subsystems such as CoBoT, control center, and mobile apps. CoBoT's major functions are autonomous driving, keeping a constant distance with the mobility handicapped and wireless charging. We introduced intelligent Braille blocks.

References

1. Ministry of Land, Infrastructure, and Transport: Act on Promotion of the Transportation Convenience of Mobility Disadvantaged Persons
2. Hwang, J.K., et al.: Planning Report, Development on Interactive Route Guidance and Supporting System Technology for the Mobility Handicapped in Railway Station. KRRI (2018)



Facilitate External Sorting for Large-Scale Storage on Shingled Magnetic Recording Drives

Yu-Pei Liang¹(✉), Min-Hong Shen¹, Yi-Han Lien²,
and Wei-Kuan Shih¹

¹ Department of Computer Science, National Tsing Hua University,
Hsinchu, Taiwan

tychen@saturn.yzu.edu.tw

² Department of Electronic Engineering, National Taipei
University of Technology, Taipei, Taiwan

Abstract. In the era of big data and cloud computing, both external data process techniques and new storage mediums are proposed to process and accommodate the sheer amount of information with data-intensive applications. For instance, external sorting algorithms perform sorting operations directly on the storage devices to lower the data transfer latency and increase system performance. On the other hand, Shingled Magnetic Recording (SMR) is proposed to increase the areal density by overlapping tracks. However, the overlapping technique also introduces the random-write restriction because writing a track also destroys the valid data on overlapped tracks. This constraint could induce significant write amplification issue when performing external sorting on SMR drives. To mitigate the write amplification issue, this paper proposes a sort-friendly SMR drive design to lower the write amount of external sorting algorithms on SMR drives. The experimental results show that the proposed design could lower the external sorting latency by 61.99% when compared with the external merge sort algorithm.

Keywords: Shingle magnetic recording · External sorting · Cloud computing

1 Introduction

Data-intensive applications aim to extract useful information from a large dataset. However, the size of the dataset continues to grow in both academic and industrial field, nowadays. As a result, both the process efficiency and the storage capacity become major issues for data-intensive applications. Due to the large amount of data within data-intensive applications, the challenges of the underlying storage medium include how to achieve short I/O latency and satisfy the capacity requirement. Recent studies introduce the concept of active disks for hosting external data process techniques. Thus, instead of seeking high-performance storage medium, the active disks process data within the storage device and output results only to lower the amount of

transferred data, thus improving the I/O latency. Furthermore, the active disks concept becomes promising as the computing power of storage devices becomes more powerful.

Take external sorting [1] as an example. Traditionally, external sorting can be used to sort a large-scale dataset on storage devices without loading the whole dataset into the system memory at once. As shown in Fig. 1, the external sorting can be classified into two phases, including partial sorting phase and merge phase. To sort data based on a specific index, traditional external sorting firstly divides data into chunks that can fit into the available system memory during the partial sorting phase. Then, each chunk is loaded into system memory and sorted with the chosen sorting algorithm (i.e., merge sort). At the end of the partial sorting phase, the partial sorted results of each chunk are written back to the storage devices. Then, the merge phase reads those partial sorted results and merges them into the final output. Finally, the sorted dataset is written back to the storage device. However, the process of traditional external sorting involves several read/write operations between the host and storage devices, thus inducing high I/O traffic. On the other hand, the active disks concept is expected to eliminate those I/O traffic by performing the sorting algorithm within the storage devices. Therefore, the active disks concept could be beneficial to external sorting since the sorting process are processed with the storage rather than the host system, thus reducing the I/O traffic [2, 3].

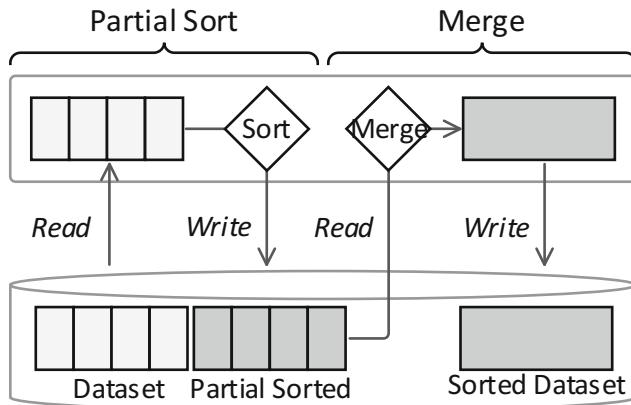
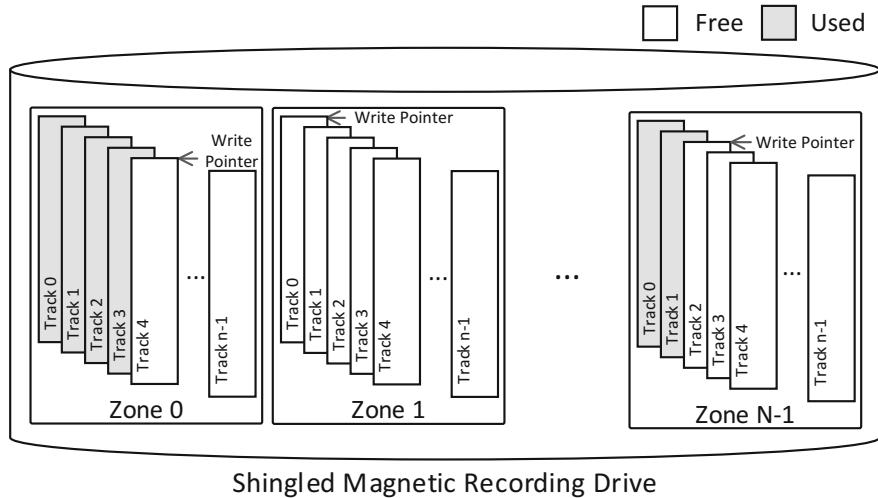


Fig. 1. The traditional external sorting.

On the other hand, in order to satisfy the capacity requirement, numerous new recording technologies [4] have been proposed, such as shingled magnetic recording (SMR), heat-assisted magnetic recording (HAMR), and bit-pattern magnetic recording (BPMR). Among these new technologies, SMR is regarded as one of the most promising candidates for providing large storage space with high data reliability and integrity. Unlike conventional hard disk drives (HDD), SMR overlaps adjacent tracks to achieve higher areal density, as shown in Fig. 2. As shown in Fig. 2, a set of



Shingled Magnetic Recording Drive

Fig. 2. The shingled magnetic recording (SMR) drive.

overlapped tracks are grouped together as a zone. Within each zone, a write pointer is maintained to indicate the next free tracks for accommodating new write traffic. However, the overlapped track layout imposes the random-write constraint on the incoming write requests because writing one track could also destroy valid data on adjacent tracks. Another issue of the random-write constraint is the write amplification issues because of the read-merge-write (RMW) operations while conducting random updates. This constraint could induce significant write amplification issue when performing external sorting on SMR drives under the active disk concept. A motivation example is given in Fig. 3 to illustrate the studied issue. Due to the overlapping track layout, any sorting operation will cause cascading update operation on adjacent tracks.

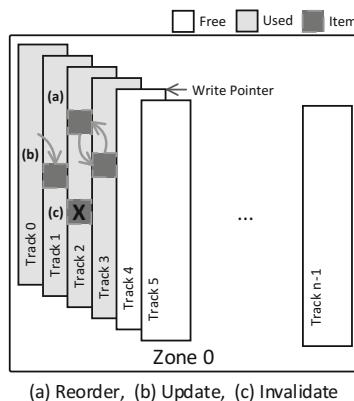


Fig. 3. The motivation example.

To mitigate the write amplification issue, this paper proposes a sort-friendly SMR drive design to lower the write amount of external sorting algorithms on SMR drives.

2 The Sort-Friendly SMR Drive Design

To lower the write amount of external sorting algorithms on SMR drives, this study proposes the sort-friendly SMR drive design to boost the performance of performing external sorting algorithms. Figure 4 shows the overview of sort-friendly SMR drive design. In the proposed design, a zone-based *B+ tree* is used to log the sorting index of the dataset. Then, each leaf node of the *B+ tree* is assigned with an SMR zone for logging the inserted data entry. When receiving insertion or update requests, the proposed design caches those requests with a small on-disk cache. With the proposed design, the number of read/write operations can be reduced to only one read and one write since logging strategy is utilized to accommodate data entries. Thus, the I/O traffic can be significantly reduced. Then, when the cache is full, those cached entries are flushed into the corresponding SMR zone at once. On the other hand, when the SMR zone is full, compact operations are issued to reclaim invalid entries.

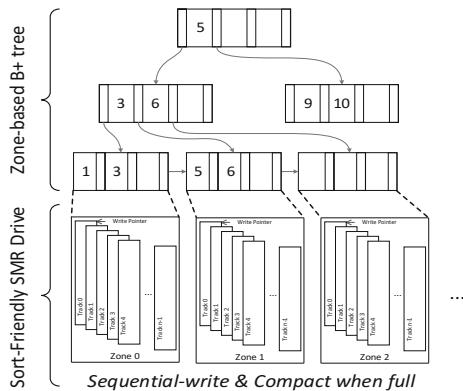


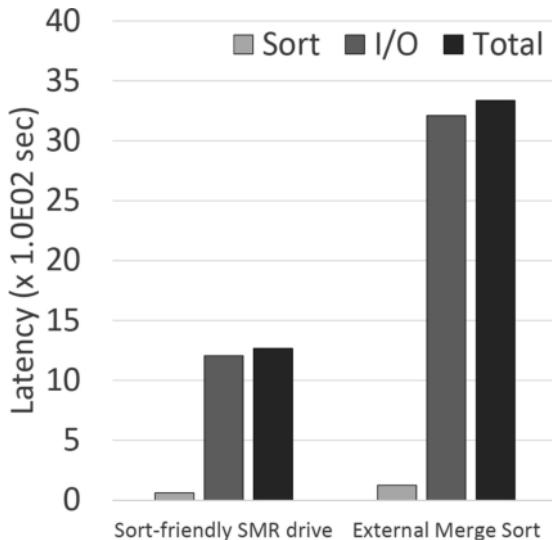
Fig. 4. The sort-friendly SMR drive.

3 Performance Evaluation

In this session, an in-house simulator was built to simulate the behavior of sorting algorithms and SMR disks. The proposed design is compared with traditional external merge sort algorithm. Table 1 summarized the I/O latency parameters of the studied SMR disks. In this evaluation, the size of the simulated SMR disks is set to 100 TB and the Yahoo! Cloud Serving Benchmark (YCSB) [5] workload is used for performance evaluation. During the evaluation, 1.9 million entries are inserted into SMR drive for sorting. Then, a sort command is issued to get the sorted dataset. Figure 5 shows that sorting performance can be improved by 61.99% according to the sort-friendly SMR drive design.

Table 1. Simulation parameters [6]

SMR data transfer speed	150 MB/s
SMR seek time	8.0 ms
SMR rotational speed	7200 rpm

**Fig. 5.** Latency comparison.

4 Conclusion

With the growing trend of data amount for data-intensive applications, both the external data process techniques systems and the active disks concept can SMR disks can be used to boost the performance of data-intensive applications. On the other hand, in order to accommodate the large amount of information, SMR drives is proposed to overlap tracks so as to increase areal density. However, directly deploying external data process techniques on SMR disks with the active disks concept could induce serious write amplification problem due to the random-write constraint. To resolve the observed issue, this study proposes a sort-friendly SMR drive design to lower the number of RMW operations. Preliminary experimental results show that the proposed design can improve the external sorting performance by 61.99%.

References

1. Knuth, D.: *The Art of Computer Programming*, vol. 3, 2nd edn. Addison-Wesley (1998)
2. Quero, L.C., Lee, Y.S., Kim, J.S.: Self-sorting SSD: producing sorted data inside active SSDs. In: Mass Storage Systems and Technologies (MSST), 2015 31st Symposium, pp. 1–7 (2015)

3. Lee, Y.-S., Quero, L.C., Kim, S.-H., Kim, J.-S., Maeng, S.: ActiveSort: efficient external sorting using active SSDs in the MapReduce framework. In: Future Generation Computer Systems (2016)
4. Shiroishi, Y., et al.: Future options for HDD storage. IEEE Trans. Magn. **45**(10), 3816–3822 (2009)
5. Yahoo: Yahoo! cloud serving benchmark @ONLINE. <https://github.com/brianfrankcooper/YCSB/wiki> (2015)
6. Seagate: Seagate archive hdd @ONLINE. <http://www.seagate.com/www-content/product-content/hdd-fam/seagate-archive-hdd/en-us/docs/archive-hdd-ds1834-5c-1508us.pdf> (2015)



LTE Geolocation Based on Measurement Reports and Timing Advance

Zaenab Shakir^(✉), Josko Zec, and Ivica Kostanic

Computer Engineering and Sciences, Florida Institute of Technology,
Melbourne, FL, USA
zshakir2015@my.fit.edu

Abstract. This paper investigates a new method for geolocating LTE cellular users. The method relies on the measurements reported which are time advance metric and Reference Signal Received Power (RSRP) together. The simple predictive model is used with the reported LTE measurement, time advance, and cell configuration to estimate the cell phone or user equipment (UE) position in flat dimension (latitude and longitude). The algorithm is assessed via comparing the estimated locations with actual positions readings in several cases distinguished by the number of cells that UE instantaneously reports. For validity of algorithm, the mean square error is determined. The algorithm results display how the accuracy of the UE coordinate rely on the number of reported cells.

Keywords: Geolocation · Time advance · LTE

1 Introduction

With the increase in the number of Long-Term Evolution (LTE) cellular subscribers, the demand for geolocating active users has increased. Most of the applications and services depend on accurate geolocation, such as the weather, transportation, cellular forensics, fleet management, etc. [1, 2]. These applications mostly rely on the Global Positioning Satellite (GPS) receivers embedded in every smartphone. Without explicit support from the GPS, various geolocation methods attempt to retrieve the position of LTE users for finding traffic hotspots and network maintenance as part of the call tracing process. Consequently, the accuracy of geolocation is highly demanded in cellular companies.

Since 1996, the FCC mandated regulatory requirements that all communication companies must deliver highly accurate estimated positions of UEs and these requirements were modified to its final arrangement in 1998 [3]. In addition, the Global Navigation Satellite System (GNSS), which includes the GPS, provides accurate location information for outdoor users. However, GPS limitations are weak signal levels, particularly for urban and indoor receivers, as well as the need for extra hardware and batteries [2, 4]. Therefore, a GPS-based solution is not applicable in all conditions.

Various techniques have been proposed to coordinate the mobile device based on the network-based method, which is still considered the default method without explicit use of the GPS. They use the data of network to localize the UE. The proposed

traditional techniques performance is affected mainly by the environment of the area. Common techniques are Enhanced Cell-ID (ECID) which the accuracy relies on the size of the cell and its coverage and angle of arrival (AOA) that bases on the direction by using multiple antenna on cellular base station [5, 6]. In addition, time-difference-of-arrival (TDOA) technique which takes advantage of the time of travelling signal between the cell and mobile device with some geometry to estimate the UE's location, but the multipath could affect the time [7–11]. Recently, academic researchers have been using the network parameter enhanced timing advance (TA). TA was used before with GSM, but it had poor accuracy because of the early technology has poor performance [4, 12]. Today, the TA with LTE cellular networks has gain attention because of the timing requirement that improves the resolution of the TA [12, 13].

The objective of this paper to utilize existing measurements of the LTE TA to limit the search area to retrieve the coordinates of LTE users on 2-dimensional plane. Our estimation algorithm is taken the advantage of TA parameter and single level which are reported by the measurement. By using our cost function as predictive model combined with reported measurements and utilizing the direction of the antenna from the information of base stations. Five scenarios are taken here in this paper for evaluation based on different number of cells that reported on the same time. Hence, by using the actual locations which are found with measurement to determine the root mean square error between our estimation and the actual locations GPS readings from the same measurement set are used to calculate the root mean square error between the actual GPS location and the algorithm estimated location.

The remaining of this paper is organized into three sections. Section 2 is for Time advance in LTE, Sect. 3 for LTE measurement parameters, Sect. 4 for algorithm description for LTE Geolocation, Sect. 5 for result description, and Sect. 6 for the conclusion.

2 Time Advance in LTE

TA is the offset in time between the beginning of received downlink and transmitted uplink subframes, as defined in the release 9 of the LTE cellular communications standard [6]. It ensures that the transmission from all UEs will be synchronized when they are received by the cell. UE that is far in distance from the cell measures larger propagation delay, which leads to larger transmission timing advance compared to a UE closer to the cell. Thus, timing advance primarily enables synchronized LTE operations while at the same time serves as estimate of the distance between the UE and the sector.

In Fig. 1, timing of the uplink and downlink frames is illustrated, and the time advance δ for starting uplink transmission is calculated as:

$$\sigma = N_{TA} \times T_s \text{ s} \quad (1)$$

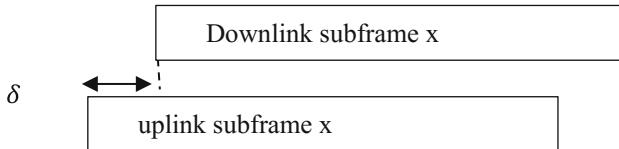


Fig. 1. Uplink downlink timing frame [13]

$0 \leq N_{TA} \leq 20,512$ which is the offset of the time that is required between the uplink and downlink subframe and give us maximum 20,512 TA units.

$$T_s = 1/(15,000 \times 2048) \text{ s} \quad (2)$$

In Eq. (2), each frequency subcarrier for LTE is 15 kHz, and for the maximum bandwidth 20 MHz the FTT size is 2048.

LTE cells determine the TA from the Physical Random Access Channel (PRACH) message, which is sent by a UE. Cells send the TA to the UE by the Random Access Response (RAR) message. TA messaging is specified in the LTE Media Access Control (MAC) protocol [13].

The MAC RAR shown in Fig. 2 contains six octets. Time advance command field contains 11 bits from first two octets and is represented by the index (0, 1, 2, ..., 1282). The TA needs to be adjusted relative to the changes in distance of the UE from the serving cell, and TA command MAC control element in octet 1 is responsible for this adjustment, as shown in Fig. 3. TA adjustment consists of 6 bits. Thus, the index of TA in this case ranges from 0 to 63.

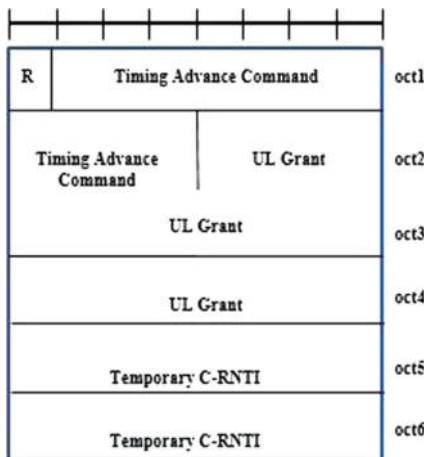


Fig. 2. MAC RAR [13]

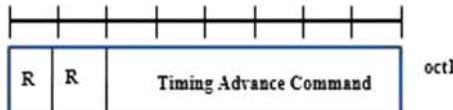


Fig. 3. TA adjustment MAC control element [13]

The TA command specifies the changes in uplink timing compared to the current timing as a multiple of $16T_s$. In the case of RAR, which contains 11 bits, the adjustment of time is expressed by $N_{TA} = TA \times 16$. Furthermore, in the case of MAC control element with 6 bits, the alignment of N_{TA} is given by

$N_{TA,new} = N_{TA,old} + (TA - 31) \times 16$. Therefore, the updated value of N_{TA} could be positive or negative value which specify advancing or delaying of timing for uplink transmission.

The calculation of the distance and time between UE and the serving cell relative to the TA unit could be illustrated by the following example: suppose the TA index has a maximum value of 1282. We applied this value in:

$$N_{TA} = TA \times 16 = 1282 \times 16 = 20,512 \quad (3)$$

Then, applied to Eq. (1), the maximum timing advance is:

$$N_{TA} = 20,512 \times (1/(15,000 \times 2048)) = 0.67 \text{ ms} \quad (4)$$

To obtain the distance per unit of TA, it could be verified by taking the reciprocal of the sampling frequency of the LTE, which is $1/3.84$ MHz, and multiply it by the speed of light, 3×10^8 m/s. Thus, TA units will be 78.125 m, and by multiplying the TA index by 78.125, the distance in meters will be determined between the serving sector and UE [13].

3 LTE Measurement Parameters

Drive testing is a common method to gather RF data in cellular communication. The measured data are utilized for coverage evaluations and optimization of networks. In this research, drive measurements will be used for geolocation validation. The area that has been chosen for validating the method proposed has been a city outdoor environment in Atlanta, GA and the total length of the measurement route has been 20 miles. The LTE measurements utilized in this research include received signal level and TA data measurements form the drive-test and the configuration of the cells from communication operators. GPS recordings from the drive test will be compared to algorithm retrievals to validate algorithm's accuracy. Data used in this study are summarized in Tables 1 and 2 [14, 15].

Table 1. Cell configuration [14, 15]

Cell configuration	Comment
Latitude/longitude	Latitude and longitude of each sector
Cell azimuth	Antenna centerline pointing angle measured clockwise from the North
Cell ERP	Effective radiated power for each sector
Cell PCI	PCI allocated to each cell

Table 2. LTE measured data [14, 15]

LTE measurement	Comment
Serving cell PCI	The serving cell is identified by its Physical Cell Identity (PCI). PCIs range between 0 and 503, allowing allocation of 504 unique PCIs before reuse becomes necessary
Serving cell frequency	LTE is deployed in several frequency bands with potentially multiple deployed carriers from the same band. This parameter indicates the carrier frequency of the serving cell
Serving cell RSRP	Received power measurements collected on the reference sequence of the serving cell. It is not subject to power control or load control and is continuously transmitted by every cell. It is the principal measurement type used in different LTE radio resource management procedures, including mobility management (handover and reselection) algorithms
Serving cell TA	Serving cell time advancement is a MAC level parameter. It is determined adaptively by the serving cell and it is provided to the UE through a MAC layer message. It is not visible from layer 3, and it may or may not exist in the measured data
Non-serving cell PCI	LTE measurement reports RSRP measurements from non-serving cells simultaneously with the serving cell. Non-serving cells are identified by their PCIs
Non-serving cell RSRP	RSRP measurements from non-serving cells coming in pair with a corresponding PCI

4 Algorithm Explanation for LTE Geolocation

The proposed UE location estimation method varies based on the number of simultaneously reported sites and sectors. Better geolocation accuracy is obtained with more reported sectors and sites. The algorithm can be divided into 9 main steps:

Step 1: Initialize the simulation setup.

Step 2: Divide the driven area to the bins and each bin size is 50 m.

Step 3: Pre-calculate the location polygons. When the location polygon is done, the area of polygon is divided into Voronoi based region for each sector. Consequently each bin will be allocated to one of the active cells regarding the strongest received signal level (*RSL*). One of the propagation models is utilized to determine the signal level, as in the following equations [14, 15]:

$$RSL = ERP[\text{dBm}] + f(\theta) - PL[\text{dB}] \quad (5)$$

$$PL = \overline{PL} + 10m \log(d/d_f) \quad (6)$$

When ERP is the effective radiated power for each site, PL is the pathloss, \overline{PL} is pathloss reference, d is the distance between the site and each bin, d_f is the reference distance, m is the path-loss exponent, and $f(\theta)$ is the function to describe the shape of the antenna pattern. On the sectorized antenna, the pattern is modeled as [14, 15]:

$$f(\theta) = \begin{cases} 10 \log_{10} \left(\cos \left(\frac{n\theta}{\pi} \right)^2 \right), & |\theta| \leq \frac{\pi}{3} \\ -20, & |\theta| > \frac{\pi}{3} \end{cases} \quad (7)$$

where n is the number of sectors on a site (typically 3) and θ is the difference between the azimuth of the cell and the direction to each geographical bin in assessment.

Step 4: Calculate the center point for each Voronoi region.

Step 5: Calculate the distance between the serving sector and the center point.

Step 6: Use the timing advance parameter (TA) to calculate the distance d between the serving sector and UE to fast the computation and reduce the search area of UE position by using the following equation:

$$d = TA * 78.125m \quad (8)$$

Step 7: Determine the distance \hat{d} , between the serving sector and UE, by using RSRP data measurement with signal level equation as follow [14, 15]:

$$\hat{d} = \min \left(d_0 10^{\left(\frac{RSL_o - RSRP}{m} \right)}, 2d_0 \right) \quad (9)$$

When d_0 is the distance between the center point of the sector region and serving cell. RSL_o is the reference of RSL and assume at $RSL_o = -80$ dBm. The slop of pathloss is $m = 45$ dB/dec. The distance should be not more than the double of the distance between the center point and the cell itself, therefore the minimum function is used.

Step 8: Calculate the cost function values for each bin under evaluation. The bin that has least value in cost function will be chosen as the estimated position for the UE with coordinates UE_{ey}, UE_{ex} . The generalized equation for determining the cost function is as follows:

$$G = \sum_{i=1}^N G_i^2 \quad (10)$$

N is the number of sectors and each individual cost is calculated by:

$$G_i = Ra \left\{ [\cos \emptyset_i - \cos \alpha_i]^2 + [\sin \emptyset_i - \sin \alpha_i]^2 \right\}^{\frac{1}{4}} \quad (11)$$

$R = 0.5$ km refers to the radius of serving cell, α_i is the azimuth of the sectors, and \emptyset_i is the look angle from sectors to the bins.

The final term in Eq. (10) is calculated as:

$$G_N = \sqrt{x^2 + y^2} - \hat{d} \quad (12)$$

When x, y are the bins coordinates in the limited area that are calculated from the TA as shown in Fig. 4.

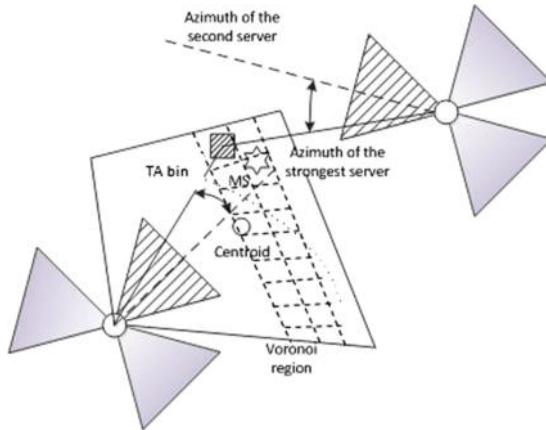


Fig. 4. Geometry of scenario for two sectors

Step 9: Determine the root mean square errors (RMSE) between the real location for each measured point and the estimated location by:

$$RMSE = \sqrt{(UE_{my} - UE_{ey})^2 + (UE_{mx} - UE_{ex})^2} \quad (13)$$

where UE_{my}, UE_{mx} are the real coordinates for UE and UE_{ey}, UE_{ex} are the predictable coordinates.

5 Results

We evaluate the efficiency of our method is explained in MATLAB by utilizing cell configuration and LTE measurements that reported by drive test on the city of Atlanta in Tables 1 and 2. In the first scenario (case 1) simultaneous measurements from two

reported cells from two various sites are available. The average of RMSE is around 53 m and a standard deviation (Std) 36 m. As shown in Fig. 5, the extreme approximation errors do not surpass 155 m. The cumulative distribution function (CDF) shows the probability estimated errors. It shows 60% of distance errors at 50 m, and 90% at around 100 m.

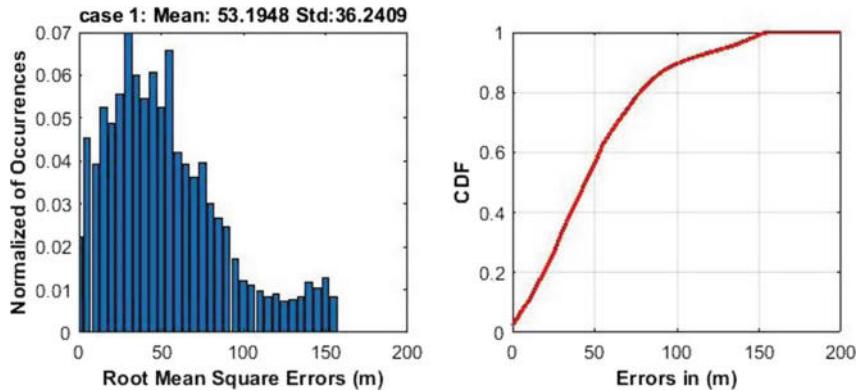


Fig. 5. Case 1 geolocation estimation error

In case 2 with three reported cells of two sites, the average RMSE is 44.6 m and Std is 33 m, as shown in Fig. 6. The maximum position errors do not exceed 130 m in this case. The CDF of this case displays 70% at approximately 60 m, and 95% below 100 m. consequently, the accuracy of scenario 2 increases around 16% over scenario 1.

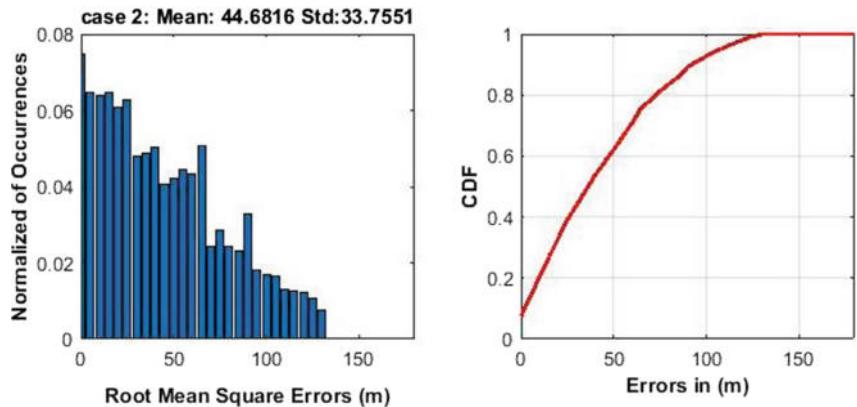


Fig. 6. Case 2 geolocation estimation error

Case 3 comes with three cells from three sites, the average RMSE in this case is 42 m and the Std around 31 m, as shown in Fig. 7. The extreme approximation errors don't exceed 125 m. The CDF displays 70% at approximately 50 m, and 92% around 90 m. This case increases the accuracy 21 and 5% compared to scenario 1 and scenario 2, respectively.

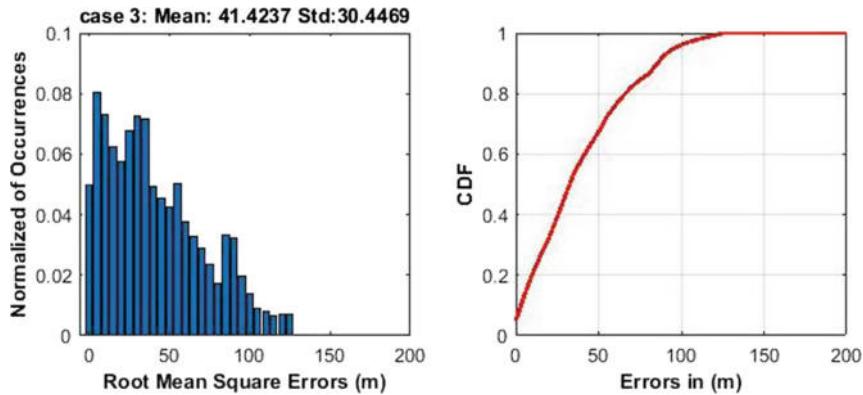


Fig. 7. Case 3 geolocation estimation error

For case 4 with four cells from two sites, the average RMSE and Std are 35 and 25 m, respectively as in Fig. 8. Results of CDF displays 62% at approximately 40 m and 85% at 60 m, and the maximum position errors do not exceed 100 m, as shown in the top panel of error statistics. Scenario 4 increases the accuracy approximately 34, 22, and 16% compared with scenario 1, scenario 2, and scenario 3, respectively.

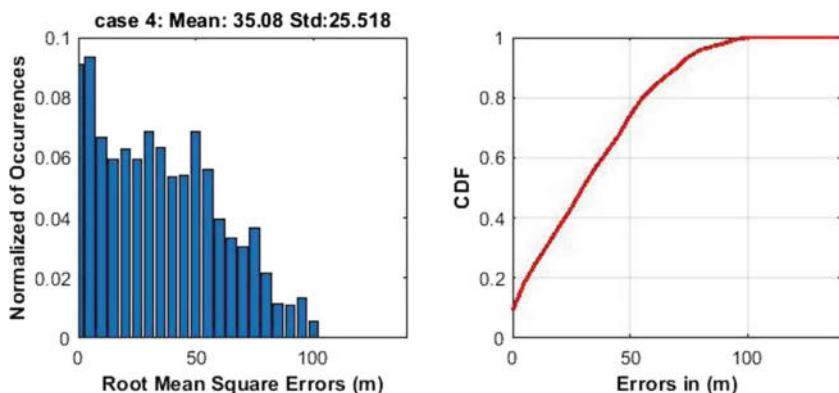


Fig. 8. Case 4 geolocation estimation error

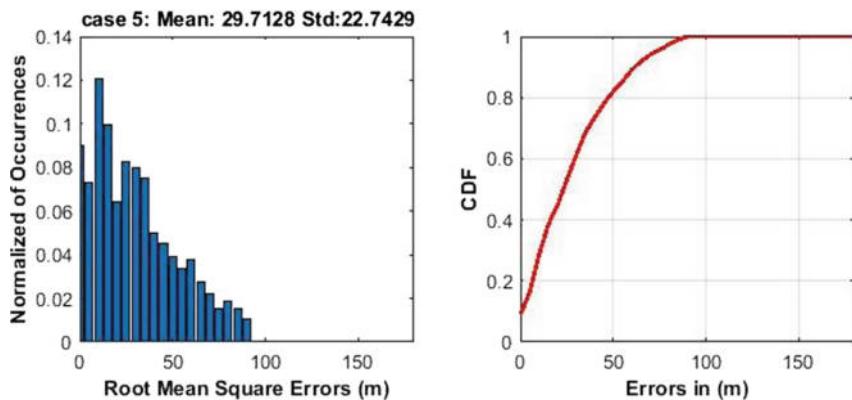


Fig. 9. Case 5 geolocation estimation error

For case 5 with five cells among up to 5 sites, the average RMSE and Std are 29.7 and 22.7 m, respectively as shown in Fig. 9. The supreme location errors do not surpass 90 m. CDF illustrates 70% at roughly 35 m and 90% at around 60 m. The accuracy in this case experiences lower measurement error than other cases. It improved in case 5 approximately 44, 33, 29, and 15% respectively over scenario 1, scenario 2, scenario 3, and scenario 4, respectively.

6 Conclusion

For this study, our developed method has been tested to estimate UE coordinate inside LTE networks. The algorithm utilizes time advance parameter and LTE measurement (RSRP) with simple cost function. Five cases are considered for evaluation; starting from 2 up to 5 simultaneously reported cells on various sites. The results display this method provides excellent accuracy to locate UE in LTE cellular networks. For future work, our method will be combined with other algorithm and optimized used different predictive models to enhance the accuracy.

References

1. Zhang, T., Xiao, D., Cui, J., Luo, X.: A novel OTDOA positioning scheme in heterogenous LTE advanced systems. In: 2012 3rd IEEE International Conference on Network Infrastructure and Digital Content, IC-NIDC 2012, pp. 106–110 (2012); Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) Conference 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016)
2. Huang, M., Xu, W.: Enhanced LTE TOA/OTDOA estimation with first arriving path detection. In: IEEE Wireless Communications and Networking Conference, WCNC, pp. 3992–3997 (2013)

3. Reed, J.H., Krizman, K.J., Woerner, B.D., Rappaport, T.S.: An overview of the challenges and progress in meeting the E-911 requirement for location service. *IEEE. Commun. Mag.* **30**-37 (1998)
4. Roth, J., Tummala, M., McEachen, J., Scrofani, J.: On mobile positioning via cellular synchronization assisted refinement (CeSAR) in LTE and GSM networks. In: *IEEE 9th International Conference on Signal Processing and Communication Systems (ICSPCS)* (2015)
5. Roxin, A., Gaber, J., Wack, M., Nait-Sidi-Moh, A.: Wireless geolocation techniques: a survey. In: *IEEE Global Telecommunication Conference* (2007)
6. Thorpe, M., Kottkamp, M., Rossler, A., Schutz, J.: LTE location based services technology introduction, white paper, Apr 2013
7. Caffery, J., Jr.: Subsciber location in CDMA cellular networks. *IEEE. Trans. Veh. Technol.* **47**(2) (1998)
8. Cotsnis, I., Le, L.: Aspects related to geolocation based on mobile measurement in WCDMA wireless networks. In: *International Conference in Computing, Networking and Communication, ICNC12*, pp. 902–906 (2012)
9. Qi, Y., Kobayashi, H., Suda, H.: Analysis of wireless geolocation in a non-line-of-sight environment. *IEEE. Trans. Wirel. Commun.* **5**(3) (2006)
10. Guan, W., Deng, Z., Ge, Y., Zou, D.: TDOA mobile location based Kalman filter in CDMA2000 cellular networks. In: *2010 6th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM* (2010)
11. Chen, Y., Yen, S.: Smart antenna with joint angle and delay estimation for the geolocation, smart uplink and downlink beamforming. In: *International Conference on Signal Processing Proceedings, ICSP*, vol. 1, pp. 393–397 (2002)
12. Roth, J., Tummala, M., McEachen, J., Scrofani, J., DeGabriele, R.: Maximum likelihood geolocation in LTE cellular networks using the time advance parameter. In: *10th International Conference on Signal Processing and Communication Systems, ICSPCS* (2016)
13. Jarvis, L., McEachen, J., Loomis, H.: Geolocation of LTE subscriber stations based on the timing advance ranging parameter. In: *IEEE Military Communications Conference MILCOM 2010*, pp. 180–187 (2010)
14. Shakir, Z., Zec, J., Kostanic, I.: Measurement-based geolocation in LTE cellular networks. In: *IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, Jan 2018
15. Shakir, Z., Zec, J., Kostanic, I.: Position location based on measurement reports in LTE cellular networks. In: *IEEE 19th Wireless and Microwave Technology Conference, WAMICON*, pp. 1–6, 23 May 2018



Hybrid Parallel Approach of Splitting-Up Conjugate Gradient Method for Distributed Memory Multicomputers

Akiyoshi Wakatani^(✉)

Faculty of Informatics and Intelligence,
Konan University, Kobe 658-8501, Japan
wakatani@konan-u.ac.jp

Abstract. This paper describes several variants of SPCG (Splitting Up Conjugate Gradient) method suitable for parallel computing and evaluates the performance and the speed of convergence on a distributed-memory multicomputer. SP (Splitting-Up) preconditioner can be easily parallelized because other dimensions except for one dimension are independent. Among the variants, one of incomplete SPCG method, which does not carry out one of three tridiagonal matrix solvers, achieves the best performance, and this method is about 20 times faster than one-process version of the SPCG method on 32 CPU cores of the multicomputer.

Keywords: Iterative methods · Tridiagonal matrix solver · Preconditioning

1 Introduction

Large-scale computer simulation can reveal the mechanism of natural phenomena, such as weather prediction and fluid dynamics. In general, these phenomena can be formalized by appropriate partial differential equations. These equations are usually difficult to be solved analytically, but the numerical solutions for them can be found by a large scale numerical simulation with high-speed computers. Such a simulation can be usually solved by not direct methods, but iterative methods. Iterative methods include stationary iterative methods such as SOR (Successive Over-Relaxation) method, and non-stationary iterative methods based on Krylov subspace method such as CG (Conjugate Gradient) method and GMRES method [1]. In recent years, the non-stationary iterative methods are frequently used because of their rapid convergence. However, since a high-speed computation using parallel processing is indispensable in order to achieve a large size simulation, we have to select an iterative method that is easy to be parallelized with keeping the convergence rate high. Usually, ICCG (Incomplete Cholesky Conjugate Gradient) method, which is based on modified Cholesky decomposition, is used, but it is difficult to parallelize the method straightforwardly.

SPCG (Splitting-Up CG) method has SP operator as a preconditioner, and it is one of preconditioned CG methods suitable for parallel processing on distributed memory computers. As mentioned in our prior art [2], a simple implementation of SPCG on distributed memory computers does not easily achieve a high performance computing because it requires a heavy communication cost due to array redistributions. One alternative is the implementation using P-scheme method that does not require array redistributions, and another alternative is a method without any communication, such as NoZSolve method and ISPCG method. Note that the latter methods may lead to degradation of convergence speed, so we evaluated total performance of our methods, which was the combination of the convergence speed and the execution time [2]. On the other hand, since recent processors consist of plural computing cores, parallel processing within a node with a shared memory must be exploited for a high performance computing.

In this paper, we will evaluate a hybrid approach that utilizes both parallel processing between nodes and parallel processing within a node for SPCG methods. In general, MPI is used for parallel processing between nodes and OpenMP is used for parallel processing within a node. We will also confirm the validity of this general approach on SX-ACE supercomputer. It should be noted here that our evaluation is carried out on SX-ACE (NEC) super computers that is installed at Osaka University (Japan). SX-ACE consists of 512 nodes that contains 4 vector CPUs and 64 GB memory (256 GB/s), and the data is transferred between the nodes at the speed of 8 GB/s. The theoretical peak performance of the CPU is 64 GFLOPS, so the total performance reaches 132 TFLOPS [3].

The rest of this paper is organized as follows: Sect. 2 presents the SPCG method for solving Laplace equation and Sect. 3 proposes the implementation policy for distributed-memory systems. Section 4 presents the experimental method and discusses the results and Sect. 5 concludes this paper with a summary.

$\mathbf{h}^n = C^{-1}(A\mathbf{u}^n - \mathbf{f})$ $\tau = \frac{(\mathbf{h}^n, Ch^n)}{(\mathbf{d}^n, Ad^n)}$ $\mathbf{u}^{n+1} = \mathbf{u}^n + \tau \mathbf{d}^n$ $\mathbf{h}^{n+1} = \mathbf{h}^n + \tau C^{-1} A \mathbf{d}^n$ $\beta = \frac{(\mathbf{h}^{n+1}, Ch^{n+1})}{(\mathbf{h}^n, Ch^n)}$ $\mathbf{d}^{n+1} = -\mathbf{h}^{n+1} + \beta \mathbf{d}^n$	$C = (D + X)D^{-1}(D + Y)D^{-1}(D + Z)$ $D^{-1} = [(\mathbf{0}, \mathbf{0}, \mathbf{0})(\mathbf{0}, (0, 1/6, 0), \mathbf{0})(\mathbf{0}, \mathbf{0}, \mathbf{0})]$ $D + X = [(\mathbf{0}, \mathbf{0}, \mathbf{0})(\mathbf{0}, (-1, 6, -1), \mathbf{0})(\mathbf{0}, \mathbf{0}, \mathbf{0})]$ $D + Y = [(\mathbf{0}, \mathbf{0}, \mathbf{0})((0, -1, 0), (0, 6, 0), (0, -1, 0))(\mathbf{0}, \mathbf{0}, \mathbf{0})]$ $D + Z = [(\mathbf{0}, (0, -1, 0), \mathbf{0})(\mathbf{0}, (0, 6, 0), \mathbf{0})(\mathbf{0}, (0, -1, 0), \mathbf{0})]$
(a) Iterative method	(b) Preconditioning

Fig. 1. Splitting Up CG method

2 Conjugate Gradient Methods

The CG method achieves an excellent performance in terms of convergence speed. However, the speed of convergence of iterative methods depends on a condition number of a matrix that should be solved. In general, when the condition number is close to 1, the speed of convergence is fast. Therefore, in order to improve the speed of convergence of the CG method, the preconditioning is applied to a matrix that should be solved. Usually, the ICCG method is used for solving a symmetric positive-definite matrix.

Although the speed of convergence of the ICCG method is fast, it is difficult to parallelize the method straightforwardly, and thus several variants have been studied for parallel processing. For example, by using a red-black ordering and a large-numbered multi-color ordering, the ICCG method can be parallelized, but the speed of convergence may be degraded [4]. Therefore, it is general that the parallelization of the ICCG method may result in the degradation of the convergence performance, although the computational performance can be improved.

On the other hand, SPCG (Splitting-Up Conjugate Gradient) method is a preconditioned CG method like the ICCG method, but the parallelization of the SPCG method is relatively easy and straightforward. According to the literature [5], it is known that the speed of convergence of the SPCG method is almost equal to that of the ICCG method.

Figure 1 shows SPCG method that applies to the following three-dimensional Laplace equation: $A\mathbf{u} = \mathbf{f}$ ($Au_{ijk} \equiv 6u_{ijk} - u_{i-1jk} - u_{i+1jk} - u_{ij-1k} - u_{ij+1k} - u_{ijk-1} - u_{ijk+1}$). Here, \mathbf{u}^n is an unknown array at the n -th iteration, τ_n and β_n are scalar parameters for the iterative method, \mathbf{h}^n is a residual vector and \mathbf{d}^n is a direction vector that is used for updating the unknown array.

3 SPCG Method

3.1 Base Method

SPCG method consists of the CG part and the precondition part (splitting-up part). The CG part in the SPCG method consist of an inner product calculation, a matrix-vector product calculation and a scalar-vector calculation, and these calculations can be easily parallelized on multicomputers if data can be distributed properly, so the details of the parallelization of this part is omitted.

On the other hand, the preconditioner (splitting-up part) solves three tridiagonal matrices for three dimensions in the case of three-dimensional Laplace equation. Namely, $(D + X)^{-1}$, $(D + Y)^{-1}$ and $(D + Z)^{-1}$ are tridiagonal matrix solvers for x-direction, y-direction and z-direction, respectively. Therefore, the solver for $(D + X)^{-1}$ can be parallelized in the y-direction and z-direction, the solver for $(D + Y)^{-1}$ can be parallelized in the z-direction and x-direction and the solver for $(D + Z)^{-1}$ can be parallelized in the x-direction and y-direction. In addition, a tridiagonal matrix is solved by Thomas method, which consists of a

forward substitution and a backward substitution. A tridiagonal matrix equation is as follows:

$$\begin{aligned} b_0x_0 + c_0x_1 &= d_0 \\ a_i x_{i-1} + b_i x_i + b_i x_{i+1} &= d_i \quad (i = 1, \dots, N-2) \\ a_{N-1} x_{N-2} + b_{N-1} x_{N-1} &= d_{N-1} \end{aligned}$$

where a , b , c and d are given in advance and x is an unknown vector. By using auxiliary vectors l and m , a forward substitution is as follows:

$$\begin{aligned} l_0 &= -\frac{c_0}{b_0}, \quad m_0 = -\frac{d_0}{b_0} \\ l_i &= \frac{-c_i}{a_i l_{i-1} + b_i} \quad (i = 1, \dots, N-2) \\ m_i &= \frac{d_i - a_i m_{i-1}}{a_i l_{i-1} + b_i} \quad (i = 1, \dots, N-2). \end{aligned}$$

On the other hand, a backward substitution is as follows:

$$\begin{aligned} x_{N-1} &= \frac{d_{N-1} - a_{N-1} m_{N-2}}{a_{N-1} l_{N-2} + b_{N-1}} \\ x_i &= l_i x_{i+1} + m_i \quad (i = N-2, \dots, 0). \end{aligned}$$

It should be noted that these substitution are carried out for adjacent array elements.

In order to distribute three-dimensional arrays on distributed-memory multicomputers, at least one dimension of three dimensions must be distributed. For example, when the distributed dimension is z-direction, tridiagonal matrix solvers in x-direction and in y-direction can be carried out without communications because all the data in x-direction and in y-direction are stored in a local memory, but a tridiagonal matrix solver in z-direction needs communications because the data in z-direction are distributed over processors.

3.2 Implementation on Distributed Memory Systems

Array Redistribution Array redistribution, which transforms the configuration of data distribution, is a simple method to implement communications in z-direction [SPCG (redist)]. For example, when an array distributed in z-direction is redistributed in y-direction, Thomas method in z-direction can be done without communications because all the data in z-direction are stored in a local memory. However, since array redistributions must be carried out before and after the tridiagonal matrix solver, the overhead of the communication must be considered. It should be noted that the total computational complexity of Thomas method is $O(N^3/P)$, and the communication cost is $O(N^3)$ where P is the number of processors.

No Array Redistribution We have two methods without array redistributions to implement the SPCG method in parallel. The first one utilizes P-scheme [6] method that authors have proposed for solving a tridiagonal matrix without array redistributions [SPCG (p-scheme)]. This method requires just three one-to-one communications between neighboring processors, but the computational complexity is a few times larger than the original SPCG method. However, since the communication time is usually dominant on multicomputers, this method is effective on distributed-memory systems. The tradeoff of computation and communication will be evaluated later. It should be noted that the total computational complexity is $O(N^3/P)$, and the communication cost is $O(N^2 \times P)$. In general, $O(N^2 \times P)$ is less than $O(N^3)$.

On the other hand, a pipelining is utilized in Thomas method in z-direction [SPCG (pipe)]. As mentioned, Thomas method consist of a forward substitution and a backward substitution. For example, when $i = 1$ and $j = 1$, a processor with rank k carries out the forward substitution using data that this processor has, and then this processor sends the last data to a processor with rank $k + 1$. After that, the processor with rank $k + 1$ carries out the forward substitution by using the received data, and the processor with rank $k + 1$ carries out the forward substitution when $i = 1$ and $j = 2$. Thus these processors run in concurrent, and other processors can run concurrently as well. The backward substitution is processed in the same way. However, in order to reduce the number of communications, several communications must be coalesced. So, in this paper, all the data in y-direction are communicated together, and the data in x-direction are processed in a pipeline manner.

Incomplete SPCG The last variants do not utilize communications in order to improve a performance, although they deviate from the original SPCG method. Because SP preconditioner incompletely solves an original problem by nature, it is not so significant that the incompleteness of the preconditioner increases a little bit. So, the first method does not carry out the tridiagonal matrix solver in z-direction (NoZsolv). Since this does not solve in z-direction, the elapsed time can be reduced, and the communication time can be also alleviated, but a speed of convergence may be degraded. Therefore, the convergence performance should be evaluated by experiments, and thus a total performance, which is a combination of a speed of convergence and a computational speedup, should be also considered.

Next, the second method incompletely solves the tridiagonal matrix in z-direction [ISPCG (Incomplete SPCG)]. Namely, like SPCG (pipe), this method solves the matrix in z-direction, but this does not send the last value of the forward substitution to the neighboring processor. Since this method is closer to the original SPCG method than NoZSolve, a speed of convergence may be superior to NoZSolve, but the computational complexity increases a little bit. For this method, a total performance should be also considered by combining a speed of convergence with a computational speed.

3.3 Total Performance

In order to show the difference in performance of the SPCG implementations, a total performance of these implementations are measured by using MPI. The total performance is a combination of the speed of convergence and the elapsed time of one iteration. Namely, a total performance is an elapsed time that is required for achieving a given value of residual for the variants of the SPCG method. Figure 2 shows the total performance when the data size is 96^3 . Note that the number of processes is 32 (8 nodes), and SPCG (1 proc) shows the result of 1 process (1 CPU core).

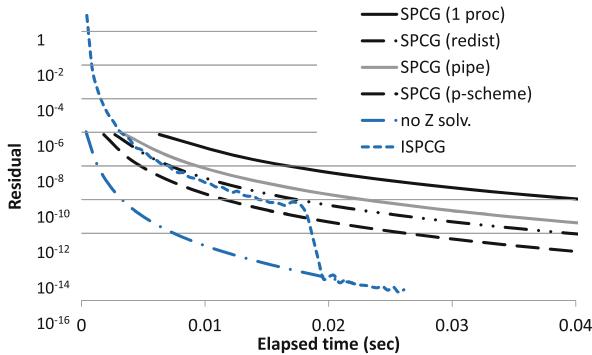


Fig. 2. Total performance (size = 96^3)

As a whole, NoZSolve exhibits the best performance. For example, NoZSolve requires about 2 ms to reduce the residual to 1.0×10^{-10} , but SPCG (1 proc) needs about 38.5 ms, so the speedup is about 20. When the residual is 1.0×10^{-10} , the elapsed times of SPCG (pipe) and SPCG (p-scheme) are about 23 ms and about 13 ms, respectively. ISPCG is somewhat worse than NoZSolve, but both require the almost same time to reduce the residual to 1.0×10^{-15} .

Figure 3 shows the total performance when the data size is 192^3 . In this case, NoZSolve exhibits the best performance, too. When the residual is reduced to 1.0×10^{-7} , NoZSolve achieves the speedup of about 18 compared with SPCG (1 proc). Other results are almost same as the previous case, except for ISPCG. Namely, SPCG (p-scheme) exhibits the second best performance. The reason why the total performance of ISPCG is not good is that the speed of convergence is bad until 1.0×10^{-8} .

According to the results of the experiments, SPCG (p-scheme) does not overcome NoZSolve, but the convergence performance of incomplete SPCG methods may be degraded when other parameters are taken. When the bandwidth of network of multicomputers is fast, the elapsed time of SPCG (p-scheme) can be improved, and thus the total performance of SPCG (p-scheme) may be comparable or superior to that of NoZSolve.

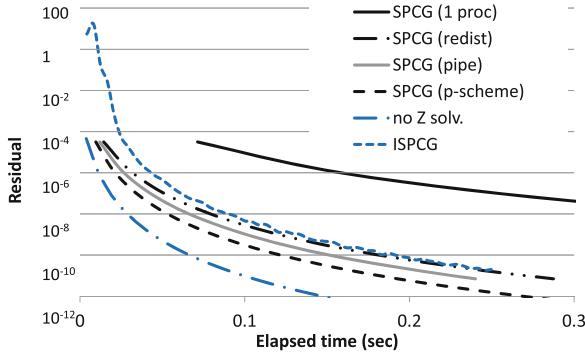


Fig. 3. Total performance (size = 192^3)

4 Evaluation

We evaluate the performance of hybrid parallelization for our SPCG methods on SX-ACE supercomputer. Since each node of the supercomputer consists of four computing cores with a vector unit and a shared memory (64 GB), up to four threads can be consumed without waste on a node.

Therefore, we evaluate elapsed time of our SPCG methods in four ways: 4P1T is that each node has 4 MPI processes that consists of 1 thread each, 1P2T is that each node has 1 MPI process that consists of 2 threads each, 1P3T is that each node has 1 MPI process that consists of 3 threads each, and 1P4T is that each node has 1 MPI process that consists of 4 threads each. When the number of nodes is 4, 8 and 16, the size of data is 192^3 and 284^3 , we measure the elapsed time of one iteration of the SPCG implementations. The results of our experiments are shown in Figs. 4, 5 and 6.

On the whole, when the size of data is small, 4P1T is superior to the others. Whereas, when the size of data is large, 1P4T is the best. For example, when the size of data is 192^3 and the size of nodes is 16, the elapsed times of 4P1T, 1P2T, 1P3T and 1P4T of NoZSolve are 1.28, 3.16, 4.93, 2.27 ms, respectively. On the other hand, when the size of data is 384^3 and the size of nodes is 16, the elapsed times of 4P1T, 1P2T, 1P3T and 1P4T of NoZSolve are 36.13, 44.3, 56.38, 28.43 ms, respectively.

As mentioned earlier, redist requires communication cost that is in proportion to the size of data, while pipe and P-scheme require communication cost that is in proportion to the number of MPI processes and one third of the data size. Thus, the elapsed times of pipe and P-scheme are generally sensitive to the number of nodes. When the size of data is small, an overhead cost of generation of OpenMP threads relatively increases, so the effectiveness of OpenMP is degraded.

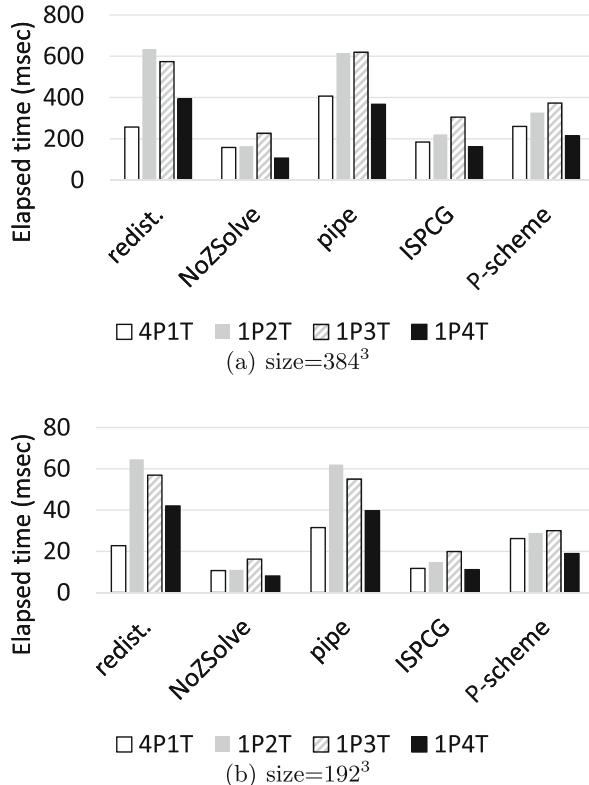
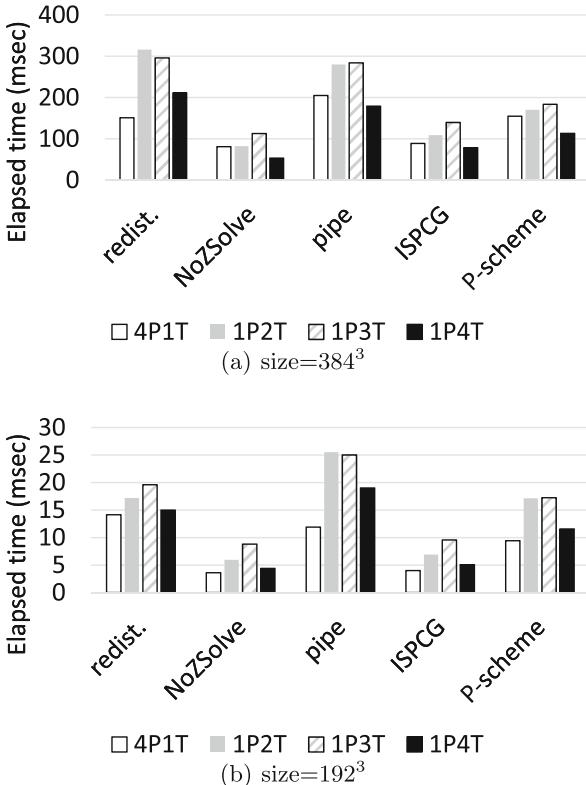


Fig. 4. 4 nodes case

When the number of nodes is 8 and the size of data is 384^3 , 4P1T is superior to the others for redist method, but 1P4T is the best for the other methods. The reason why 4P1T is the best for redist method is that the communication cost is irrelevant to the number of MPI processes. When the size of data is 192^3 , 4P1T is the best in all the cases. Each thread is in charge of 8 data when a dimension is divided by 32 threads. So, the effectiveness of multi-threading (1P4T) is degraded since the cost of generation of OpenMP threads relatively increases. However, when a dimension is divided by 32 processes (4P1T), the runtime overhead of thread generation is not required, so 4P1T outperforms 1P4T in the small data case.

When the number of nodes is 16, 1P4T of redist method is superior in the case of large data since OpenMP thread generation cost relatively decreases, and 4P1T of redist method is fast in the case of small data since no thread generation cost is required.

The results of our experiment is summarized as follows:

**Fig. 5.** 8 nodes case

- For redist method, 4P1T is mostly superior to the others, since the communication cost is irrelevant to the number of MPI processes and 4P1T does not require OpenMP thread generation cost.
- For pipe method, 4P1T is mostly faster than the others, since the method requires thread synchronization between data communications.
- For the other methods, 1P4T is the best when the computational complexity per thread is large, and 4P1T is the best otherwise.
- Among all the cases, 1P4T of NoZSolve is the fastest. In addition, 1P4T of P-scheme is the fastest in the cases where the convergence speed is fast.

5 Conclusion

This paper describes several variants of the SPCG method suitable for parallel computing and evaluates the performance and the speed of convergence on a distributed-memory multicomputer.

Among the variants, one of incomplete SPCG method, which does not carry out one of three tridiagonal matrix solvers, achieves the best performance, and

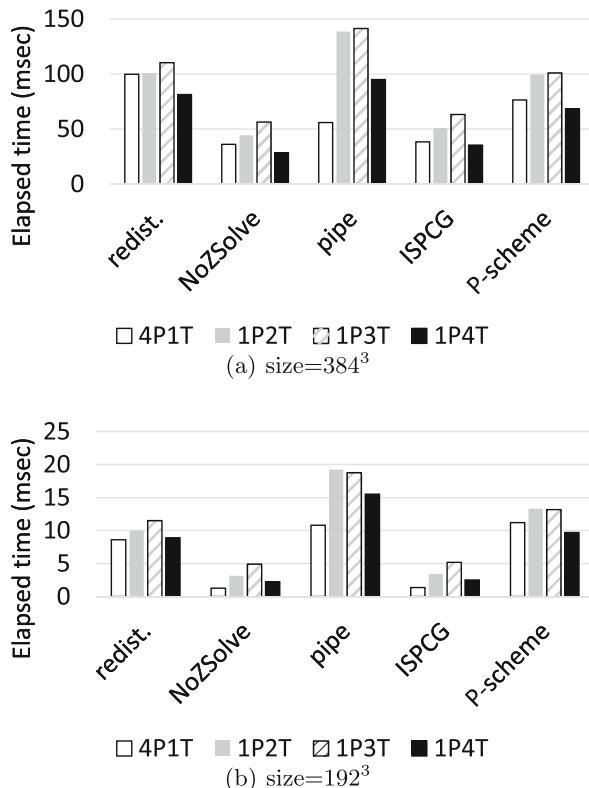


Fig. 6. 16 nodes case

this method is about 20 times faster than one-process version of the SPCG method on 32 CPU cores of SX-ACE supercomputer.

In the future, we will compare the performance of the SPCG method with that of a parallelized ICCG method and we will also evaluate our methods on other multicomputers in which the speed of network is faster than the SX-ACE.

Acknowledgements. We are grateful to Professor Tatsuo Nogi of Kyoto University for helpful discussions. I would like to express my gratitude to both professors. This work was supported by JSPS KAKENHI Grant Number 18K02920. This research was also partially supported in part by MEXT, Japan.

References

1. Saad, Y., Schultz, M.H.: GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.* **7**(3), 856–869 (1986)
2. Wakatani, A.: An incomplete splitting-up conjugate gradient method for parallel computing. *Int. J. Comput. Technol. Appl.* **7**(5), 236–243 (2016)

3. Osaka University. <http://www.hpc.cmc.osaka-u.ac.jp/sx-ace/>
4. Duff, I.S., Meurant, G.A.: The effect of ordering on preconditioned conjugate gradients. *BIT Numer. Math.* **29**(14), 635–657 (1989)
5. Odanaka, S., Nogi, T.: Massively parallel computation using a splitting-up operator method for three-dimensional device simulation. *IEEE Trans. Comput. Aided Des. Integr. Circ. Syst.* **14**(7), 824–832 (1995)
6. Wakatani, A.: A parallel and scalable algorithm for ADI method with pre-propagation and message vectorization. In: *Parallel Computing*, vol. 30, pp. 1345–1359 (2004)



Exceeding the Performance of Two-Tier Fat-Tree: Equality Network Topology

Chane-Yuan Yang^(✉), Chi-Hsiu Liang, Hong-Lin Wu,
Chun-Ho Cheng, Chao-Chin Li, Chun-Ming Chen, Po-Lin Huang,
and Chi-Chuan Hwang

Department of Engineering Science, National Cheng Kung University,
701 Tainan, Taiwan
eurekacyyang@gmail.com

Abstract. A high-performance interconnect topology system named Equality is introduced in this paper for general purpose applications including supercomputing, data center, cloud service, and industrial cluster solutions. Equality is designed based on chordal ring networks. It advances previous discussed chordal ring topologies by a set of systematic linking strategies and routing rules. The Equality topology can be used construct low diameter networks with reasonably low router radices. Equality interconnects are highly symmetric and hence cabling rule and routing logic are simple. Compared with other networks, the Equality topology is flexible in total number of routers, where any even number is allowed. This paper introduces the evaluation of Equality performance using open-source BookSim 2.0 package. The benchmarks of ten traffic models for systems constructed using 36- and 48-port switch are presented to assess the network performance, compared with the very popular 2-tier fat-tree structure. The results show that the Equality networks are resilient in the scenario similar to practical computation.

Keywords: Equality network topology · Fat-tree · High speed computation · Data center · BookSim

1 Introduction

THE capability of computers has kept growing steadily for the last few decades [1] depending on the cooperatively developments on chip design, system integration, and community efforts. The way all components of the system is connected, the network, plays an important role to make the system more coherent for parallel computation. High performance computing (HPC) communities have long been searching for better networking topologies matching the need for all purposes. However, there are only a few topologies stand out from the others and survive in the industrial and academic implementations. Furthermore, these topologies still suffer performance bottlenecks, routing difficulties and scaling barriers derived from the fundamental designs.

In the language of computer network, the cost of network communication is related to: 1. network topology; 2. programming model semantics; 3. data handling and routing, and 4. communication software protocols. Especially, time for communicating

a message between two nodes is the sum of time to prepare a message for transmission and time taken by the message to traverse the network to its destination. That is, network topology dominants the efficiency of communication.

The other question in the modern computer network is that the number of nodes in all state-of-art topologies has large gaps between designs of different network radices, dimensions and server concentrations. The reason for this is that most of the topologies have total router number being the product of integers. For instance, a full-scale nonblocking 2-tier fat-tree (2-T FT) using 80-port switches can support up to 3200 servers, with 80 leaf and 40 core switches. One can also configure a half-scale fat-tree with 1600 endpoints, using 40 leaf and 20 core switches. However, if one likes to scale down the system a nonintegral fraction, he/she has to sacrifice some ports in the second layer switches. Similar situations are in the popular topologies such as torus [2, 3], flattened butterfly [4] or even the latest low diameter topologies Dragonfly [5], orthogonal fat-tree [6] and Slim Fly [7].

We aim to provide an adjustable network topology for general purpose HPC, from standard engineering work-loads to artificial intelligence, from small clusters to supercomputers, and hopefully help to achieve the milestone of the whole human brain simulation.

Recently, we have introduced a new flexible yet high-performance interconnect topology named Equality for general purpose applications including supercomputing, data center, cloud service, industrial cluster solutions and chip design, see Fig. 1. Details can be found in the paper submitted to IEEE journal [8]. Equality provides a way to construct networks of flexible sizes, adaptable computation-to-communication ratios, low latency, high throughput, high resilience, high path diversity, simple routing logic, simple cabling and reasonable cost, matching the requirements of the next-generation general purpose HPC.

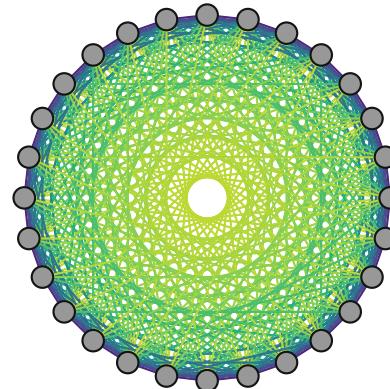


Fig. 1. Schematic example of Equality network: number of 36-port switch = 28 with 336 servers.

In the aspects of high-speed computing and data center, FT [9] topologies are popular for their nonblocking nature, providing many redundant paths between any 2 hosts. Such topologies have been used to build 90% fast and efficient super computers and commercial data centers, such as Google [10] and Facebook [11]. However, there are many topology inherited inconveniences in the FT topologies. For example, a FT topology requires high radix routers for large networks if a 2- or 3-tier FT being the constraint. In addition, FT networks perform extremely well on all permutation traffic patterns, only that the cost to build for large network is considerably higher, and the zero-load latency of a fat-tree network is determined by the number of the layers. Facebook devoted to improve performance of FT in view of the natural limitation of FT [11]. The improvement, however, is very limited.

On the other hand, practical problems in the high-speed computation and data center are 1. the communication modes are nonuniform, and 2. the size of message packets distributes in a wide range [12]. In this work, we construct the Equality network by using 36-port (e.g. Mellanox) and 48-port (e.g. Intel) switches and evaluate the performance of Equality, compared with that of two-tier Fat-Tree (2T-FT) with various packet sizes under nonuniform traffic patterns. Simulations are carried out by using open-source BookSim 2.0 package [13]. Below we first interpret the specific rules for construction of Equality topology.

2 Equality Topology

Here we briefly introduce the construction of Equality network topology. Note that the mathematical details are not provided herein.

2.1 Connection of Links

To make the interconnects, every member of the routers make links through the specific rules that we developed. Details of the rules can be found in the paper that has been submitted to IEEE journal [8].

For an Equality network has N routers, where N is an even number, the link rules are described below.

Step 1: The routers are sequentially numbered from zero to $N-1$, i.e. r_0, r_1, \dots, r_{N-1} .

Step 2: The routers are firstly connected to form a ring, and later connected with other members in the ring just like in the chordal ring topologies.

Step 3: A set of positive integers, C , starting from 2 to $N-3$, excluding any even numbers greater than $N/2$, are used as the candidates for making the physical links.

Step 4: From C , a subset of integers, S , are selected each time for an Equality topology, which are used to complete the interconnections.

Step 5: The connections are made for every router r_i with $r_{(i+S_j)\bmod N}$ if i is even, or with $r_{(iS_j)\bmod N}$ if i is odd, for every number S_j . The definition above defines the Equality topology.

The general syntax of Equality is shown in Table 1, which involves a ‘ N ’ denoting the total number of routers, followed by a number and a ‘ K ’ followed by another number denoting the network radix. The notation of ‘ p ’ can be omitted if the number of

attached endpoints is not yet specified. The notation of ‘N’, ‘K’ and ‘p’ allows both uppercase and lowercase as long as they are in a sequence to describe the constraints of the target network.

Table 1. Symbols and notations used in the paper.

General network terms	
Notation	Explanation
N2048K38P8	Notation for the configuration of a group of Equality networks. In this example, it represents a network configuration consists of 2048 switches, where each switch has network radix of 38 and attached with 8 endpoints. This notation can also be written as n2048k38p8 (allowing both uppercase and lowercase). Each of the networks in the group can be specified in more detailed notation
Variable	Explanation
N	Total number of nodes/routers in the network
K	The number of inter-node links per node/router (network radix)
p	The number of endpoints per router
P	Router radix, i.e. $P = K + p$

A simple example of Equality networks is presented in Fig. 2, named N14K6 [1,1,3,9](4). A pair of square brackets and a pair of parenthesis with numbers are for detailed specification of an Equality network. The numbers in the square brackets are an array of odd numbers S_A for odd and alternative links, whereas the numbers in the parentheses are an array of even numbers S_B for even links. For instance, Equality networks in Fig. 2 is described by the ‘N’ and ‘K’ that mark the number of router nodes and network radix constraints, respectively. Therefore, the number 14 means there are in total 14 routers in the network, and the number 6 indicates that the network radix of the routers is six. The notation is for the designers to have a rough idea of what configuration the network has followed rather than the full specification of the network.

2.2 Network Optimization

We optimize Equality networks with genetic algorithm implemented in TopologyWeb. The goal of the optimization is to minimize the product of the average distance and network diameter. In the end of the optimization, a series of best results from generations of evolution are reported. If the search space being explored is large, the optimized results are not necessarily to be the global minimum; however, the results are usually low enough for application. If sufficient evolution time is given, one can get results close to the global minimum.

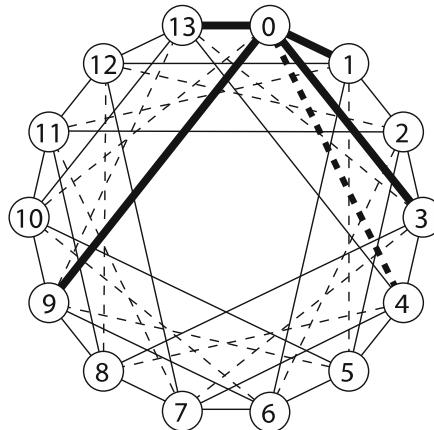


Fig. 2. A simple Equality network named N14K6[1,1,3,9](4). The bold lines marks the links initiate from r_0 . The solid lines marks odd links and the dashed lines marks even links.

3 Results and Discussions

We have evaluated the dynamic communication performance of Equality networks using deadlock-free dimension order routing with minimum number of virtual channels under the hot-spot traffic pattern.

Network traffic is an important metric that has to be considered when designing an optimized parallel computation system. The performance of a specific topology may vary depending upon the generated traffic pattern, thus making traffic pattern one of the most important parameters during design of application specific routing topologies and routing algorithms.

Under uniform traffic each node sends message to other nodes with an equal probability and destination nodes are chosen randomly using a uniform distribution. This traffic model is considered as a standard benchmark in network studies. FT topology performs very well under uniform traffic in both latency and throughput.

In real application, however, the traffic is non-uniform. On the other hand, the packet size of message is various in the communication. A long message is generally cut into pieces and the message pieces are sent through the network one followed by another.

3.1 Performance with Various Packet Size

System of 36-port switch

In the system using 36-port switch there are 28 switches and 336 servers. The performance of such system is shown in Fig. 3 along with that of 2-tier fat-tree for comparison. In the 36-port switch system of 2-tier fat-tree, there are 27 switches and 324 servers in total, i.e. system of Equality topology is slightly larger than that of 2-tier fat-tree structure.

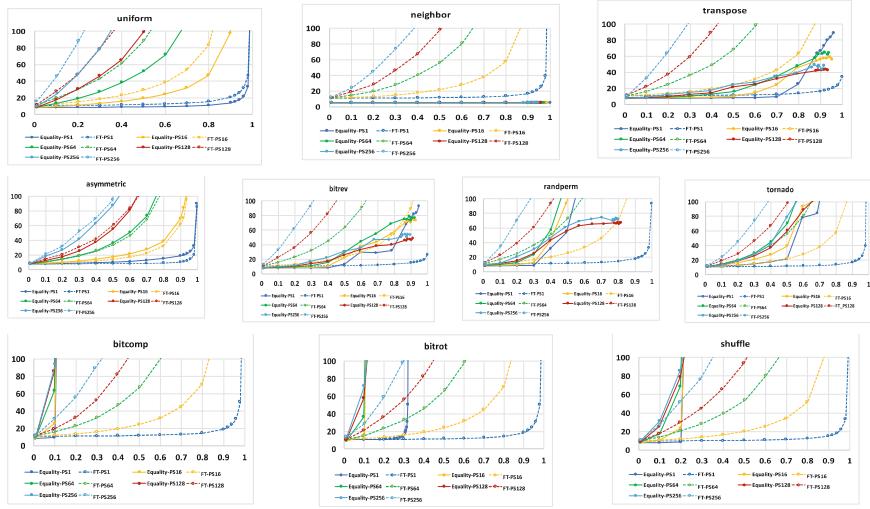


Fig. 3. Latency of 36-port switch system compared with 2-tier fat-tree. In the caption, the number after PS is packet size.

The first row of Fig. 3 includes the results of uniform, neighbor and transport traffic patterns. Equality performs better in these cases.

The second row includes the results of asymmetric, bit reverse (bitrev), rand permutation (randperm), and tornado. Equality performs better with large packet size in these cases.

The third row includes the results of bit compliment (bitcomp), bit rotation (bitrot), and shuffle. 2T-FT performs better in these cases.

System of 48-port switch

In the system using 48-port switch there are 36 switches and 576 servers. The performance of such system is shown in Fig. 4 along with that of 2-tier fat-tree for comparison. In the 48-port switch systems of both Equality and 2-tier fat-tree, there are 48 switches and 576 servers in total. With packet size distributes from 1 to 256 flits, performance of Equality is comparable with that of 2T-FT under 10 traffic modes.

The first row of Fig. 4 includes the results of uniform, neighbor, transport and bit reverse (bitrev) traffic patterns. Equality performs better in these cases.

The second row includes the results of asymmetric, tornado and rand permutation (randperm). Equality performs better with large packet size.

The third row includes the results of bit compliment (bitcomp), bit rotation (bitrot), and shuffle. 2T-FT performs better in these cases.

2-T FT performs better than Equality in bitcomp, bitrot and shuffle in 36- and 48-port switch systems.

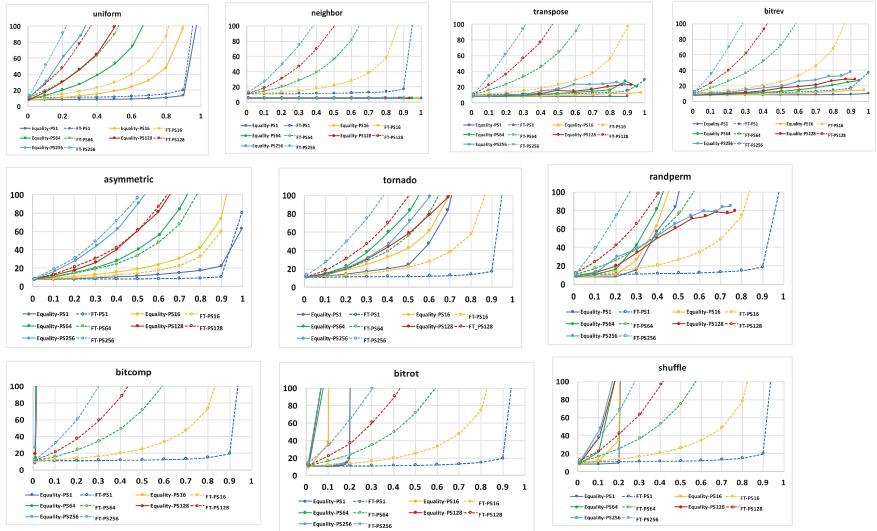


Fig. 4. Latency of 48-port switch system compared with 2-tier fat-tree.

3.2 Combination of Various Packet Size and Hot-Spot Traffic

In this section, we evaluate the performance of the Equality network topology under nonuniform hot spot traffic pattern. One of the most important non-uniform traffic models is hot spot pattern [14]. Hot spot traffic nodes are very busy nodes in a network. In the hotspot traffic pattern, each node sends messages to other nodes with an equal probability except for a specific node i.e. the hotspot that receives message with a greater probability. The percentage of messages that a hotspot node receives beyond the usual nodes is indicated after the hot spot name, e.g. hot spot 10%.

System of 36-port switch

Latency of 36-port switch system with various packet size under hot-spot traffic pattern, compared with 2-tier fat-tree is shown in Fig. 5. It is clear to see that under 10, 50 and 100% hot-spot traffic, Equality topology (solid lines) perform better than 2-T FT (dashlines) in all the cases.

System of 48-port switch

Latency of 48-port switch system with various packet size under hot-spot traffic pattern, compared with 2-tier fat-tree is shown in Fig. 6. Again, it is clear to see that 10, 50 and 100% hot-spot traffic, Equality topology (solid lines) perform better than 2-T FT (dashlines) in all the cases.

3.3 Combination of Various Packet Size and Nonuniform Traffic

In the simulations for nonuniform traffic, the system is divided into two groups. The traffic loading of the first group is heavier than that of the second group. The percentage of messages that the first group receives beyond the second is indicated before the name, e.g. 10% nonuniform.

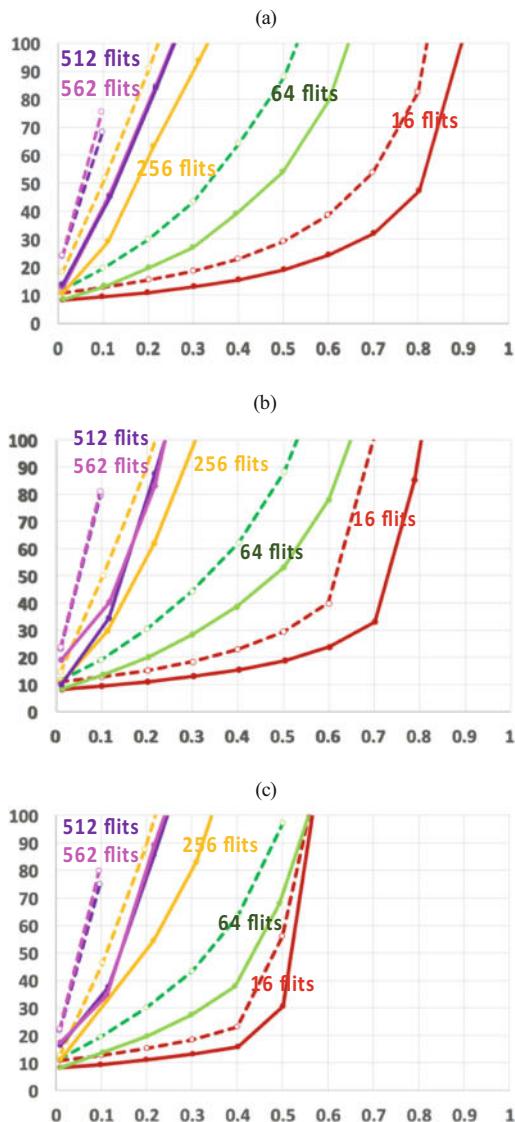


Fig. 5. Latency of 36-port switch system with various packet size, compared with 2-tier fat-tree.
a 10% hot-spot; **b** 50% hot-spot and **c** 100% hot-spot.

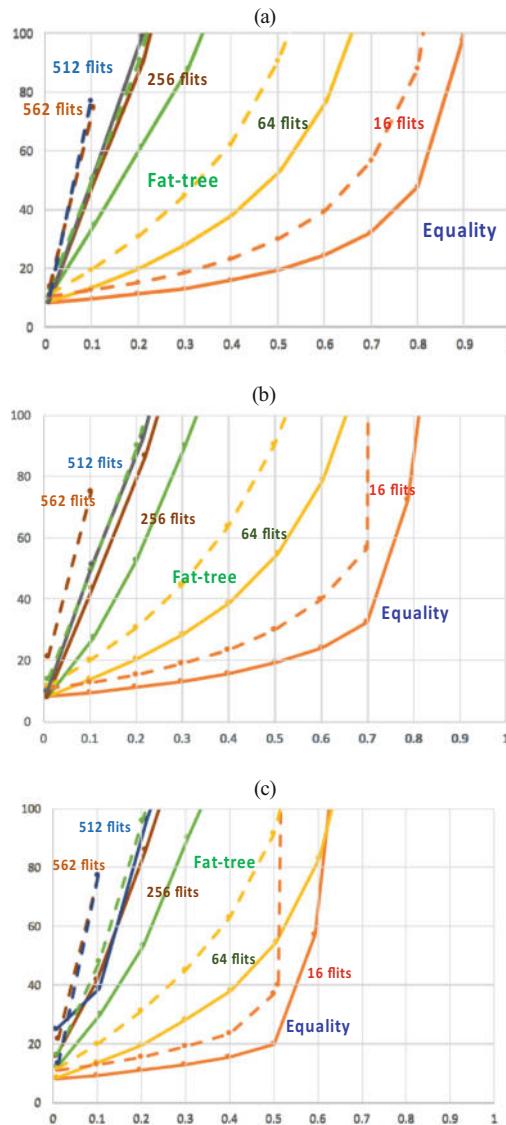


Fig. 6. Latency of 48-port switch system with various packet size, compared with 2-tier fat-tree. **a** 10% hot-spot; **b** 50% hot-spot and **c** 100% hot-spot.

System of 36-port switch

Latency of 36-port switch system with various packet size under nonuniform traffic pattern, compared with 2-tier fat-tree is shown in Fig. 7. It is clear to see that under 10, 50 and 100% nonuniform traffic, Equality topology (solid lines) perform better than 2-T FT (dashlines) in all the cases.

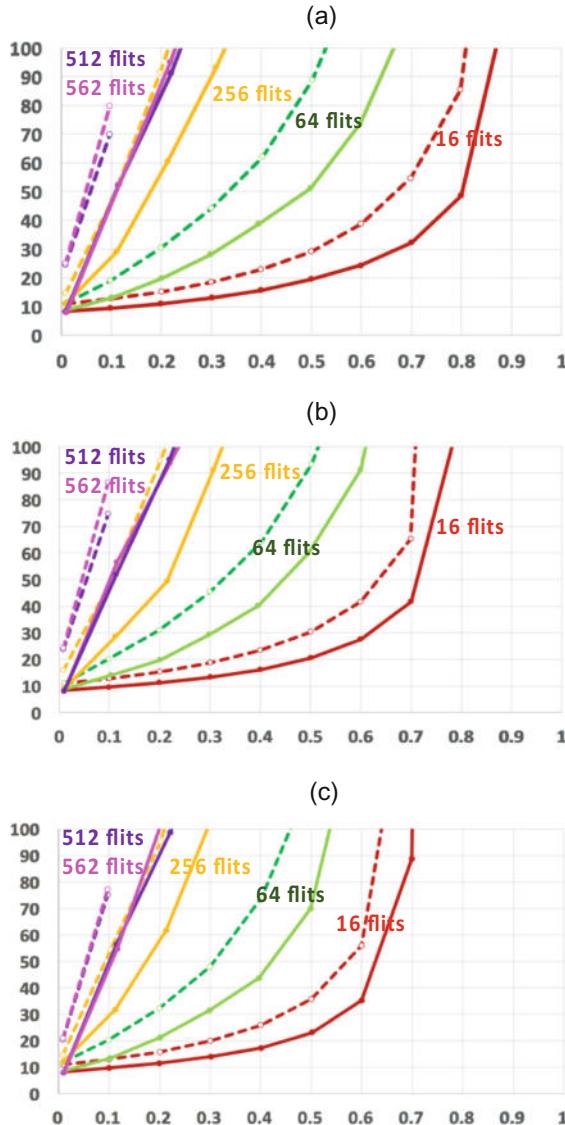


Fig. 7. Latency of 36-port switch system with various packet size, compared with 2-tier fat-tree.
a 10% nonuniform; **b** 50% nonuniform and **c** 100% nonuniform.

System of 48-port switch

Latency of 48-port switch system with various packet size under nonuniform traffic pattern, compared with 2-tier fat-tree is shown in Fig. 8. It is clear to see that under 10, 50 and 100% nonuniform traffic Equality topology (solid lines) perform better than 2-T FT (dashlines) in all the cases.

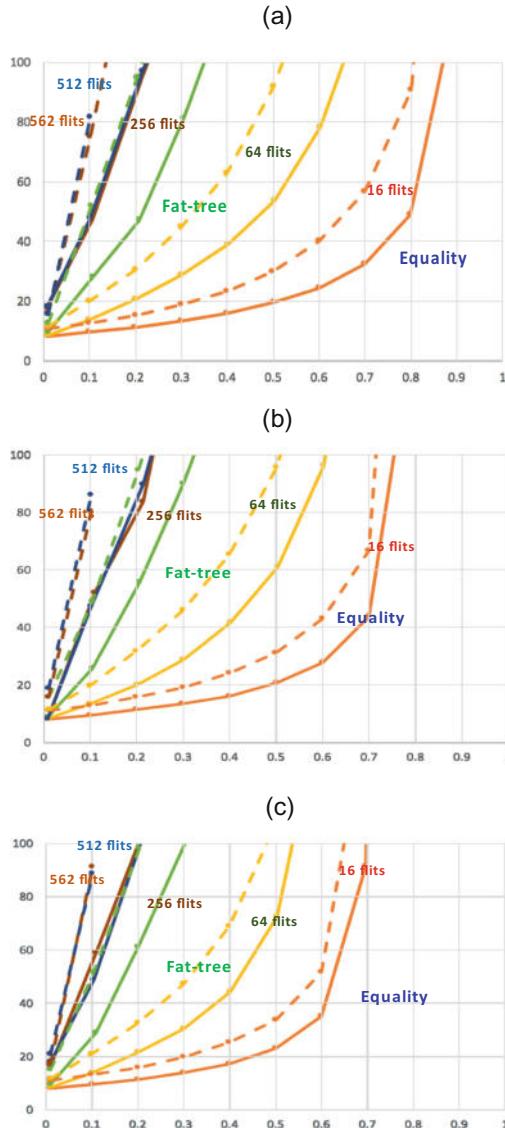


Fig. 8. Latency of 48-port switch system with various packet size, compared with 2-tier fat-tree. **a** 10% nonuniform; **b** 50% nonuniform and **c** 100% nonuniform.

4 Conclusion

We have proposed a novel Equality topology for general purpose applications including supercomputing, data center, cloud service, industrial cluster solutions and chip design. In this study the performance of Equality topology is evaluated in the scenario similar to scientific computation, in which various sizes of packets and

nonuniform communication traffics happen. Two metrics including latency and throughput are used for the evaluation. Also present is the performance of the popular 2-tier fat-tree network for the purpose of comparison. Simulations are carried out by the open-source BookSim 2.0 package.

With various packet sizes ranging from 1 to 256 flits, performance of Equality is comparable with that of 2T-FT under 10 traffic modes. While under hot-spot traffic and nonuniform traffic with various packet size, which is similar to the practical scenario of scientific computation, Equality topology performs better than 2-T FT in all the cases.

In conclusion, to tackle the problems that happen in practical scientific computations and data center, i.e. combination of various packet sizes and nonuniform communication patterns, our Equality topology can totally replace the popular fat-tree systems.

References

1. Denning, P.J., Lewis, T.G.: Exponential laws of computing growth. *Commun. ACM.* **60**(1), 54–65 (2016)
2. Kessler, R.E., Schwarzmeier, J.L.: CRAY T3D: a new dimension for Cray Research. In: *Compcon Spring'93, Digest of Papers*, pp. 176–182, 22 Feb 1993. IEEE
3. Ajima, Y., Sumimoto, S., Shimizu, T.: Tofu: a 6D mesh/torus interconnect for exascale computers. *Computer* **42**(11), 36–41 (2009)
4. Kim, J., Dally, W.J., Abts, D.: Flattened butterfly: a cost-efficient topology for high-radix networks. In: *ACM SIGARCH Computer Architecture News*, vol. 35, no. 2, pp. 126–137, 9 June 2007. ACM
5. Kim, J., Dally, W.J., Scott, S., Abts, D.: Technology-driven, highly-scalable dragonfly topology. In: *ACM SIGARCH Computer Architecture News*, vol. 36, no. 3, pp. 77–88, 21 June 2008. IEEE Computer Society
6. Valerio, M., Moser, L.E., Melliar-Smith, P.M.: Fault-tolerant orthogonal fat-trees as interconnection networks. In: *IEEE First International Conference on Algorithms and Architectures for Parallel Processing*, 1995. ICAPP 95. IEEE First ICA/sup 3/PP., vol. 2, pp. 749–754, 19 Apr 1995. IEEE
7. Besta, M., Hoefer, T.: Slim fly: a cost effective low-diameter network topology. In: *International Conference for High Performance Computing, Networking, Storage and Analysis*, SC14, pp. 348–359, 16 Nov 2014. IEEE
8. Liang, C.-H., Yang, C.-Y., Li, C.-C., Cheng, C.-H., Wu, Y.-C., Chen, C.-M., Huang, P.-L., Hwang, C.-C.: Equality: a flexible topology system ready for high performance computing. *IEEE Trans. Parallel Distrib. Syst.* (2018) (to be confirmed)
9. Leiserson, C.E.: Fat-trees: universal networks for hardware-efficient supercomputing. *IEEE Trans. Comput.* **100**(10), 892–901 (1985)
10. Al-Fares, M., et al.: A scalable, commodity data center network architecture. In: *ACM SIGCOMM Computer Communication Review*, vol. 38
11. Introducing data center fabric, the next-generation Facebook data center network. Andreyev (2014) [Online]. Available: <https://code.facebook.com/posts/360346274145943/>

12. Klenk, B., Fröning, H.: An overview of MPI characteristics of exascale proxy applications. In: International Supercomputing Conference. Springer, Cham (2017)
13. Jiang, N., et al.: A detailed and flexible cycle-accurate network-on-chip simulator. In: 2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)
14. Pfister, G.F., Norton, V.A.: “Hot spot” contention and combining in multistage interconnection networks. *IEEE Trans. Comput.* **C-34** (10) (1985)



Energy Aware LEACH Protocol for WSNs

Nahid Ebrahimi Majd^(✉), Pooja Chhabria, and Anjali Tomar

Department of Computer Science and Information Systems, California State
University San Marcos, San Marcos, USA
nmajd@csusm.edu, {chhab002, tomar001}@cougars.csusm.edu

Abstract. LEACH is a clustering routing protocol in wireless sensor networks. This protocol frequently forms clusters of nodes and selects one of the cluster members as the cluster head. The two-tier structure of LEACH divides the network into two layers: (1) cluster members, which collect raw data from the environment and (2) cluster heads, which receive the collected data from cluster-members, fuse the received data and transmit that to the base station. LEACH reduces the energy consumption of nodes since the most energy-consuming task, which is the long-distance transmission to the base station takes place only in cluster heads. LEACH selects the cluster heads uniformly using a probabilistic calculation. However, there are other parameters that affect the energy consumption of nodes. As a result, the selected nodes may not be that strong to handle the high workload of a cluster head; their energy will be depleted very soon; the network lifetime and throughput decline. To overcome this tradeoff, in this paper, we propose EA-LEACH (Energy Aware LEACH), a new clustering routing protocol, which selects cluster heads using their residual energies. This method provides an appropriate selection of cluster heads that are strong enough to handle the expected workload of cluster heads to reach high throughput. We validated the effectiveness and efficiency of our protocol through simulations. The analysis of our results shows that the cluster heads selected by our proposed protocol prolong the network lifetime by 60% in comparison to those selected by LEACH. As a result, the nodes transmit significantly more amount of data during their lifetime.

Keywords: LEACH · Wireless sensor networks · Adaptive clustering · Energy efficiency · Routing protocols

1 Introduction

A wireless sensor network (WSN) is a spontaneous formed cooperative wireless network consisting of several sensor nodes spatially scattered across the network. WSNs are often used to observe and monitor a desired physical condition. The nodes of a WSN possess various processing capabilities and contain sensors, transceivers, power sources, and memory. While a wireless communication takes place among the nodes, a gateway is established as a connection between the wireless network and the end user. This gateway is named the base station. A WSN can consist of numerous nodes ranging from hundreds to thousands, depending on the type and functionality of the network.

These networks are ad hoc since they have no pre-existing infrastructure. Routing protocols, which select a path to transmit data from sensor nodes to the base station, play vital role in establishing an efficient communicating wireless network.

Clustering is a conventional methodology to improve the efficiency of routing protocols in WSNs. The network forms several clusters, each contains a cluster head and a group of cluster members associated with the head. The cluster head collects data from cluster members, aggregates (fuses) them and transmit the aggregated to the base station. Thus, clustering reduces the energy consumption of sensor nodes in the network as it cuts down on direct data transmission from each sensor node to the base station. Also, the cost of local processing to aggregate data at cluster head is much less than the cost of separate transmissions of raw data. Since the data aggregation and transmission to the base station is performed only at the cluster heads, the energy of the member nodes is conserved.

Data aggregation, also known as data fusion, combines multiple data signals and create a more accurate signal. It enhances the common signal and reduces the uncorrelated noises. The method of data aggregation depends on the application. For instance, beamforming algorithm is used for acoustic signals aggregation.

The energy dissipated to transmit data from a member to the short distance cluster head is much less than direct transmission to the base station. If the cluster heads were chosen fixed throughout the network lifetime, those nodes would die quickly due to the high load of data transmission and energy loss. This would end the valuable lifetime of the rest of the member nodes of that cluster.

Low Energy Adaptive Clustering Hierarchy (LEACH) protocol [1, 2] selects the cluster heads uniformly based on a defined probability function and changes the cluster heads over the rounds. Thus, in each round a new set of nodes are selected to be the cluster heads, which increases the overall network lifetime in comparison to direct transmission, meaning the network nodes remain active for a longer time.

SEP (Stable Election Protocol) [3], ESEP (Enhanced Stable Election Protocol) [4], and TSEP (Threshold-Sensitive Stable Election Protocol) [5] proposed clustering protocols based on LEACH. These protocols assume different levels of initial energies on the nodes. SEP assumes two levels of initial energies and uses normal and advanced nodes. ESEP adds an intermediate level of energy and uses intermediate nodes as well. TSEP uses the three levels of ESEP but enhances the performance by a pre-process at the source and reducing the noise at cluster members. Our proposed method assumes the same level of energy for all nodes. Thus, we compare our method with LEACH, which uses the same assumption.

Some recently polished papers explored energy efficiency in LEACH protocol. EE-LEACH (Energy Efficient LEACH) [6] proposed a protocol that uses the conditional probability theorem to aggregate data based on spatial density function and forms clusters based on the energy availability in the neighborhood. Although the results show less energy consumption in the network, the complexity of this technique is high, which results in lack of scalability and data integrity. EEM-LEACH (Energy-Efficient Multi-hop LEACH) [7] forms clusters based on residual energy and energy consumption at each node. In this protocol, cluster heads send data to the base station through other cluster heads. CL-LEACH (Cross Layer LEACH) [8] also proposes an enhanced multi-hop routing technique that detects the broken links and substitute them

with new paths. Multi-hop routing in these two protocols [7, 8] significantly reduces the energy consumption. However, it highly increases the message overhead and the time required for the algorithms to converge.

The remaining of the paper is organized as follows. Section 2 describes the LEACH protocol. Section 3 explains the proposed protocol. Section 4 presents simulations setup and results. Section 5 presents the conclusion.

2 LEACH Functionality

In LEACH, each round contains two phases: Set up phase and Steady state phase. The clusters are formed in the first phase, and the data is transmitted in the second phase. In the set up phase, each node runs the predefined probability function, and if selected as the cluster head, advertises to the network. Then each non-cluster head node associates with the closest cluster head as a member of that cluster. Each node becomes a cluster head once in an epoch of $1/P$ rounds. P is the desired percentage of nodes selected to be cluster heads. The probability function works as follows: each node generates a random number between 0 and 1. If it is less than threshold $T(n)$, given in Eq. (1), then the node n self-elects to be a cluster head. In this equation, r stands for the current round and G represents the set of nodes that have not been cluster heads in previous rounds.

$$T(n) = \begin{cases} \frac{P}{1-P*(r \bmod \frac{1}{P})} & \text{if } n \in G \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

G is a set of nodes that has not been selected to be cluster heads in the current epoch. If node n does not belong to G , i.e. the node has already been selected as a cluster head in the previous rounds in the current epoch, it cannot be chosen as a cluster head until all the remaining nodes are all selected as the cluster heads in the same epoch. The function exponentially increases the probability for the remaining nodes to become cluster heads. At the very last round of epoch, the function value will be one for all remaining nodes and they become cluster heads.

Once clusters are formed, members transmit their data to their cluster heads using a TDMA protocol. Once a cluster head receives data from all members, it aggregates them and sends the aggregated data to the base station.

The radio model of nodes dissipates $E_{elec} = 50$ nJ/bit to run the transmitter/receiver circuitry. The amount of energy consumed to receive k bits data is

$$E_{RX}(k) = k * E_{elec} \quad (2)$$

and the amount of energy consumed to transmit k bits data to a node in distance d is

$$E_{Tx}(k, d) = \begin{cases} k.E_{elec} + k.E_{fs}.d^2 & \text{if } d \leq d_0 \\ k.E_{elec} + k.E_{mp}.d^4 & \text{if } d > d_0 \end{cases} \quad (3)$$

where d_0 is a threshold to separate two channel models: free space routing (fs) for short distance transmissions and multipath routing (mp) for long distance transmissions. The

E_{fs} and E_{mp} are the energy dissipations in these two channels respectively (Table 1). It is expected that free space model is required for in-cluster transmissions and multi path model for far transmissions between cluster heads and base station. The threshold d_0 is defined as

Table 1. Parameter settings

Parameters	Values
Topology area	400 m × 400 m
Number of nodes (n)	100
Base station position	200 m × 200 m
Percentage of desired cluster heads in LEACH (P)	0.1
Energy supplied to each node	0.5 J
Energy spent by the radio for transmission/reception (E_{elec})	50 nJ/bit
Data aggregation energy (EDA)	5 nJ/bit
Maximum number of rounds	2500
Energy dissipated in free space routing (E_{fs})	10 pJ/bit/m ²
Energy dissipated in multi path routing (E_{mp})	0.0013 pJ/bit/m ⁴
Packet size	4000 bits

$$d_0 = \sqrt{\frac{E_{fs}}{E_{mp}}} \quad (4)$$

Each node becomes a cluster head once in an epoch. In this way, LEACH distributes the cluster head workload among nodes. However, the important impact of distance to the base station is neglected here. The cluster heads that are close enough to the base station to use free space channel consume much less energy to transmit data to the base station (in the order of d^2) in comparison to the further cluster heads that need to use multipath channel (in the order of d^4). After a few rounds, this phenomenon causes a huge gap between the residual energies of nodes. To address this problem, we proposed EA-LEACH, which considers the residual energies of nodes in the cluster head selection function.

3 Proposed Protocol

Our proposed protocol EA-LEACH (Energy Aware LEACH), uses the following threshold function

$$T(n) = \begin{cases} \frac{P}{1-P*(r \bmod \frac{1}{P})} * \frac{E(n)}{E(total)} & \text{if } n \in G \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where $E(n)$ is the residual energy of node n and $E(\text{total})$ is the total residual energies of all nodes in the network. Each cluster member calculates the expected residual energy it will have after transmission to the cluster head. The member sends this value and its data packet to its cluster head. Then, each cluster head calculates the expected residual energy it will have after transmission to the base station, adds that to the total residual energies of all members in the cluster, and sends that with the data packet it transmits to the base station. At the end of each round, the base station calculates the total residual energy of the network and broadcasts that to the network nodes.

The proposed threshold function increases the probability for the nodes with higher residual energies to become cluster heads. Thus, those nodes will be selected more frequently as cluster heads. The cluster heads close to the base station dissipate less energy to transmit data to the base station. Thus, their residual energies will be more, and be selected more frequently as cluster heads. It is also important to consider the energy dissipated by member nodes. Since E_{elec} is relatively high, a node that has been a member for several rounds, has dissipated a large amount of energy, which is close to the energy required at a cluster head. This issue is also neglected in LEACH, where we addressed with our new threshold function.

4 Simulations and Discussions

To model and simulate the proposed protocol, we used MATLAB. We evaluated the performance of EA-LEACH based on the network lifetime, residual energy and throughput. The metrics that describe the performance of network are:

- (1) Number of alive nodes per round.
- (2) Number of dead nodes per round.
- (3) Number of cluster heads per round.
- (4) Number of packets sent to the base station.
- (5) Average residual energies of the nodes.

We Simulated a network consisting 100 nodes uniformly distributed in region of 400×400 . The base station is in the center of network. The desired percentage of cluster heads (P) is 10%. Table 1 presents the parameters we used in our simulations.

Figure 1 shows the cluster formation of network. Asterisks and dots represent heads and members respectively. Figures 2 and 3 show the number of alive and dead nodes in each protocol respectively. The figures show that although the first node dies earlier in EA-LEACH, it outperforms LEACH by 60% in terms of network lifetime, meaning the last node dies much later in EA-LEACH, and the network remains stable for significantly more amount of time. The reasons are:

- (1) EA-LEACH more frequently selects cluster heads from nodes with high residual energies. However, LEACH does not consider the amounts of residual energies at nodes and selects cluster heads among nodes with the same frequency;

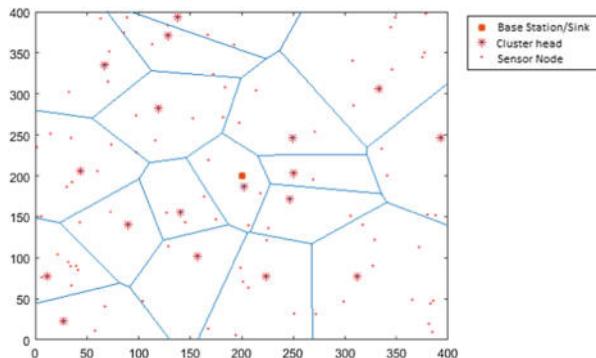


Fig. 1. Cluster formation

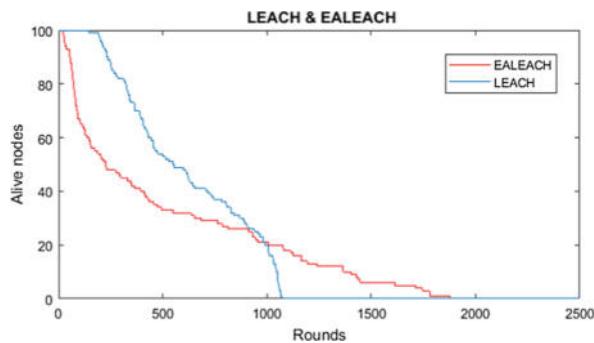


Fig. 2. Number of alive nodes per round

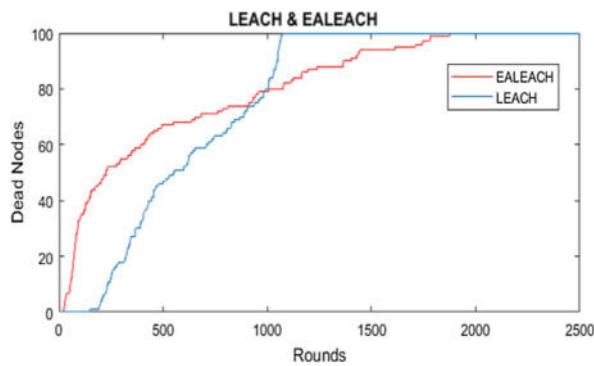


Fig. 3. Number of dead nodes per round

(2)

EA-LEACH avoids long distance transmissions to the base station by selecting cluster heads from nodes that are closer to the base station. The further nodes still get

selected, but with a lower frequency. This conserves a lot of energy in the network. Notice that if a node transmits to a distance further than d_0 , the transmission cost switches from the d^2 (free space routing) to d^4 (multipath routing);

- (3) In EA-LEACH, the cluster head selection function tends to select almost the same number of cluster heads during the simulation, which is on an average less than the number of cluster heads in LEACH. Also, LEACH tends to reduce the number of cluster heads when the nodes die during the time. Therefore, In LEACH, at the beginning of simulation that number of cluster heads is high, a lot of energy is dissipated in the nodes. Thus, as shown in Fig. 2, almost all nodes reach a low residual energy after a certain number of rounds and quickly die after first node dies.

Figures 4 and 5 show the number of cluster heads for the two protocols in each round respectively. LEACH selects a large number of cluster heads at the beginning and the average number of cluster heads decreases during the time. LEACH is not aware of nodes residual energies. On an average, it always selects the desired percentage of nodes to be cluster heads. As the nodes die during the time, the number of nodes decreases, and the protocol selects a smaller number of cluster heads to keep the desired percentage the same. The strength of EA-LEACH is the fact that it selects the strongest nodes to perform the duties of cluster heads at each round. This significantly improves the processing and transmission capacity of network at that round. At the next round, that node might not be strong enough to perform the heavy duties of a cluster head, thus EA-LEACH switches to other nodes and selects the strongest ones as the new cluster heads. As a result, not that many cluster heads are required at each round, and the network lasts for a longer period of time.

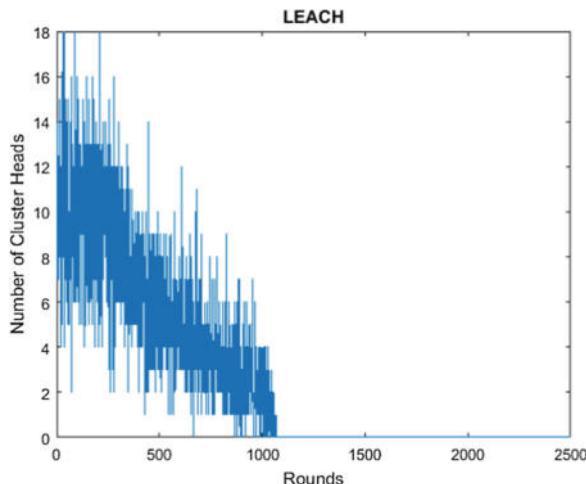


Fig. 4. Number of cluster heads per round (LEACH)

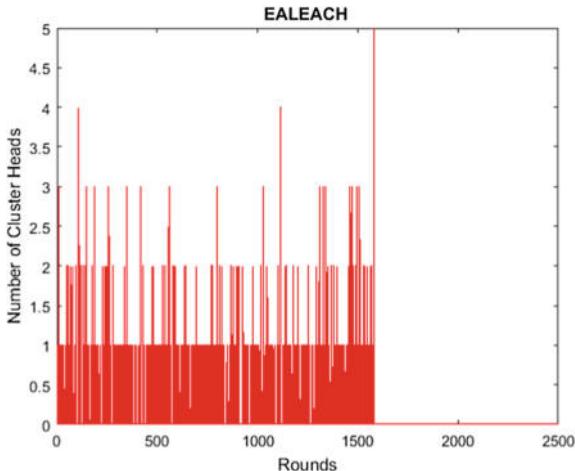


Fig. 5. Number of cluster heads per round (EA-LEACH)

Figure 6 depicts the cumulative number of packets sent from the cluster head to the base station. The network throughput of EA-LEACH is significantly higher than LEACH. The data transmission in LEACH stops at around 1100 rounds when all the nodes die. For EA-LEACH, the data transmission to the base station lasts for almost 1800 rounds as the nodes are still alive in EA-LEACH. The longer lifetime provided in EA-LEACH results in significant more throughput in comparison to LEACH.

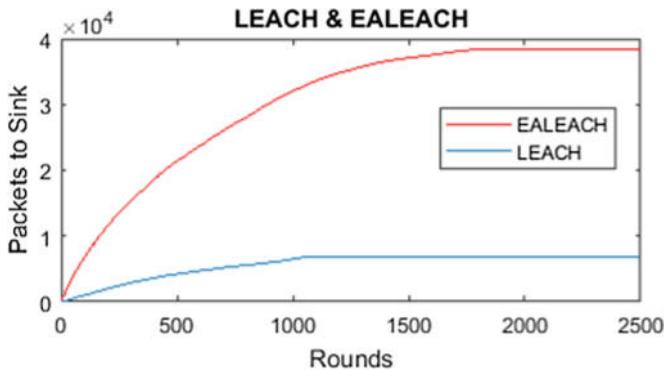


Fig. 6. Packets transmitted to base station per round

Figure 7 shows the average residual energies of nodes in the network for the two protocols. Referring to Figs. 1 and 2, the first node dies later in LEACH, but then all the nodes die quickly. That is because the nature of LEACH that uniformly dissipate energy in all nodes. We see the same effect in Fig. 7. The average energy for LEACH remains high for a while, and then quickly falls. However, in EA-LEACH, the nodes

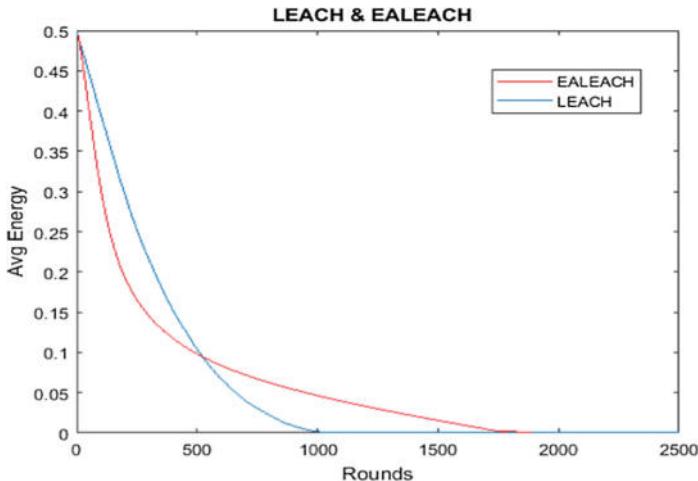


Fig. 7. Average energy per round

average energy initially declines, but with the passage of rounds, the network gains stability and the residual energy lasts until almost 1800 rounds, which is the long network lifetime in EA-LEACH.

5 Conclusion and Plans for Future

In this paper, EA-LEACH, an energy aware adaptive clustering protocol for WSNs is proposed where cluster heads are selected according to their residual energies. This feature of the protocol significantly improves the network lifetime and the amount of data reached the base station.

Our plan for future work is to propose a method in which the first node dies later. When we use the probability function, all the nodes must become a cluster head at least once in an epoch. Even if a node is not that strong to handle the high workload of cluster head, it is still selected once per epoch. Therefore, their batteries are quickly depleted, and they die soon. An approach to handle this issue is adjusting the probability function with the distance parameter. Other approach could be adding a condition where a node cannot be a cluster head in the current epoch if its energy is below some threshold.

References

- Heinzelman, W.R., Chandrakasan, A.P., Balakrishnan, H.: Energy-efficient communication protocols for wireless microsensor networks. In: Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33) (January 2000)

2. Heinzelman, W.R., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. *IEEE Trans. Wireless Commun.* **1**(4), 660670 (2002)
3. Smaragdakis, G., Matta, I., Bestavros, A.: SEP: a stable election protocol for clustered heterogeneous wireless sensor networks. In: Second International Workshop on Sensor and Actor Network Protocols and Applications (SANPA 2004) (2004)
4. Aderohunmu, F.A., Deng, J.D.: An enhanced stable election protocol (SEP) for clustered heterogeneous WSN. Department of Information Science, University of Otago, New Zealand (2009)
5. Kashaf, A., Javaid, N., Khan, Z.A., Imran, A.K.: TSEP: threshold-sensitive stable election protocol for WSNs. In: IEEE International Conference of Frontiers of Information Technology (2012)
6. Arumugam, G.S., Ponnuchamy, T.: EE-LEACH: development of energy-efficient LEACH protocol for data gathering in WSN. *EURASIP J. Wireless Commun. Netw.* **2015**(1), 1–9 (2015)
7. Antoo, A., Mohammed, A.R.: EEM-LEACH: energy efficient multi-hop LEACH routing protocol for clustered WSNs. In: Proceedings International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 812–818 (July 2014)
8. Marappan, P., Rodrigues, P.: An energy efficient routing protocol for correlated data using CL-LEACH in WSN. *Wireless Netw.* **22**(4), 1415–1423 (2016)



Evaluation of Parameters Affecting the Performance of Routing Protocols in Mobile Ad Hoc Networks (MANETs) with a Focus on Energy Efficiency

Nahid Ebrahimi Majd[✉], Nam Ho, Thu Nguyen, and Jacob Stolmeier

Department of Computer Science and Information Systems, California State University San Marcos, San Marcos, USA
nma.jd@csusm.edu, {ho047, nguyen007, stolm003}@cougars.csusm.edu

Abstract. This paper provides a comprehensive evaluation and analysis of two classes of routing protocols optimized for MANETs: reactive (DSR, AODV), and proactive (DSDV, OLSR). These protocols are particularly designed for the dynamic nature of MANETs where nodes actively move, connections between nodes are regularly broken, and paths need to be reconstructed. MANETs are used in a wide range of applications including rescue operations, military areas, and oceanography. However, the nodes have limited batteries, and typically, it is not applicable or cost effective to replace the batteries of nodes. Other than the dynamic nature of these networks, which leads to more processing and rerouting requirements, the movements also speed up the batteries' depletions. Thus, the energy efficiency of routing protocols significantly affects the performance of these networks. In this research, we study the two classes of routing protocols in MANETs, investigate the effect of several parameters on the network performance through excessive simulations, and analyze how the variations of these parameters affect the performance. We evaluate the network performance in terms of energy consumption, routing overhead and Quality of Service metrics including throughput and delay.

Keywords: Routing protocol · DSR · AODV · DSDV · OLSR · Mobile ad hoc network · Performance · Energy efficiency

1 Introduction

There are two classes of routing protocols in MANETs: reactive and proactive protocols. While proactive protocols periodically broadcast routing messages to keep the routing tables updated on the nodes, reactive protocols discover new paths only if an existing connection is broken and a new path is required. Therefore, reactive protocols save bandwidth and energy by eliminating the need for periodical floods. However, these protocols need more time to transmit data if a path construction is required, which may cause longer delays and less throughput.

The remaining of the paper is organized as follows. Section 2 describes the routing protocols evaluated in this paper. Section 3 explores the related work, evaluated the energy efficiency of these protocols. Section 4 presents simulations setup and results. Section 5 presents the results and discussion. Section 6 presents the conclusion and plans for future.

2 Routing Protocols

DSR, Dynamic Source Routing [1], is a reactive protocol, which uses shortest hop forwarding paths to route the packet to the destination. This protocol works in three phases: (1) discover a route when no path exists, or the current path is broken, (2) record routes on nodes, and (3) if a path is broken, broadcast a broken route message to inform the other nodes that the route is not available anymore. This protocol is proven to perform well in static environments. Low nodes mobilities leads to lower connection breaks and route reconstructions. However, the performance degrades rapidly when mobility increases since the links break regularly, and the protocol needs to reconstruct the paths, which causes long delays. Also, the protocol has no mechanism to remove the old paths information cashed in the nodes routing tables, which causes inconsistencies in the rerouting process.

AODV, Ad hoc On-Demand Distance Vector [2], is a reactive distance vector routing protocol, which uses shortest hop forwarding method to find paths. AODV eliminates three drawbacks in the DSR design: (1) DSR carry the path in the packet header, thus the header will be very large when the path is long. However, AODV stores the information to access the paths in the intermediate nodes tables, thus the packet header is always in the standard size; (2) DSR keeps all the stale paths even if new paths have been constructed due to nodes movements and broken links. However, AODV assigns timestamps to the paths and discards the old route information from the nodes routing tables. Thus, it always uses fresh paths; and (3) In DSR, there is no way to check whether the same route reply message sent by a node has arrived multiple times in a loop. However, AODV uses sequence numbers to avoid recording stale information that has been in loops.

DSDV, Destination Sequenced Distance Vector [3], is a proactive distance vector routing protocol optimized for MANETs. This protocol solves the routing loop problem (count to infinity) using incremental updates of sequence numbers. DSDV proactively updates all the paths, thus the paths to all destinations are always known. However, it regularly updates the routing tables even if the network topology has not changed. This significantly increases the bandwidth and energy consumption in the network. On the other hand, if the network is highly dynamic and the links regularly break, DSDV required to create several new sequence numbers before it can re-converge the paths. Thus, DSDV has a high amount of routing overhead in a dynamic network, which consumes significant amount of energy and bandwidth.

OLSR, Optimized Link State Routing [4], is a proactive link state routing protocol. A link state protocol may cause problems like receiving the same HELLO message multiple times on a node. However, OLSR avoids unnecessary rebroadcasts that may lead to such redundancies. In OLSR, each node purposefully limits the number of

neighbors that can rebroadcast its HELLO messages. Therefore, other nodes receive this node's HELLO message only once. This mechanism of OLSR significantly reduces the amount of routing overhead and energy consumption. However, a link state protocol needs complete information about the network topology, and the cost to collect and process that much data will significantly increase when the number of nodes is large.

3 Related Work

Many researchers have evaluated and compared routing protocols in MANETs [5, 6] but only some of them have studied the effect of routing protocols on energy consumption of nodes, and how the energy consumption pattern of each protocol affects the performance of network. Also, there is a need for a comparative evaluation of proactive versus reactive classes of protocols based on both energy and Quality of Service metrics.

In [7], the authors compared two reactive routing protocols DSR and AODV in MANETs in terms of energy consumption for different scenarios.

In [8], the authors compared four routing protocols DSR, AODV, DSDV, and OLSR for MANETs in terms of energy consumption for a variety of scenarios including increasing number of nodes, rates of traffic sources, number of connections, speeds, and pause times. Their results showed that DSR always consumes much less energy than other protocols, and proactive protocols consume more energy than reactive ones especially in high speeds. Also, among proactive protocols, DSDV generally consumes more energy than OLSR unless the number of nodes is high.

In [9], the authors compared two reactive (DSR, AODV), one proactive (DSDV), and one hybrid (ZRP) protocols in terms of energy consumption and also a dual metric defined in the paper based on residual energy and amount of data successfully delivered at destination. Their results proved that DSR outperforms other protocols in terms of throughput and energy consumption especially in a network with extremely large number of nodes. This research does not investigate the performance of protocols in terms of delay.

In [10], the authors studied the performance of three protocols DSDV, DSR and AODV for MANETs in terms of energy consumption, throughput and packet delivery ratio. Their simulations studied exponential traffic generators where packets are generated at a constant burst rate when the generator turns on. In this way they combined exponential and constant bit rate traffic generators and studied the performance of network for this particular type of traffic. The results showed that for this type of traffic (1) DSR has a better performance in terms of throughput and packet delivery ratio, (2) DSDV outperforms the other two protocols in terms of energy utilization, and (3) AODV shows higher throughput.

The research on the evaluation of performance of routing protocols in MANETs is limited to the effect of different scenarios on energy consumption and throughput. To the best of our knowledge, there has been no comprehensive research on the evaluation of performance of these protocols in all aspects containing energy consumption,

throughput, end-to-end delay and routing overhead. This was our motivation for this research.

4 Simulations, Parameters and Metrics

We used network simulator ns2 for our modeling and evaluations. In our simulations, we assumed there is one base station, which is the destination of all traffics generated in the network. The nodes are deployed and move in the field in a uniform distribution. We investigated a variety of network densities assuming all nodes are sources of data constantly generate CBR traffic. A summary of the parameters used in our simulations is given in Table 1.

Table 1. Simulation parameters

Parameter	Value
Network area	750 × 750 m
Number of nodes	10-20-30-40-50
Antenna model	Omni antenna
Radio-propagation model	Two ray ground
Mobility model	Random waypoint
MAC layer	IEEE 802.11
Interface queue type	Drop tail
Queue length	50
Communication range	250 m
Simulation time	150 s
Traffic type	CBR
CBR interval	0.125 s
Packet size	500 bits
bandwidth	1 Mbps
Initial energy	250 J
Idle power	1 W
Sleep power	0.001 W
Transition power (sleep to idle)	0.2 W
Rx power	1 W
Tx power	5 W
Pause time	85 s
Moving speed of nodes	15 m/s

We analyzed the performance of network in a purposeful series of scenarios to find out the best performance that could be achieved under different circumstances including network density, nodes speeds, pause time and transmission power. We evaluated the network performance based on several metrics including energy consumption, routing overhead, throughput, and delay. For each of these metrics, we

measured the amount periodically in each 0.5 s and then calculated the average values for 150 s simulation. Then we calculated the standard deviation to quantify the amount of variations in the set of data values. The standard deviation bars are depicted in the bar charts.

Average consumed energy at each time interval 0.5 s is defined as sum of all energy consumed in nodes over the number of nodes n.

$$\text{Avg consumed } E = \frac{\sum_{i=1}^n (\text{last } E_i - \text{current } E_i)}{n} \quad (1)$$

The average energy consumption of network displayed on the charts is the average of all these values measured for each 0.5 s time interval.

The average throughput in each interval is defined as the total amount of data arrived at the destination during that time interval over the time interval duration.

$$\text{Avg Throughput} = \frac{\sum_{j=1}^x \text{PacketSize}_j}{\text{time interval}} \quad (2)$$

The average end-to-end delay in each interval is defined as sum of (received time – sent time) for all x data packets reached the destination during that time interval over the total number of those packets x.

$$\text{Avg Delay} = \frac{\sum_{j=1}^x \text{RcvdTime}_j - \text{SentTime}_j}{x} \quad (3)$$

We define a new metric to analyze the ratio of routing overhead required to provide the achieved throughput. This ratio shows how much routing information has been exchanged in the network to reach the achieved throughput. If this overhead ratio is almost one, it means the routing protocol needs to exchange extremely high amount of routing information such that the overall throughput is negligible in comparison to that. On the other hand, if this ratio is almost 0.5, it means the routing protocol exchanges almost the same amount of routing information as the overall throughput. A reactive routing protocol might have lower overhead ratio, consume less bandwidth at links and less energy at nodes, however be slower to find a path or not successful to deliver all the packets to the destination when the links are broken.

$$\text{Routing Overhead ratio} = \frac{\text{routing packets}}{\text{routing packets} + \text{throughput}} \quad (4)$$

5 Results and Discussion

At the first step, we investigate the energy consumption of the four routing protocols for a range of network densities: 10, 20, 30, 40, 50 nodes in the same size field. Figure 1 shows that the average energy consumptions of nodes decline from low

density to high density, and their reduction follow a similar pattern. For instance, from 10 nodes to 50 nodes density, the average energy consumption of DSR, AODV, DSDV, and OLSR decline 32, 32, 27, and 25% respectively. The reason is in a dense network where all nodes generate traffic, more alternative paths are also available. Thus, once a link is broken due to nodes movements or other phenomena, the nodes can find a new path with a minimum effort at each node. It is also important to note that Fig. 1 shows the average energy consumption of nodes in the network, and the total energy consumption of all nodes exponentially grow when the number of nodes increase.

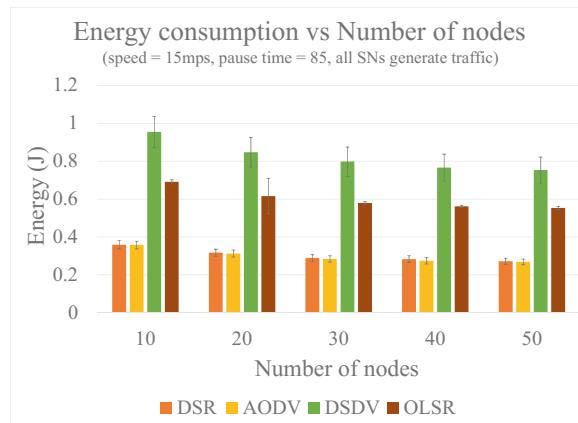


Fig. 1. Energy consumption

Also, as expected, proactive protocols consume more energy than reactive protocols with a similar pattern. For instance, in the lowest density when there are 10 sources of traffic in the network, DSDV and OLSR consume 2.6 and 2 times more than reactive protocols respectively, and in the highest density when there are 50 sources of traffic in the network, DSDV and OLSR consume 2.7 and 2 times more than reactive protocols respectively. The main reason is proactive protocols constantly exchange routing tables among nodes to keep the tables updated. However, reactive protocols update the routes only when an existing connection in a path is broken, and a new path is required. As a result, we have higher energy consumptions and routing overheads in proactive protocols. It affects the throughput and end-to-end delays of the network, which we will investigate more in the next steps.

Figure 1 also shows better performance for OLSR in comparison to DSDV in terms of energy consumption. Again, the pattern is almost the same from low density to high density networks. For instance, DSDV consumes 38 and 36% more energy than OLSR in a network of 10 and 50 nodes respectively. The reason is the Multipoint Relay Broadcasting method used in OLSR, where each node transmits the routing messages to only a selected group of neighbors sufficient to distribute routing information to the entire network. In this way, OLSR reduces the amount of routing overhead required to establish new paths and distribute the load on the nodes in a uniform manner.

We reach to a conclusion that when all nodes are sources of traffic, proactive protocols always consume more energy than reactive ones, and among proactive protocols, DSDV consumes more than OLSR. We investigated this conclusion under different circumstances with different nodes' movements speeds, simulation's pause times (the moment the nodes start to move in the simulation), and transmission powers (power required to transmit one unit of data) and achieved similar results with the same patterns.

Figure 2 compares throughputs of the four protocols for different network densities. In low densities, the gap between the throughputs of four protocols is low. For instance, when there are only 10 nodes in the network, DSR has 24% more throughput than DSDV. The maximum performance gap occurs in a dense network of 50 nodes, where proactive protocols have 20 and 5 times more throughput than DSR and AODV protocols respectively. The reason is in a dense network with high amount of traffics and movements reactive protocols cannot efficiently reroute, and packets get lost.

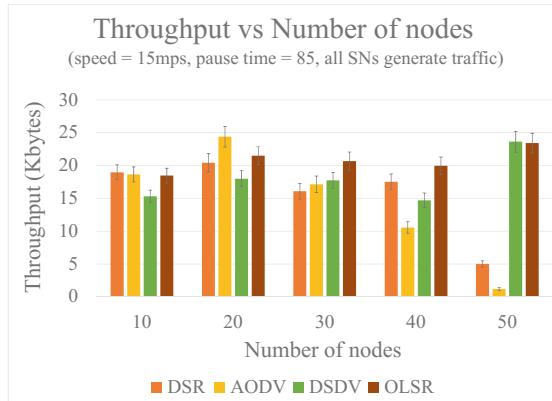
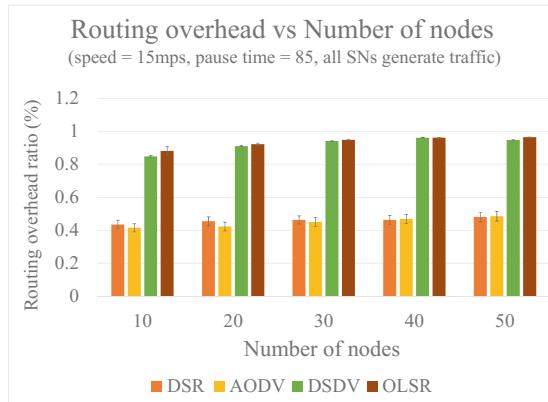
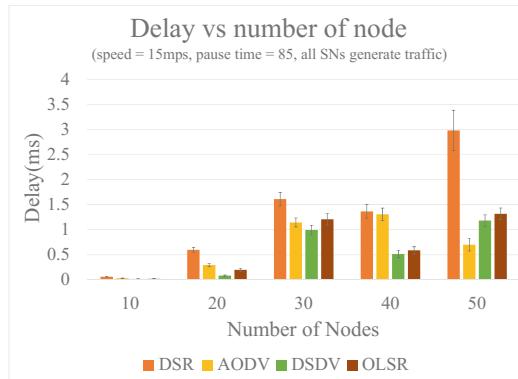


Fig. 2. Throughput

Proactive protocols appropriately handle this situation, find the best paths before the packets reach the bottlenecks and forward the packets through proper paths towards the base station. Thus, they exchange more routing messages to keep the routing tables updated all the time. Figure 3 shows the routing overhead ratios of the protocols. The figure shows that proactive protocols have higher (almost twice) overhead ratios than reactive ones. Also Fig. 4 shows that proactive protocols have relatively lower end-to-end delays in most cases. Particularly, in a network of 50 nodes, the delay of DSR is almost 2.5 times the delay of proactive protocols.

Also, among reactive protocols, AODV has less delay than DSR, which is due to the sequence numbers that AODV uses to eliminate stale routing information and loops while DSR has no mechanism to avoid them.

We reach to a conclusion that in a dense network with high amount of generated traffic, OLSR has a better performance than DSDV. Although OLSR has a slightly

**Fig. 3.** Routing overhead**Fig. 4.** Delay

more end-to-end delay, and almost the same throughput and routing overhead, its energy consumption is 36% less than DSDV, which is a great improvement in performance. Also, in these networks, proactive protocols have a better performance than reactive ones. Although their energy consumption is almost twice reactive ones, their throughputs are significantly higher (5 and 20 times more) and their delays are also relatively low. Also, among DSR and AODV, DSR has a higher throughput and delay. Thus, depends on the application, AODV might be selected as a balance with reasonable throughput and delay while consumes almost the same energy.

On the other hand, the results show that in a network with low number of nodes and traffic, there is no significant difference between performances of reactive and proactive protocols in terms of either throughput or end-to-end delay, however, both energy consumptions and routing overhead ratios of proactive protocols are at least twice reactive protocols. We reach to a conclusion that in these networks, reactive protocols have better performance in terms of energy consumptions and routing overheads while

there is no significant difference between proactive and reactive protocols in terms of throughput or delay.

6 Conclusion and Plans for Future

When all nodes are sources of traffic, proactive protocols always consume more energy than reactive ones, and among proactive protocols, DSDV consumes more than OLSR. If the network is dense, the proactive protocols have significantly more throughput and relatively low delay. In such applications OLSR provides a balance towards the best performance since its energy consumption is less. However, in a low-density network, reactive protocols are more efficient, and AODV with less delay and more throughput has the best performance.

Our plan for future is to propose and evaluate a hybrid protocol that eliminates the drawbacks of the current protocols and improves the network performance and lifetime. Proactive protocols always keep the paths updated while Reactive protocols update the routes only when a new path is required. Therefore, the former suffers from redundant routing information frequently distributed in the network, and the latter has long delays due to lack of updated routes. A hybrid protocol would estimate the network required paths and purposefully distribute enough routing information sufficient for rerouting with a short delay.

References

1. Johnson, D., Hu, Y., Maltz, D.: The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4. No. RFC 4728 (2007)
2. Perkins, C., Belding-Royer, E., Das, S.: Ad hoc On-Demand Distance Vector (AODV) Routing. No. RFC 3561 (2003)
3. He, G.: Destination-Sequenced Distance Vector (DSDV) Protocol. Helsinki University of Technology, Networking Laboratory (2002)
4. Clausen, T., Jacquet, P.: Optimized Link State Routing Protocol (OLSR). No. RFC 3626 (2003)
5. Mohapatra, S., Kanungo, P.: Performance analysis of AODV, DSR, OLSR and DSDV routing protocols using NS2 simulator. In: Procedia Engineering, pp. 69–76 (2012)
6. Sharma, A., Kumar, R.: Performance comparison and detailed study of AODV, DSDV, DSR, TORA and OLSR routing protocols in ad hoc networks. In: 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), pp. 732–736 (2016)
7. Barati, M., Atefi, K., Khosravi, F., Daftari, Y.A.: Performance evaluation of energy consumption for AODV and DSR routing protocols in MANET. In: Computer & Information Science (ICCIS), 2012 International Conference, vol. 2, pp. 636–642. IEEE (2012)
8. Er-Rouidi, M., Moudni, H., Mouncef, H., Merbouha, A.: An energy consumption evaluation of reactive and proactive routing protocols in mobile ad-hoc network. In: Computer Graphics, Imaging and Visualization (CGIV), 2016 13th International Conference, pp. 437–441. IEEE (2016)

9. Ourouss, K., Naja, N., Jamali, A.: Efficiency analysis of MANETs routing based on a new double metric with mobility and density models. In: Computer Systems and Applications (AICCSA), 2016 IEEE/ACS 13th International Conference of, pp. 1–8. IEEE (2016)
10. Razouqi, Q., Ahmed, B., Mohamed, G.: Performance analysis for diverse simulation scenarios for DSDV, DSR and AODV MANET routing protocols. In: Computer Engineering Conference (ICENCO), 2017 13th International, pp. 30–35. IEEE (2017)



Integrating User Opinion in Decision Support Systems

Saveli Goldberg^{1(✉)}, Gabriel Katz², Ben Weisburd³,
Alexander Belyaev⁴, and Anatoly Temkin⁴

¹ Massachusetts General Hospital, Boston, MA 02114, USA
sigoldberg@mgh.harvard.edu

² Massachusetts Institute of Technology, Cambridge, MA 02142, USA
gabkatz@gmail.com

³ Broad Institute, Cambridge, MA 02142, USA
weisburd@broadinstitute.org

⁴ Boston University Metropolitan College, Boston, MA 02215, USA
{abelyaev, temkin}@bu.edu

Abstract. We propose an approach to decision support systems (DSS) that starts with the user first making their own unassisted decision α_U and providing this as an input to the algorithm. Then, if the algorithm disagrees with the user's initial decision, it iteratively works with the user to converge on a common decision or at least make the user reconsider input values that are inconsistent with α_U . We provide a detailed description of this approach along with examples, and then discuss potential benefits and limitations.

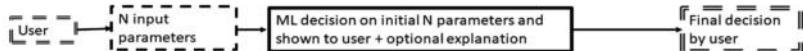
Keywords: Decision support · Machine learning · User interaction

1 Introduction

With rapid progress across a broad range of machine learning (ML) applications in recent years [1–4], some implications of these advances are also causing concern. One set of issues that may arise as people increasingly rely on these systems is that they diminish the users' sense of responsibility for decisions and outcomes [5], and that by reducing the need for human expertise, they gradually lead to a loss of human expertise in certain important areas. Current approaches to addressing these issues focus on improving the explainability of decisions generated by ML algorithms [1, 6] or by requiring that humans confirm or approve ML decisions. These measures are indeed very helpful, but the better ML systems become, the more likely it will be that users will stop putting much effort into analyzing or critically evaluating the algorithms' decisions, even if automated explanations are also provided.

Here we propose an alternate approach that mitigates the potential loss of expertise and restores a fuller sense of responsibility to users—Fig. 1. This approach introduces additional steps to a user's interaction with the decision support system. One key aspect is that it requires the user to first make an unassisted decision and provide this as an input parameter to the algorithm before the algorithm generates its own automated decision.

User interaction flow in traditional decision support systems



Proposed user interaction flow - full

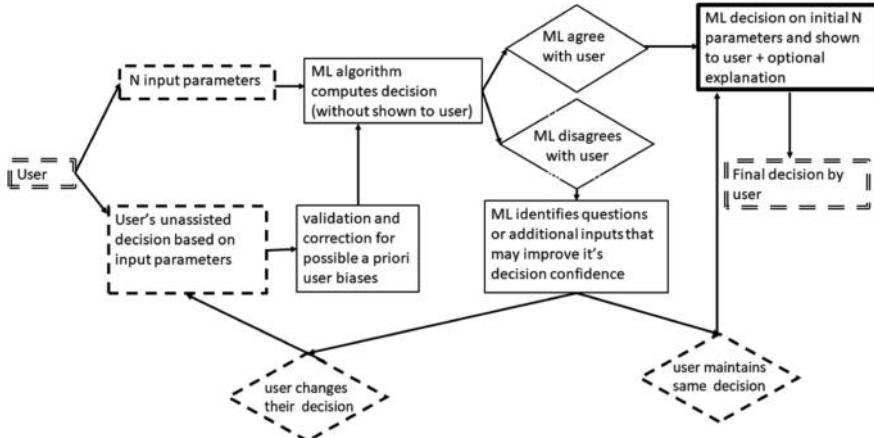


Fig. 1. Traditional versus proposed user interaction flow

2 Formal Description

Let $X_n = (x_1, x_2, \dots, x_n)$ be the n input parameters to the algorithm.

x_i can be continuous or categorical variables.

Let $v = (v_1, \dots, v_n)$ be the particular input values entered by the user.

v_i may have bias or error so we define $\Omega_i(v_i)$ as the set of values which are considered within the error bounds for v_i .

Let $D = \{\alpha^j\}, j = 1, \dots, k$ be the set of k possible decisions or output classes.

Let $\alpha_U \in D$ be the initial unassisted decision of the user.

Additionally we allow the user to mark a subset of input parameters $(v_1, \dots, v_m), m \leq n$ as being particularly important to their decision α_U .

For example, a decision support system that assists with diagnosis of respiratory infections may have $n = 5$ input parameters $x = (\text{patient temperature, cough, presence of chest pain, degree of sweating, loss of appetite})$.

A particular patient may have $v = (101^\circ\text{F, dry cough, sharp chest pain, heavy sweating, mild loss of appetite})$.

The doctor might also mark parameters 1, 2 as particularly important to their initial decision $\alpha_U = \text{bronchitis}$.

We define the decision function \mathbf{f} which maps an input vector \mathbf{v} to a class $\alpha \in D$ with confidence $c \in [0, 1] : f(\mathbf{x}) : X \rightarrow (\alpha, c)$

Let α_{ml} = the algorithm's decision based on the user-provided input values \mathbf{v} .

To determine the stability of a given decision α with respect to \mathbf{v} , we can perturb the input values to ones that are atypical for α .

Let $v_i^{atyp}(\alpha^j)$ = value of the x_i input parameter which has the lowest conditional probability given a particular decision $\alpha^j : \min(p(x_i|\alpha^j))$ out of all $v_i \in \Omega(v_i)$.

Let $\mathbf{v}^{atyp}(\alpha^j)$ = the vector $(v_1^{atyp}(\alpha^j), v_2^{atyp}(\alpha^j), \dots, v_m^{atyp}(\alpha^j), v_{m+1}, \dots, v_n)$

Let (α_{ml}^*, c^*) be the algorithm's decision after perturbing (v_1, \dots, v_m) by replacing each value v_i for $i \leq m$ with $v_i^{atyp}(\alpha_U)$.

Alternatively, we can perturb an input value v_i "in favor" of a given decision α^j .

Let $v_i^{typ}(\alpha^j)$ = value of the x_i input parameter which has the highest conditional probability given a particular decision $\alpha^j : \max(p(x_i|\alpha^j))$. In contrast to the $v_i^{atyp}(\alpha^j)$ definition, here we consider all possible values of x_i , and not just those within $\Omega(v_i)$. The reason for this will be explained below.

Let $(\alpha_{ml}^\dagger, c^\dagger) = \mathbf{f}(v_1, \dots, v_m, \dots, v_{m+1}, \dots, v_i^{typ}(\alpha_U), \dots, v_n)$. By this definition, α_{ml}^\dagger represents a decision based on one single input value being perturbed "in favor" of α_U .

Formalized user interaction flow:

1st Step:

User input1: $\mathbf{v} = (v_1, \dots, v_n) \in \mathbf{X}$.

2nd Step:

User input2: the initial unassisted decision α_U

3rd Step:

User input3: user marks m out of n input values as being particularly important to their decision α_U

4th Step:

In order to determine how stable α_U is relatively to perturbations of \mathbf{v} within error bounds Ω , we compute α_{ml}^* .

If α_{ml} doesn't match α_U , go to Step 5.

If α_{ml} matches α_U , go to Step 6.

5th Step:

Since $\alpha_U \neq \alpha_{ml}^*$ we iteratively work with the user to see if we can converge on a stable decision. We could, at this point, just show α_{ml} (or perhaps α_{ml}^*) to the user, but we specifically avoid doing this in order to prevent the user from unthinkingly changing their decision to α_{ml} . Instead we use a more nuanced, indirect approach where we try to find the parameter whose value v_i most deviates from the typical value for the given α_U .

To do this, we go through each $i \in (m, n]$ and, one at a time, replace v_i with $v_i^{typ}(\alpha_U)$ and use this to compute $c^* - c^\dagger$

The goal is to find the parameter i for which this perturbation leads to the maximum $c^* - c^\dagger$.

After finding this parameter, we report to the user that the value they provided for this parameter is to some degree inconsistent with α_U . We then give the user the option to change their initial α_U .

If the user maintains the same decision α_U , go to Step 6.

If user changes their decision, go to Step 2 (unless this point is reached a 3rd time, in which case go to Step 6 to avoid an overly long interaction loop).

There are several subtle details about this approach that are intended to make the user think about their initial decision α_U from a new angle without causing annoyance.

First, the interaction takes the user's point of view (as encoded in α_U and other input parameters) as the starting point and works from there.

Second, we try to only point out parameters $i \in (m, n]$ because the user doesn't consider these as particularly important to α_U , so it may be more productive to point to inconsistency in these parameters rather than parameters $[1, m]$ which the user has marked as supporting α_U .

6th Step:

We have reached the end of the interaction flow.

Compute decision α_{ml} based on unperturbed input values $f(v)$.

Display α_{ml} to the user.

3 Example Application

This proposed approach was implemented in the decision support system “Dinar-2” which assisted physicians in establishing the pathology and severity of cases when triaging emergency calls at the Center for Child Air-Ambulance Services in Yekaterinburg, Russia [7, 8]. One of the goals of this Center was to provide remote consultation to regional medical centers and doctors involved in treating seriously ill children, and thereby reduce the need to airlift children to larger or more specialized hospitals.

The Center was responsible for a large geographic area, which meant that air-ambulance services, when dispatched, could still take a long time to reach their destination. Given the volume and complexity of requests for consultation and air-ambulance services, a computerized decision support system was key to the efficient functioning of the Center. Dinar-2 was developed to fill this need. This system provides assistance in diagnosing the type of pathology (8 distinct classes of pathology), and in determining its severity (between 3 and 5 levels of severity—depending on the class). It also assists in selecting the best course of action, and in selecting the healthcare center that's best suited for treating a given patient.

The Dinar-2 decision support algorithm consists of 3 stages:

1. Identification of informative patterns and groups of symptoms.
2. Determination of the likely pathologies based on 1.
3. Determination of severity.

These steps were implemented using rule-based machine learning algorithms.

Besides objective measurements and test results, the system had to take into account a significant amount of subjective information about the patient's condition. This made the decision support task more complicated because the subjective information was susceptible to conscious and subconscious biases on the part of the reporting physicians. Specifically, these biases tended to skew the provided information toward making a patient's condition appear either more or less severe than it actually was.

Due to this, the Dinar-2 decision support system assigned an a priori confidence interval to every input parameter that was based on subjective information. Then, the system perturbed the inputs within the bounds of these confidence intervals, and checked whether its computed diagnosis was consistent with the diagnosis initially proposed by the user (in this case a physician at the Center, in consultation with the regional doctor). If, under these perturbations, Dinar-2's diagnosis of the pathology or severity did not match that of the user, Dinar-2 would follow the proposed interaction flow (described in Sect. 2 above) to clarify the diagnosis.

After its initial deployment in 1993, Dinar-2 was soon adopted by 39 air-ambulance centers across Russia, Kazakhstan, and Belarus, and has been in continuous use since then. For example, available statistics for the Yekaterinburg region for 2017 indicate that during that year, the system assisted in evaluating 537 cases. In 131 of these cases (24%), effective remote diagnosis and consultation proved sufficient for resolving the patient's crisis, and the need to dispatch an air-ambulance was avoided.

4 Discussion

There are a number of benefits and opportunities afforded by the proposed approach. Requiring the user to first reach their own decision serves to counteract the loss of users' expertise and sense of responsibility that often occurs when users delegate decisions to a DSS. It prevents the user from becoming complacent, and motivates them to give more thought to their initial decision. It provides continued opportunity for user to revisit and refresh their domain knowledge. When the user and the algorithm don't agree, it forces the user to reconsider their decision in light of parameters highlighted by the algorithm. In the end, it makes it more likely that the user will critically evaluate the machine's decision. In applications where the algorithm is more accurate than human users, this even allows the user to challenge themselves to anticipate the algorithm's answer—either on their own, or explicitly, by adding game-playing elements to the interaction.

Additionally, in terms of explainability, it allows the algorithm to focus on the pairwise difference between its decision and the user's decision, rather than having to take into account all other possible decisions. Finally, if some of the inputs provided by the user are based on subjective information, it's likely that these will contain biases that support the user's unassisted decision. By having access to this decision, the algorithm can take steps to correct for these biases.

As far as limitations, the proposed approach works best when the following assumptions are met.

First, it's intended for DSS applications with more than 2 output classes ($k > 2$).

Also, it assumes that many of the input parameters are imprecise or subjective, making it possible to perturb them within error bounds.

Finally, the proposed approach assumes that the decision support algorithm and the user are not too far apart in their ability to arrive at the correct decision based on the input parameters. If the algorithm by itself significantly outperforms the user in a particular application, then requesting that the user provide their own unassisted decision as input to the algorithm may contribute little.

In the case of applications where the final decision and responsibility for that decision cannot be delegated to an algorithm, a modified version of the proposed approach could still be useful if it leaves out Step 6. That way it never shows α_{ml} to the user and instead iteratively draws the user's attention to inconsistencies or important details in the input values—Fig. 2. A system called Summary Page serves as an example of this last approach. [9]

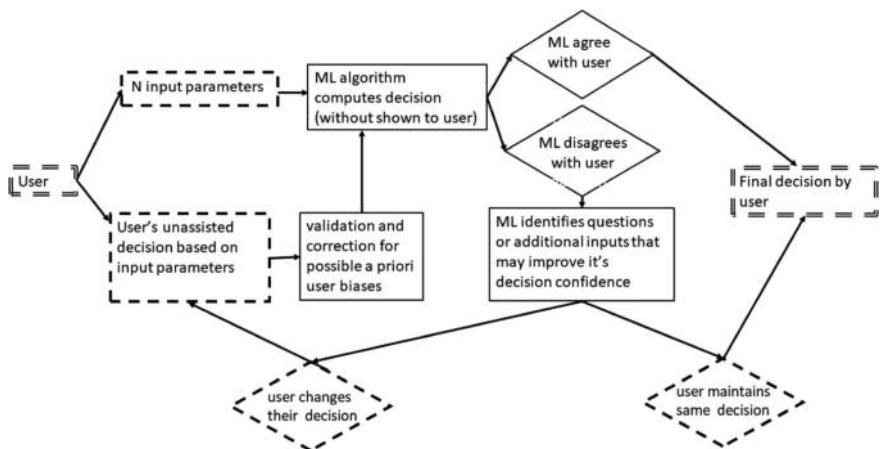


Fig. 2. Variation of proposed approach that never shows α_{ml} to the user

5 Conclusion

In conclusion, there are a number of benefits to structuring decision support systems in a way that makes the user provide their own unassisted decision to a decision support system as a first step. It allows the user to increase their accuracy by encouraging them to think more critically through an iterative decision making process. It also makes it possible to increase the accuracy of the DSS by giving it information on potential biases in any subjecting input parameters which the user may unknowingly skew in support of their initial conclusion.

Ongoing work includes developing additional applications of this approach and designing ways to quantitatively evaluate its effectiveness. Additionally, we are working on approaches to improve the explainability of the DSS algorithm's decisions.

References

1. Deo, R.C.: Machine learning in medicine. *Circulation* **132**(20), 1920–1930 (2015)
2. Rajpurkar, P., et al.: Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning (Dec. 2017)
3. Poplin, R., Newburger, D., Dijamco, J., Nguyen, N., Loy, D., Gross, S.S., McLean, C.Y., DePristo, M.A.: Creating a universal SNP and small indel variant caller with deep neural networks (2016)
4. Chan, C.-Y.: Advancements, prospects, and impacts of automated driving systems. *Int. J. Transp. Sci. Technol.* **6**(3), 208–216 (2017). <https://doi.org/10.1016/j.ijtst.2017.07.008>
5. <https://www.theverge.com/2018/4/29/17298750/tesla-autopilot-british-driver-charged-driving-sleeping>
6. Qi, Z., Li, F.: Learning Explainable Embeddings for Deep Networks. Available: <http://www.interpretable-ml.org/nips2017workshop/papers/04.pdf> (2017)
7. Goldberg, S.I., Makhanek, A.O., Novikov, I.D.: Dispatcher—consultative expert systems. *Bull. All Union Commun. Inf. Comput. Facil. BTHMU* **1**, 46–52 (1991)
8. Goldberg, S.I.: Inference Engine the Systems of the Dr. Watson Type, pp. 33–45. DIMACS Workshop, Rutgers University (1997)
9. Goldberg, S.I., Niemierko, A., Shubina, M., Turchin, A.: “Summary Page”: a novel tool that reduces omitted data in research databases. *BMC Med. Res. Methodol.* **10**, 91–97 (2010)



SEAKER: A Tool for Fast Digital Forensic Triage

Eric Gentry¹, Ryan McIntyre¹, Michael Soltys^{1(✉)}, and Frank Lyu²

¹ California State University Channel Islands, Camarillo, CA 93012, USA
michael.soltys@csuci.edu
<https://compsci.csuci.edu/>

² SoCal High Technology Task Force (SCHTTF), Camarillo, USA

Abstract. Faced with a preponderance of high capacity digital media devices, forensic investigators must be able to review them quickly, and establish which devices merit further attention. This early stage of an investigation is called *triage* and it is a chief part of *evidence assessment*; see [1, Chap. 2]. In this paper we present a digital forensic device, which we named SEAKER (Storage Evaluator and Knowledge Extraction Reader), which enables forensic investigators to perform triage on many digital devices very quickly. Instead of imaging the drives, which takes hours, SEAKER does a search for files with names that conform to pre-established patterns. The search is done by mounting the devices in read-only mode (to preserve evidence) and listing the contents of the device. Unlike imaging, this approach takes minutes rather than hours. Also, SEAKER’s hardware consists principally of a Raspberry Pi (RP) and so it is very inexpensive—this is crucial in this era of budgetary constraints; see [2]. Once SEAKER has identified media devices of interest, those can be confiscated for further investigation in a lab. But devices that do not have hits can be left at the scene. This has two principal benefits: forensic examiners can concentrate on those devices that are promising in terms of evidence for the given investigation, and devices without hits are not confiscated from legitimate users.

Keywords: Triage · Digital evidence assessment · Automation · Raspberry Pi · Storage forensics · Digital evidence and the law · Digital evidence preservation

1 Introduction

In this paper we are going to present a digital forensic tool (SEAKER) that was designed and implemented as a result of a collaboration between law enforcement, specifically Southern California High Technology Task Force—Ventura DA, and academia, specifically California State University Channel Islands Department of Computer Science. In the words of [2]:

Law Enforcement investigators and forensic analysts have extensive knowledge of what is required to progress an investigation and secure a conviction. Academia is able to provide scientific support, recognized testing procedures and computer science specialists.

The collaboration that we describe in this paper has been especially fruitful, as we designed a device that is able to significantly speed up *Digital Forensic Triage (DFT)*.

We named our device SEAKER, which stands for *Storage Evaluator and Knowledge Extraction Reader*, and its intended application is to quickly scan storage drives in read-only mode, forwarding hits to handhelds of investigators, according to predefined regular expression searches.

SEAKER performs what [3] calls “enhanced preview” for Law Enforcement digital forensic investigations. Also, quoting from [2]:

Due to budgetary constraints and the high level of training required, digital forensic analysts are in short supply in police forces the world over. This inevitably leads to a prolonged time taken between an investigator sending the digital evidence for analysis and receiving the analytical report back. In an attempt to expedite this procedure, various process models have been created to place the forensic analyst in the field conducting a triage of the digital evidence.

SEAKER is an example of such a process model, designed to help with the triage of digital evidence.

2 Digital Triage

Hitchcock et al. [2] covers the initial phase of an investigation, from issuing a search warrant to acquisition of evidence. As there are more Detectives and Investigators than Computer Forensic Examiners (CFEs), CFEs will always have a long queue of cases. A solution might be to create a tiered system for forensic examiners, consisting of examiners who attend search warrants and use approved tools to conduct triage, and in-lab examiners who will analyze the results of the triage. In this model SEAKER’s role will be that of a well understood and trusted tool to be used by the triage personnel. SEAKER would fill this role as it is:

1. *Simple to use*: in the fluid and chaotic environment of a search warrant, the examiners operate a simple triage system.
2. *Tested*: can be operated with confidence at the scene with assurance that evidence thus obtained will stand in court.
3. *Fast*: many devices can be examined, and actionable intelligence can be provided to the investigators at the scene.

SEAKER also acts as a distributed system, as many media devices can be connected to the Raspberry Pi (RP) simultaneously, and all those devices can be examined by many investigators working concurrently. For each connected device, SEAKER creates a unique file containing the complete listing of files on that device. This file can be then searched by several investigators simultaneously, each for a different (if needed) set of patterns.

3 SEAKER

3.1 SEAKER Functionality

It was a conscious design decision to make SEAKER very easy to use. After collecting all the media devices at the scene, the investigator will triage them with SEAKER. Here are the steps:

1. Connect the RP to the power, and after waiting for about a minute to let it start, connect to the RP's hotspot (WiFi network). Depending on the setup, the WiFi's SSID will be "SEAKER01" or "SEAKER02" etc. (if there were to be more than one SEAKER at the scene). See Figure 1. Note that the hotspot is password protected; again, the password may be configured.
2. At the same time the investigator may connect all the media devices to the RP. This may be done concurrently with the previous step.
3. Once connected to the hotspot, the investigator will open any web browser on their handheld, and direct it to go to <http://seaker01.local>. We decided to allow access through a web browser as this is the most universal way to connect; any device (iPhone, iPad, Android, laptop, etc.) can connect to a hotspot and open a browser. Once the browser establishes the connection, the user will see Fig. 2. Note that the keywords (or regular expression patterns) present in the "Type in Search Terms:" can be pre-loaded before arriving at the scene, or changed/updated at the scene.

The regular expression can be given using the syntax of the `grep` utility. For example, if we want to find occurrences of either 'two' or 'too', we use `t[w|o]o`; if we want to find every word that start with capital letters, we use `[A-Z]`; if we want to find words where number 9 is the last character of the line, we use `9`. There are a vast number of possibilities; we can also replace `grep` with `egrep` that has an even richer syntax. Check the `grep` and `egrep` man pages for all the details.



Fig. 1. Investigator's handheld view: connecting to the RP's WiFi (hotspot)



Fig. 2. Investigator's handheld view: using a browser, connect to <http://seaker01.local>

- Once any storage devices that are found at a search warrant scene are connected to the RP, the investigator will typically wait for a few minutes (we have seen times up to 10 min for 1 Tb disks with millions of files) for the file list to be built. The search will then be carried out very quickly: essentially, grep browses the file list, line by line, outputting those lines that conform to at least one pattern specified in the “Type in Search Terms.” window. Once this finishes, the investigator will have the results presented as in Fig. 3.

The filenames themselves can be incriminating evidence, such as in Child Pornography (CP) cases, where the material has a commonly used naming convention, e.g., “lolita” which can be found with the grep pattern `.*lolita.*` (`.*` means the following: ‘.’ (period) matches any single character of any value, except a newline, and ‘*’ (asterisk) matches zero or more of the preceding character or expression) or simply `lolita`. This can be used by the investigators to question the suspects. The questioning usually takes place at the same time as the forensic examiners triage the evidence, and one of the requirements of SEAKER was to be fast so that investigators can start getting intelligence quickly from the initial processing of the scene. This can be seen in the User Process Flow shown in Figure 4.

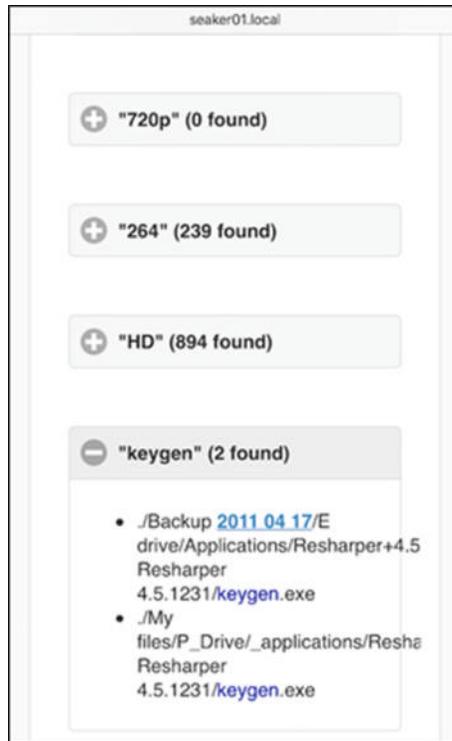


Fig. 3. Investigator's handheld view: the results of the search of a particular device

3.2 SEAKER Design

SEAKER has been custom designed to meet a particular need of forensic examiners; quoting [2]:

One of the benefits of a custom built tool is that changes are able to be made from a grass roots level.

SEAKER consists of standard, inexpensive and easy to obtain components. The heart of the device is a Raspberry Pi (RP) together with ancillary components such as SATA, USB and power cables. All these components can be purchased for under \$150. The RP has to be set up for usage; this is accomplished with a single long bash script that installs all the necessary software components.

The design principles, indeed the requirements of the solution were the following:

1. *Fast*: digital forensic examiners already have at their disposal methods for reviewing hard disks (hds), thumb-drives and other devices in a way that prevents tampering with evidence. The problem is the proliferation of hds

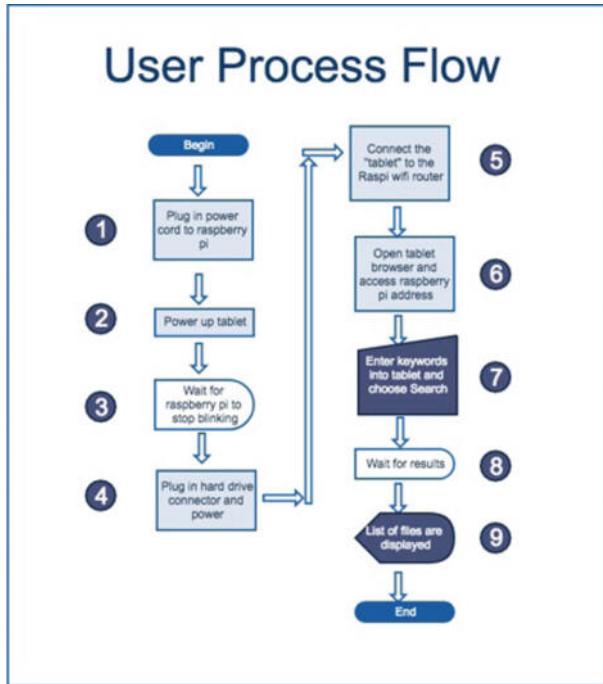


Fig. 4. The functionality of SEAKER from the user perspective

with terabytes of data. Imaging such hds takes hours, and when there are many hds this is infeasible. SEAKER provides a way to triage devices quickly at the scene, and aids in selecting devices that warrant in-depth examination.

2. *Robust*: the device will be used in the field, in chaotic settings, where the examiners cannot be spending time tinkering with the device, and following complicated instructions; it has to be simple and reliable.
3. *Inexpensive*: given the budget cuts facing the public sector all over the world, the aim was to create a working prototype with off-the-shelf components that can be purchased under \$200 per unit.
4. *Compact*: the system must be brought to the scene, so portability is key. The small nature of the RP and accompanying plugs and cables makes SEAKER ideal for transport.
5. *Easy to use*: once configured in the lab, the system is ready to work at the scene. In fact, it can be deployed by investigators with little knowledge of IT.

The process flow of SEAKER (Fig. 5) is very simple. After powering the RP, two processes happen simultaneously. Immediately when the drive(s) are connected to the RP, the file names and their paths are copied to a text file. Meanwhile, the user must connect to the RP via WiFi (usually on a handheld device, but any device with a browser works). So in fact, the RP becomes a hot spot. The user must then open the SEAKER web page. As shown in Fig. 5,

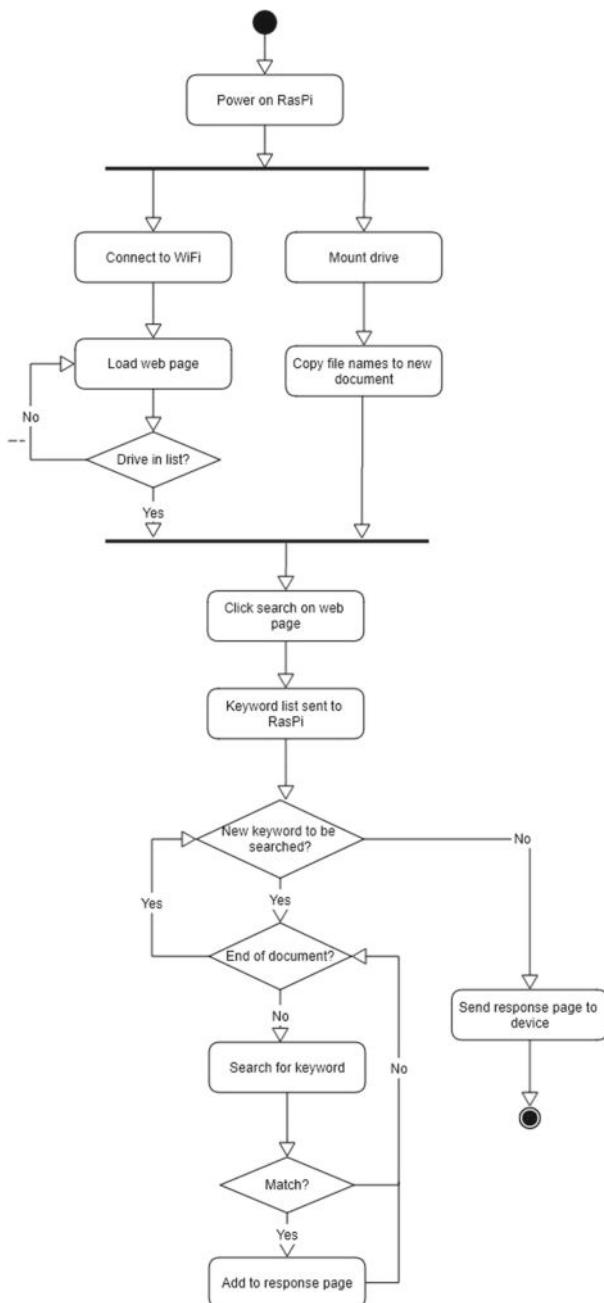


Fig. 5. SEAKER flowchart. Note the left-most “No” arrow indicated with two dashes: the page refreshes every 3 s, and the first drive connected is selected automatically; other drives must be selected manually

the web page will refresh every 5 s, checking each time if new drives have been connected (and their file system recognized) by the RP.

All selected drives will then go through the search process as shown in the lower section of Fig. 5. Each drive will be processed sequentially. For example, the list of files for the drive will be scanned for any matches to the first keyword in the list. The search is done using the UNIX `grep` utility. All files that are found to match that keyword will be added to a text file. Once the whole list of files has been searched for that keyword, the process will begin with the next keyword in the list. This process will continue until there are no more keywords to be searched. If multiple drives have been selected to be searched, this process will continue for each drive. Finally, the text file is used to create the response page and is sent back to the user's mobile device.

While the listing of all the files on the drive can be accomplished with standard UNIX utilities such as `find` or `ls -R`, we opted for a simple C program that does so faster as it does not implement all the additional functionality of those standard utilities. Our C code is contained in `collect_files.c` and it implements a recursive depth-first search of a given directory tree structure (outputting only file names):

```

1 #include <unistd.h>
2 #include <sys/types.h>
3 #include <dirent.h>
4 #include <stdio.h>
5
6 void listdir(const char *name) {
7     DIR *dir;
8     struct dirent *entry;
9     if (!(dir = opendir(name))) {
10         return;
11     }
12     while ((entry = readdir(dir)) != NULL) {
13         if (entry->d_type == DT_DIR) {
14             char path[1024];
15             if (strcmp(entry->d_name, ".") == 0 ||
16                 strcmp(entry->d_name, "..") == 0) {
17                 continue;
18             }
19             snprintf(path, sizeof(path), "%s/%s",
20                     name,
21                     entry->d_name);
22             listdir(path);
23         } else {
24             printf("%s/%s\n", name, entry->d_name);
25         }
26     }
27     closedir(dir);
28 }
29
30 int main(void) {
31     listdir(".");
32     return 0;
33 }
```

4 SEAKER Technical Specifications

4.1 Hardware Requirements

The setup of a SEAKER requires the following hardware:

1. *Raspberry Pi (RP) 3 Model B*: Inexpensive, compact computer hardware that runs a version of Linux. It comes with built-in WIFI, 4 USB ports, and other typical computer connections. This also doubles as the WIFI router for connecting the handheld phone or tablet. External power must be supplied to the RP.
2. *MicroSD card*: This is the hard drive for the RP that will hold the operating system, web server, WIFI router software, and the searched contents of the suspect hard drives.
3. *USB hard drive adapter*: Plugs into the RP's USB port and enables hard drives to be read. External power must be supplied to the USB hard drive adapter.
4. *Handheld phone or tablet with WIFI*: This could be anything from an Apple iPad to a smart phone to a laptop computer. Several devices are able to connect at the same time to the RP via the WIFI hot spot capability.
5. *Ethernet internet connection*: Required for initial setup only. The connection is necessary to download the initial prep script, obtain software updates during the setup phase, and allow for the wireless NIC to be setup for WIFI.

4.2 RP Preparation Script

SEAKER is designed for portability and ease of use, given the chaos and complexity of a crime scene digital forensic triage analysis. The initial setup and configuration is also designed with similar simplicity in mind, but may require a person with IT knowledge.

The prep script is a simple `bash` script written for the September 2017 release of Raspbian Jessie Lite operating system, but is very likely to continue to work with newer versions. The first step for setting up the initial state of the RP is to load the standard “Lite” version of this image onto the microSD card.

The following steps are to customize and finish setting up the RP. Editing the top few lines of the script will enable a customized setup for the particular instance of SEAKER. Here are the editable lines:

```

1 PL_PASSWORD="raspberry"
2 WIFI_NAME="SEAKER01"
3 WIFI_PASSWORD="raspberry"
4 WIFI_ROUTER_IP="192.168.101.1"
5 WIFI_ROUTER_DHCP_RANGE="192.168.101.50 192.168.101.100"
```

After editing the script, it must be copied to the RP and run. Using the *pi* account and home path is best for this. The script will automatically reboot when it is finished and the RP is then setup for use as a SEAKER. The script is also designed to clear itself out and clear out the history to ensure that the passwords are not able to be recovered from the script file or the history.

Once the setup is complete and the RP is powered on, a new WIFI access point will be available using the broadcast name that was set in the *WIFI_NAME* setting from the script. Using the handheld phone or tablet to connect to this WIFI access point enables a web-page to be called up using the same web address as the name of the access point. This web-page is the main mechanism for searching for specific information from the connected suspect hard drives.

4.3 Mount Rules

When a drive is connected, SEAKER automatically detects and mounts it, then calls the collect and search procedures. This is done with a *.rules* file in the

/lib/udev/rules.d/

directory; these files can be used to automatically perform actions when certain events occur (in this case, mounting and listing the contents of storage devices upon connection). In SEAKER's mounting rules, below, each line begins with conditions, characterized by “==” and “!=.” If all conditions are met, then the actions described in the rest of the line is taken. Indented lines are actually continuations.

```

1 KERNEL!="sd [a-z][0-9]" ,
2   GOTO="auto_mount_usb_storage_by_label_end"
3
4 # Import FS infos
5 IMPORT{program}="/sbin/blkid -o udev -p %N"
6
7 # Specify a label
8 ENV{dir_name}="usbhd-%k"
9
10 # Global mount options
11 ACTION=="add", ENV{mount_options}="ro"
12
13 # Filesystem-specific mount options
14 ACTION=="add", ENV{ID_FS_TYPE}=="ntfs|exfat",
15   ENV{mount_options}="$env{mount_options},"
16   user_id=1000,group_id=1000,ntfs-3g"
17 ACTION=="add", ENV{ID_FS_TYPE}!="ntfs|exfat",
18   ENV{mount_options}="$env{mount_options},"
19   uid=1000,gid=1000"
20 ACTION=="add", ENV{ID_FS_TYPE}=="ext3|ext4|ntfs",
21   ENV{mount_options}="$env{mount_options},noload"
22
23 # Mount
24 ACTION=="add", ENV{ID_FS_TYPE}!="ntfs|exfat",
25   RUN+="/bin/mkdir -p /mnt/%E{dir_name}",
26   RUN+="/bin/mount -o $env{mount_options}"
27   /dev/%k /mnt/%E{dir_name},
28   RUN+="/home/pi/seeker_collect.sh %E{dir_name}
29   | at now"
30 ACTION=="add", ENV{ID_FS_TYPE}=="ntfs|exfat",
31   RUN+="/bin/mkdir -p /mnt/%E{dir_name}",
32   RUN+="/home/pi/mount_ntfs.sh %k %E{dir_name}"
33
34 # Clean up after removal
35 ACTION=="remove", ENV{dir_name}!="",
36   RUN+="/bin/umount -l /mnt/%E{dir_name}",
37   RUN+="/bin/rmdir /mnt/%E{dir_name}"
38
39 # Exit
40 LABEL="auto_mount_usb_storage_by_label_end"

```

Note that NTFS and EXFAT are mounted using different rules. Some file systems have issues (well known among Linux users) with being mounted by custom rules, and require said rules to mount them indirectly through an external script. Details like this provide the primary barrier to filesystem support.

It is important that drive contents are not changed when SEAKER searches them; the drive contents may be evidence in an investigation, after all. SEAKER utilizes two mounting options to avoid tampering: `read only` and `noload`. `read only` is straightforward: SEAKER can only read the contents of the drive—that is, it cannot write or execute. `noload` is applicable to journaling filesystems. A journal serves as a change-log for the drive. If it is accessed, it will change, which will result in a change in the drive's hash. Though such a change would

not affect the stored contents, the change in hash value could cast doubt on the integrity of the evidence. `noload` prevents journaling from occurring, allowing for access without any change to the hash and thereby making it possible to quickly demonstrate that SEAKER leaves the drive contents unchanged.

5 Future Work

SEAKER is currently a prototype; there are many improvements to be made, and we will discuss some of them in this section. These improvements may be implemented by future students, or by digital forensics professionals. We encourage anyone who implements them to share their work.¹

First, SEAKER supports a select few filesystems. Namely, NTFS and FAT (`exfat`, FAT32, FAT16, ...). There are likely bugs to be worked out in the supported filesystems, and there is certainly work to be done in expanding the list of supported systems.

If an unsupported filesystem is in use, it may simply fail to mount, and not show up on the SEAKER site at all. Similarly, drives from which files are currently being collected do not appear; the site displays them only when files have been collected. Here there is opportunity for improvement; as opposed to displaying drives for which collection is complete, SEAKER could display all drives, and a status next to each. This status only requires three states: failed search (for unsupported systems), collection in progress, and collection complete.

It can be difficult to match a hard drive to its corresponding search results. Partitions are uniquely identified by a UUID, and some properties (capacity, for example) are displayed with the search results, but these properties do not provide a perfect way to determine which physical device corresponds to which mounted partition or device. Storage devices generally have a serial number of sorts, but this serial number is not visible to SEAKER. This is a problem which requires some creativity to solve well. SEAKER could take a picture when a drive is plugged in, and associate that picture with the search results, for instance, but this solution requires that investigators position each storage device in front of a camera; this approach requires a camera, and moreover it is tedious and error-prone.

When a search finds a hit (i.e., a matched expression or file extension), investigators may want to view the corresponding file. Currently, this would require them to manually find and open the file. Speed and ease of use are priorities, so it would be best if investigators could select a file in the search results and have SEAKER fetch a copy of it for them. This function inevitably requires that the storage device being queried is still connected—assuming that this condition is met, copying and viewing a file should not be too complex.

Similarly, it would be useful if investigators could view thumbnails of images and videos in the search results. One example of the motivation here is child

¹ The main `bash` script for SEAKER is available on GitHub at <https://github.com/michaelsoltys/seaker>.

pornography cases; incriminating images may have innocuous names, but thumbnails would indicate the true content.

This leads to another issue: as incriminating files may be named innocuously, investigators will often want to search simply for all images, videos, etc. SEAKER could minimize the work necessary by allowing for preset groups of search terms, which can be created and edited by administrators. For example, an admin could create an “images” group which causes SEAKER to include jpg, pdf, png...

We are very interested in a “Data Carving” option. *Data carving* is the identification and extraction of files from unallocated clusters using file signatures. A file signature, also commonly referred to as a magic number, is a constant numerical or text value used to identify a file format. The object of carving is to identify and extract (carve) the file based on this signature information alone. We are interested in hidden files (which are sometimes easy to locate, as for example in UNIX with ‘ls -a’ command) and deleted files, which is more tricky as the files can partially overwritten. A partially overwritten file may still constitute valuable evidence: for example, a portion of an image can be taken as solid evidence that the entire image was on the disk at some point. How can one establish whether a portion of an image comes from a particular image? It seems that the only way to accomplish that is by visual inspection, and having an investigator recognize the original image. In order to automate this process one could attempt one of two things: build a massive database of frequently circulating (say, CP) images, and hashing different formats of these images (.pdf, .jpg, .giff, .tiff, etc.), as well as different resolutions, and chunks of standard sizes (say, 64Kb). This still seems like a shot in the dark. The second approach is to define something akin to *fuzzy hashes*, the type of hashes that are used to recognize variants of the same malware. This new type of fuzzy hashing would be invariant under different formats, or standard resolutions, and chunks of an image could be identified by close proximity to the original hash. Hits would be still confirmed visually to avoid false positives; a bigger issue would be false negatives.

Finally, documentation is important in any investigation. When triage reveals media which motivates investigators to confiscate the corresponding storage device, they should document this motivation. As such, it would aid investigators if SEAKER could generate a search report for a selected drive from the search results screen. This report could be downloaded to the investigator’s device or saved on the SEAKER unit for later access by an administrator. It should contain the search results along with some circumstantial information, such as the date, the name(s) of investigator(s) requesting the report, and their reason for confiscating the device.

6 Development Tools

An important component of this project was the learning experience for the students. While we all worked toward a working prototype for the digital forensics lab, for the majority of the students this was the first time working on a

project for which the outcome would be more than the satisfaction of a course requirement. To work in a large team (18 students) different talents, abilities, personalities and work habits had to come together in order to produce a working device. The students used a large set of tools that enabled the development of SEAKER. Here is a list of the principal tools, with short descriptions of their features. It should be mentioned that most these tools are Open Source and free to use.²

1. *BASH*: shell scripting was at the heart of the project. While some creative work went into the hardware arrangements (discussed in Sect. 4.1), most of the work consisted in developing a long BASH script. The goal of the script was to set up a RP as SEAKER (the script, called `prep.sh` is discussed in Sect. 4.2). Learning to code in BASH was a big part of the project for many of the students.
2. *Slack*: this is a fantastic collaboration tool that was introduced to the team by the third author of this paper. It allows for an easy and convenient exchange of ideas and brainstorming while developing a product. One of the best features is the ability to divide the conversation into different channels; there were channels for discussion of hardware, software development, testing and documentation.
3. *Dropbox Paper*: this became a de facto Wiki for maintaining the project documentation. It is simple to use; it requires little beyond familiarity with a standard Markdown language. It facilitates a distributed documentation development effort. Some participants had “edit rights” while all participants could read and post comments. Eventually, two documents emerged: a set up installation guide, and usage documentation.
4. *Github*: a fundamental tool, known to all software developers. About one third of the students requested GitHub collaborative access to help in the development of `prep.sh`. While there are many tools for collaborative software development, it is hard to find something better than GitHub.³ Anyone can preview the history of the development, and read the annotations.
5. *AWS*: we used an Amazon Web Services (AWS) S3 bucket to have a place with beta versions of the software. The `prep.sh` script, in its most recent version, as well as the most recent version of the documentation, are maintained in the bucket. We learned to mount the bucket onto an AWS EC2 instance, so we can update the staging folder with the latest version with `git`.
6. *C*: while relatively little has been done with the C programming language, a core functionality of the project has been developed in C. (See

² Technically, free for the students. Some services required nominal payments; for example, the fourth author has a GitHub subscription which allows for development with private repositories—anyone can open a GitHub account, but a free plan only allows public repositories. Similarly, the fourth author has an AWS subscription; we used Amazon Web Services (AWS) S3 buckets to have a staging repository for ready to use software, our beta versions.

³ We used GitHub to collaborate on this paper—which allows us to work together while hardly meeting in person.

- `collect_files.c` described on page 1234.) A good reminder that when one wants system performance one has to work with C (the `find` and `ls -R` functions were too slow, due to all their features, to list the entire disk).
7. `grep`: this tool is applied in the final stage of the capture, when all the files downloaded by `collect_files.c` are examined for the pre-assigned patterns.
 8. *Raspbian Linux*: a free OS, based on Debian Linux, and optimized for running on Raspberry Pi (RP) hardware. We used the “Lite” version, as we just needed the basic functionality, without the 35K+ packages that are present in the full version. It is indeed a revolution in controllers technology to be able to have a hardware controller furnished with a complete Linux OS. For more on the hardware, see Sect. 4.1.

It is important to remember that the principal contribution of this project rests in the performance of the device. Obviously digital forensics had methods to examine data in a sound manner; the augmentation offered by this device is the speed at which triage can be performed in the field. The advantage of our system is so obvious (minutes to list and search all the file names, rather than the hours it takes to image a disk) that we did not need to justify the benefits of our solution. Still, as the device will be used in the field by practitioners, we plan to collect data to keep track of the performance as bigger disks and new filesystems are being added.

7 Conclusion

The SEAKER project was a successful collaboration between two different institutions in the public sector: law enforcement and academia. The former has many interesting problems to offer, but as they are overwhelmed with cases they typically do not have the man power to do research and development. The latter is happy to do R&D, as it enhances the educational experience of the students to be learning in the context of applications to real life problems. It is a fortuitous and symbiotic relationship, and we plan to embark on other such projects in the future.

SEAKER is also the testament to the fact that supremely useful devices, meeting the needs of practitioners, can be constructed from relatively simple components; what is required is expertise and enthusiasm, which in the best cases academia possesses in ample measure. RPs are a revolution in embedded controllers, and we are just scratching the surface of their applicability. They are inexpensive, but wield the power of the Linux OS.

For the students, the experience was invaluable. Perhaps the most important aspect was non-technical: how to work well in a large team. There were eighteen students in the group; a composition of different backgrounds, talents and strengths. We divided the task into five different but interconnected teams: Task #1 was connecting the external devices to the RP; Task #2 was searching the contents of the devices; Task #3 was responsible for sending the query and retrieving the results of the search to the handheld; Task #4 was responsible for

the documentation of the project (both a user set up and guide, as well as the technical documentation of the solution); Task #5 was responsible for testing.

Digital forensics and academia would both benefit greatly from increased collaboration; students can offer relatively inexpensive development in exchange for real-world experience and the opportunity to create something which will be used. As a side effect more students would consider digital forensics as a career, resulting in some level of alleviation of the problems mentioned in the second quote in the introduction [2]. Everyone wins.

Acknowledgements. This work arose from a fruitful collaboration between SoCal HTTF (Southern California High Technology Task Force, Ventura County) and CSUCI (California State University at Channel Islands). We are very grateful for the opportunity to work on such an interesting and eminently applicable problem. We are especially grateful to Senior Investigator Adam Wittkins who facilitated this collaboration. The SEAKER development work was undertaken as a final project for a graduate course in Cybersecurity at CSUCI (COMP524: “Cybersecurity”). The first and third authors were students in this course, and they emerged as leaders of the project, but we are very grateful for the contribution of the rest of the class (in alphabetical order): Geetanjali Agarwal, Nick Avina, Jesus Bamford, Jack Bension, Apurva Gopal Bharaswadkar, Amanda Campbell, Christopher Devlin, Nicholas Dolan-Stern, Manjunath Narendra Hampole, Mei Chun Lo, Christopher Long, Clifton Porter, Deepa Suryawanshi, Mason U’Ren and Zhe Zhang (see <http://soltys.cs.csuci.edu/blog/?p=2713>).

A Instructions for Setting Up SEAKER

This section contains step by step instructions to build a SEAKER:

1. Prepare the MicroSD card
 - (a) Download latest version of Raspbian Lite Image to a local computer (<https://goo.gl/eNvdMu>)
 - (b) Download Etcher software for writing the image to the MicroSD card (<https://goo.gl/f6LHBU>)
 - (c) Download PuTTY if using a Windows based local computer (<https://goo.gl/Tvifot>)
 - (d) Write the image to the MicroSD card (at least 8GB) using Etcher (<https://goo.gl/FTvTVx>)
 - (e) Before removing the MicroSD card from the computer, add a file named ‘ssh’ (no quotes, no extension, no contents) to the root of the MicroSD card (<https://goo.gl/tTs2vd>).
2. Plug in and boot the Raspberry Pi (RP)
 - (a) Connect the RP to your network using the Ethernet port (Do not connect using WiFi)
 - (b) Plug in power to the RP and wait 10–20 s for the Raspbian Lite operating system to boot.

3. Find the RP's IP address and connect to it
 - (a) Find and make a note of the IP Address and substitute it in the rest of setup when RASPBERRYPI_IP is used; this can be done by tools like "Advanced IP Scanner" or by accessing your router administration page
 - (b) Use ssh (or PuTTY for Windows) to start a secure shell for example:
`ssh -l pi RASPBERRYPI_IP`
 - (c) When logging in, the default login is
 username: 'pi', password: 'raspberry'.
4. Get the prep script and run it
 - (a) At the RP prompt, download the prep.sh script:
`wget -O prep.sh https://goo.gl/5RU1Yv`
 - (b) Modify the first few lines to prevent collisions with other SEAKERS:
 - `PI_PASSWORD` (line 18) - Sets the RP's password
 - `WIFI_NAME` (line 19) - Sets the WiFi access point name
 - `WIFI_PASSWORD` (line 20) - Sets the WiFi WPA2 password
 - `WIFI_ROUTER_IP` (line 21) - Sets the WiFi access point IP address (must always end in .1)
 - `WIFI_ROUTER_DHCP_RANGE` (line 22) - Sets the DHCP address range (must have the same prefix)
 - (c) Set the permissions of `prep.sh` to 744:
`chmod 744 /prep.sh`
 - (d) Run the prep script:
`./prep.sh`
 - (e) The script will automatically reboot when finished.
5. Verify that SEAKER is working
 - (a) After the reboot, use a separate WiFi enabled handheld phone or tablet (look for a new WiFi access point named using the `WIFI_NAME` setting in the `prep.sh` script)
 - (b) Type in the WiFi password (from the `WIFI_PASSWORD` setting)
 - (c) Use a web browser from the handheld phone or tablet and type in the `WIFI_NAME` or new SEAKER IP address after "http://"; for example:
`http://SEAKER03.local.`

References

1. Hart, S.V.: Forensic Examination of Digital Evidence: A Guide for Law Enforcement. U.S. Department of Justice (2004)
2. Hitchcock, B., Le-Khac, N., Scanlon, M.: Tiered forensic methodology model for Digital Field Triage by non-digital evidence specialists. *Digit. Investig.* **16**, S75–S85 (2016)
3. James, J.I.: A survey of digital forensic investigator decision process and measurement of decision based on enhanced preview. *Digit. Investig.* **10**, 148–157 (2013)



Cyber-Physical Network Mapping Attack Topology

Glenn Fiedelholtz^(✉)

UMBC, IT/Engineering, 21250 Baltimore, MD, USA
gfiedelholtz@gmail.com

Abstract. This Cyber-Physical Network Mapping Attack Topology paradigm provides cyber analysts with appropriate background information to underpin efforts to provide accurate and comprehensive assessments in the development of cyber analytic products. In addition, this framework will assist in providing information regarding cyber threats, vulnerability and consequence analysis for the network assets that are being attacked by an adversary. The Cyber-Physical Mapping Network Topology will dramatically enhance the vulnerability and consequence analysis of cyber threats by improving the monitoring, detection, analysis, and mitigation capabilities in responding to cyber incidents in the United States. Network systems that control the critical infrastructure in most case operate constantly and the impact of downtime from a cyber exploit of the control systems that potentially could endanger public health and safety can range from inconvenient to catastrophic.

Keywords: Monitoring · Detection · Threat analysis · Vulnerability · Consequences

1 Introduction

The purpose of the *Cyber-Physical Network Mapping Attack Topology* framework is to standardize requirements for cyber-physical operational analysts and response practitioners in the event of cyber incident.

2 Pre-Incident Planning and Analysis

2.1 Steady-State Monitoring

The organizations under this cyber-physical incident framework process and monitoring will, in advance, have the mechanisms and facility to allow the development of a common operational picture of the incident in question based on participation and data input from the cyber analysts. Preparation also includes training to ensure cyber teams, and organization leadership are trained in cyber-incident response procedures and internal/external reporting mechanisms.

In the context of “steady state” monitoring is defined as the posture for routine, normal, day-to-day operations, as contrasted with temporary periods of heightened alert or real-time response to threats or incidents.

3 Incident Detection and Characterization

The United States Federal departments and agencies; State, local, tribal, and territorial governments; the private sector; and international partners will respond to cyber incidents employing this cyber framework. This process facilitates requests, receives, shares, and analyzes information on cyber attack Tactics, Techniques and Plans (TTP’s) and vulnerabilities among stakeholder’s critical infrastructure network assets.

Moreover, this cyber technical approach will enhance situational awareness and crisis monitoring of critical infrastructure, and information sharing on threat information and collaboration, assessment and analysis, and decision support pre- and post-incident.

3.1 Detection

The Cyber-Physical Network Mapping Attack Topology will enhance prevention and protection efforts to detect malicious or unauthorized activity on networks. In particular, this process will provide cyber operators with a systematic and real time approach to identify and contain malicious and unauthorized activity of intrusions on the critical networks.

This cyber framework provides the stakeholders with comprehensive network information pertaining to unauthorized activity, including any critical details on the who, what, where, when, why and how of the incident.

- **Conduct open source monitoring**—Monitor the Open Source Reports, a summary of open-source published information collected each business day concerning significant cyber intrusions. Each Cyber Report is divided by critical infrastructure sectors and key assets defined in the Federal cyber documentation and discusses relevant physical and cyber incidents across the United States.
- **Collect datasets (security incidents alerts, and events)**—During the cyber incident, review and collection of major cyber reports will be generated. For example, USCERT Open Source Infrastructure Reports.
- **Identify and analyze suspicious activity**—Identify and analyze suspicious activity reporting (SAR) from the Nationwide Suspicious Activity Reporting Initiative (NSI) database. The SAR initiative is a collaborative effort led by the U.S. Department of Justice (DOJ), Bureau of Justice Assistance, in partnership with the U.S. Department of Homeland Security (DHS), the Federal Bureau of Investigation (FBI), and State, local, tribal, and territorial law enforcement partners. The program establishes a national capacity for gathering, documenting, processing, analyzing, and sharing SAR information among law enforcement agencies. SAR reports are vetted by the fusion centers and shared as appropriate among NSI participants.

- **Detect and verify unusual and network traffic**—The cyber operational analysts will review, detect, and verify real-time cyber data with automatic collection tools, and conduct deep packet inspection of traffic coming to or from Federal Internet protocol (IP) addresses ending in “.com” or “.gov” to detect signs of suspicious or malicious activities.
- **Review and collect critical sector reports of essential information systems**—Review and collect relevant critical infrastructure sector reports and information through existing partnership agreements with critical infrastructure owners/operators, Federal agencies, and State/local governments:
 - **Critical infrastructure owners/operators**—Through their collaborative agreement with the critical infrastructure sector owners and operators, will coordinate and communicate directly with the appropriate leadership of critical infrastructure owners/operators. This may include crucial communication with the multiple stakeholders during a cyber incident.
 - **Federal agencies**—Significant cyber incidents may require nationally-coordinated rapid response actions based on differing authorities and priorities. Numerous organizations may provide essential data and capabilities: DHS, CS & C, Department of Defense, National Security Agency (NSA), DOJ, FBI, Department of State, and other Federal departments and agencies.
 - **State and local government**—Personnel from State, local, tribal, or territorial governments also play a major role in providing relevant information about the sector or entities, through media such as the Multi-State Information Sharing and Analysis Center (MS-ISAC).

3.2 Threat Analysis

Cyber threat analysis is the practice of effectively fusing incident information, knowledge of an organization’s network and vulnerabilities—both internal and external, including essential IT and industrial control systems (ICS)—and matching these against other actual cyber attacks and threats that have been observed or reported.

As part of this Cyber Physical Network Mapping Topology process, the analysts are responsible for threat analysis, which will characterize the attack, including scope and scale, from forensic information, and will attempt to ascertain the level of sophistication of the attack and the potential impact to the sector(s). This will include identification of other potentially vulnerable sector(s) and possible detection mechanisms for the attack. The attack’s level of severity will be based on the seven layers of the Open Systems Interconnection (OSI) Model.

- **Identify Tactics, Techniques, and Procedures (TTPs)**—The cyber analyst identifies tactics, techniques, and procedures that pertains to cyber attacks. Attacks such as distributed denial of service (DDoS), cyber espionage.
- **Define scope/scale**—The analyst should identify the scope and scale of the cyber-physical incident in terms of the organizations’ business enterprise and ICS.
- **Determine the intent and capabilities**—The analyst may seek information from on the intent, scale, and capabilities of the cyber attack (if available) to determine the type of attack and to understand the impact the disruption will have on the

network and the potential defense and detection mechanisms that will be required for mitigation.

- **Identify the sector affected (16 sectors) and any additional sector(s) with the potential to be affected**—The analyst should identify the sector(s) affected by the cyber-physical exploit to understand the upstream and downstream disruption impact on the sector supply chain and other sector(s) dependencies. This should include identification of other sectors that may be similarly vulnerable, so that the appropriate notification can be developed and communicated.
 - **Identify systems affected: cyber or physical**—To understand the impact of a cyber-physical exploit on critical infrastructure, it is imperative that information and network systems of the affected infrastructure be identified and understood by the analyst in terms of cyber-physical operational vulnerabilities and consequences.
 - **Determine severity**—Review seven layers of the Open System Interconnection Reference Model. OSI to understand the severity of the cyber attack on the IT and Systems Network, the analyst should review the seven layers of the OSI Reference Model as a guide for how data are transmitted over the network. The OSI Reference Model is an abstract representation of the data pathway that an adversary can exploit. For more information on the OSI Reference Model.
- **Identify and review forensics (Infrastructure Protection Packet Capture, Security Information and Event Management [SIEM] Forensic Integration Tool)**—The most common goal of performing network forensics or digital media analysis is to gain a better understanding of an event of interest by finding and analyzing the facts related to that event.
Computer and network forensics, or digital media analysis, has evolved to assure proper representation of computer crime evidentiary data in court.
The **National Institute of Standards and Technology (NIST) Cyber Digital Media Analysis Process** as shown in Fig. 1 describes the forensic or digital media analysis process in terms of collection, examination, analysis, and reporting.

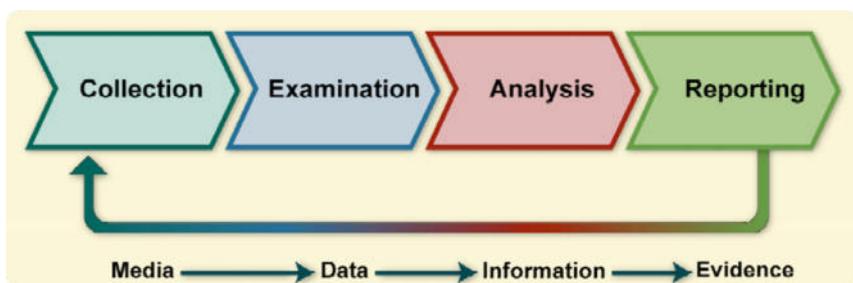


Fig. 1. NIST digital media analysis process [1] cyber

- **Review and identify network configuration vulnerabilities (e.g., sensors, firewalls, routers, host, IDS/intrusion protection system [IPS], host anti-virus [AV])**—The analysts should consider reviewing and identifying critical network vulnerabilities during the initial stage of the cyber-incident analysis (if available);

these may include misconfigured firewalls, sensors, routers, IDSs and IPSs, and host AV software, as well as risks associated with vendor-supplied software, risks associated with the network, and systems administration errors. If the information is not available in the initial analysis, efforts should be made to seek this information during a latter stage of an analysis, such as digital media analysis, to aid in the development of response and recovery plans. The following are other vulnerabilities that should be considered:

- *Network vulnerabilities*—Review of network vulnerabilities across information and control systems includes computers, network hardware/systems, OSs, and software applications that may have originated from a vendor system, system administration activities, and/or user activities.
- *Vendor vulnerabilities*—Vendor-originated vulnerabilities include software bugs, missing OS patches, vulnerable services, insecure default configurations, and Web applications.
- *System administration vulnerabilities*—System administration originated vulnerabilities include incorrect or unauthorized system configuration changes and lack of password-protection policies.
- *User vulnerabilities*—User-originated vulnerabilities include sharing of directories with unauthorized users, failure to run virus scanning software, and malicious activities such as introducing system backdoors [2].

3.3 Malware Analysis

Malware analysis entails comprehensive review of the network risk from an infrastructure disruption from cyber threats. The analysis provides a holistic view of the problem to assist the critical sector(s) in response to cyber threats. The analysis should consider the following:

- **Collect and analyze the data profile**—Collection and analysis of data should include information needed to analyze and manage critical infrastructure risks. The dataset should include addresses, points-of-contact, asset geo-location, and other information. This data should leverage geographic information system to visually represent the data on a map; it can be sorted by sector, risk, and priority. Output should provide tabular results of the critical infrastructure in question.
- **Perform security and vulnerability assessments**—This step should provide analysts a way to assess the vulnerabilities and consequences of the threat's impact on infrastructure assets.

4 Vulnerability/Consequence Analysis

4.1 Information Sharing

Secure, functioning, and resilient critical infrastructure requires the efficient exchange of information, including intelligence, between all levels of government and critical infrastructure sector owners and operators. This must facilitate the timely exchange of

threat and vulnerability information as well as information that allows for the development of a situational awareness capability during incidents [3].

To conduct vulnerability/consequence analysis of any particular sector or potential exploit, the analyst will collect relevant data on credible cyber threats and combine it with cyber-physical system information from owner/operators. Collection of data includes understanding potential cyber exploit details, potential impacts on the entity based on their information, and relevant mitigation steps.

4.2 Vulnerability/Consequence Analysis

To conduct a vulnerability/consequence analysis, a method for assessing and rating the risk (low, medium, and high) of possible cyber vulnerability for a specific critical infrastructure facility is needed. The risk is a function of the likelihood (probability) that a defined threat agent (adversary) who has the intent and capabilities can exploit specific cyber-physical vulnerability and create an impact (consequence). The risk induced by any given cyber vulnerability is influenced by a number of related indicators, including the following:

- **Policy and procedures**—Vulnerabilities are often introduced into network because of incomplete, inappropriate, or nonexistent security documentation, including policy and implementation guides, proper change management, and procedures.
- **Computer architecture and conditions**—Vulnerabilities in the network can occur due to flaws, misconfigurations, or poor maintenance (change management process) of their platforms, including hardware, OS, and network updated software/applications.
- **Network architecture**—Vulnerabilities in network may occur from flaws, misconfigurations, or poor administration of the networks and their connections with other networks.
- **Installed countermeasures**—Care must be taken before using tools such as an IDS or IPS to identify and protect networks from malware. Without assessment and proper software updates, these countermeasures may have an adverse impact on the production ICS.
- **Technical difficulty of the attack**—Attacks on the network (reliant on third-party contractor support due to technical complexities of the project) by sophisticated malware such as Stuxnet, which is designed to be stealthy, are examples of the technical difficulty of an attack.
- **Probability of detection**—Probability of detection is important because security is built on a layered defense. If there is a 3% chance that an exploit can make it into a network and a 4% chance an exploited system will go undetected, then there is only a 0.12% chance that someone will manage both feats at the same time. It is impossible to reduce the probability all the way to 0%, but it is possible to make the probability very small. Another example of probability of detection is the amount of time the adversary can remain in contact with the target or network without detection.
- **Consequences of the incident**—A cyber breach in some critical infrastructures can have significant physical impacts (personal injury and loss of life), as well as

economic (greater economic loss on the facility and or organization), and social (loss of national or public confidence in an organization) impacts.

- **Cost of the incident**—Undesirable incidents of any sort detract from the value of an organization, but safety and security incidents can have longer-term negative impacts than other types of incidents on all stakeholders, including employees, shareholders, customers, and the communities in which an organization operates [4]. For example, the *Hudson Valley Times* reported on February 28, 2013, the Central Hudson Gas & Electric Company on the cyber attack to their financial system. It was suspected that the cybersecurity breach may have compromised financial data of more than 100,000 of the utility's customers. The company offered potentially impacted customers a year of free credit monitoring and advised them to keep an eye out for suspicious activity on their bank accounts to avoid loss of public confidence [5].
- **Network and host information**—Cyber data collection involves review of network and host information collected from application servers, system log files, firewall log records, IDS, AV, malware detection automated tools, ISP log files, and interviews of people and departments involved in the incident. For example, the Network Mapper (Nmap) open-source tool for network exploration and security auditing can provide information about hosts available on the network, types of services.

5 Summary

This standard Cyber-Physical Mapping Framework will substantially enhance the cyber capabilities of organizations through Pre-Incident Planning Analysis, Incident Detection and Characterization, Vulnerability/Consequence Analysis and Incident Response and Recovery to prevent asymmetric attacks.

References

1. Kent, K., Chevalier, S., Grance, T., Dang, H.: Special Publication SP800-86, Guide to Integrating Forensic Techniques into Incident Response, NIST. <http://csrc.nist.gov/publications/nistpubs/800-86/SP800-86.pdf>. Accessed 25 March 2013
2. An Overview of Vulnerability Scanners. HKSAR (The Government of the Hong Kong Special Administrative Region). <http://www.infosec.gov.hk/english/technical/files/vulnerability.pdf>. Accessed 25 March 2013
3. Presidential Policy Directive—Critical Infrastructure Security and Resilience. Whitehouse.gov. <http://www.whitehouse.gov/the-press-office/2013/02/12/presidential-policy-directive-critical-infrastructure-security-and-resil>. Accessed 25 March 2013
4. Stouffer, K., Falco, J., Scarfone, K.: Special Publication SP800-82: Guide to Industrial Control Systems (ICS) Security, NIST. <http://csrc.nist.gov/publications/nistpubs/800-82/SP800-82-final.pdf>. Accessed 25 March 2013
5. Hack Attack. Hudson Valley Times. http://www.ulsterpublishing.com/view/full_story/21844700/article-Hack-attack-?. Accessed 25 March 2013

Author Index

A

- Aadil, Farhan, 131
Abrar, Sundus, 400
Adda, Mehdi, 892
Ahmad, Maaz Bin, 801
Ahmed, Musharif, 341
Ahmed, Syed Shahbaaz, 84
Ahmed, Zeeshan, 326
Ahn, Tae-Ki, 1153
Aiston, Jack, 759
Akogo, Darlington Ahiale, 152
Al-Dayel, Reham, 881
Alhwaiti, Yousef, 248, 299
AlJarraah, Abeer, 1016
Allen, Lee, 523
Allison, Mark, 422
Ambrosini, Luca, 453
Anwar, Farhat, 1106
Arafah, Mohammad, 881
Arora, Amrinder, 634
Asao, Daiki, 384
Askay, David, 373
Aslam, Zohra, 341
Atieh, Mirna, 892
Avramovic, Ivan, 586
Ayhan, Mahir, 616
Ayubi, Salahuddin, 75
Azam, Sheikh Shams, 55

B

- Baber, Junaid, 46
Bai, Yu, 739
Bakhtyar, Maheen, 46
Bakry, Saad Haj, 881

Balraj, Jeshreen, 224

- Barabasi, Stephan, 547
Barrera, James, 547
Beck, Micah, 667
Belyaev, Alexander, 1220
Bhalani, Prashant, 547
Booher, Duane, 781
Burhan ul Islam Khan, 1106

C

- Cambou, Bertrand, 781
Cao, Meng, 162
Cao, Wenxuan, 988
Chakravarty, Sumit, 1144
Chan, S., 912
Chandell, Sonali, 988, 1050
Chen, Chun-Ming, 1187
Chen, Hsiang-Chun, 11
Chen, Jim Q., 1
Cheng, Chun-Ho, 1187
Chhabria, Pooja, 1200
Choi, Hyeong-Ah, 616, 634
Choi, Yoonsuk, 739
Corti, Giancarlo, 453

D

- Dall, Zachary, 118
Dalvi, Preeti, 547
Danalis, Anthony, 667
Date, Prasanna, 98
Date, Susumu, 384
De Ceglia, Manuel, 1095
Deeka, Boriboon, 142
Deeka, Tanyaluk, 142

Devi, Varsha, 46

Dimiecik, Ryan, 547

Domnauer, Colin, 373

E

Ebrahimi Majd, Nahid, 1200, 1210

Endo, Arata, 384

Estoperez, Noel R., 284

F

Faheem, Osama, 881

Farook, Cassim, 224

Fatima, Taskeen, 341

Fiedelholtz, Glenn, 1244

Flak, Olaf, 479

G

Gentry, Eric, 1227

Ghalwash, Haitham, 691

Gkioulos, Vasileios, 1079

Goldberg, Saveli, 1220

Gonzalez, Carlos, 1069

Goryashchenko, Alexey, 1136

Graf Plessen, Mogens, 188

Gruhn, Volker, 1095

Guidi, Roberto, 453

H

Hamandi, Lama, 174

Harris, Erick, 373

Hawang, Seokhyun, 25

Hayajneh, Thaier, 868

Heal, Maher, 603

Ho, Nam, 1210

Hsu, Chia-Te, 11

Huang, Chun-Hsi, 691

Huang, Hsien-Chung, 11

Huang, Po-Lin, 1187

Hwang, Chi-Chuan, 1187

Hwang, Jong-Gyu, 1153

I

Ikram, Naveed, 951

Ishida, Kazuya, 384

Ivan, Berlocher, 25

J

Jayson, Renbert Jay R., 284

Jingji, Zang, 1050

Jingyao, Sun, 1050

K

Kagita, Mohan Krishna, 1038

Kamimura, Ryotaro, 211

Kammüller, Florian, 271

Kamruzzaman, Abu, 248, 299

Kashmar, Nadine, 892

Kasivajjala, Vamsi Chandra, 55

Katt, Basel, 801, 1079

Katz, Gabriel, 1220

Kazimirova, Eva, 1123

Khalid, Madiha, 728

Khan, Muhammad Fahad, 131

Khan, Naurin Farooq, 951

Kido, Yoshiyuki, 384

Kim, Jai-Eun, 25

Kim, Kyeong-Hee, 1153

Kim, Seonhghyun, 25

Kiong, Loo Chu, 400

Klaylat, Samira, 174

Klein, James, 713

Kostanic, Ivica, 1165

Kumar, Ayush, 847

L

Lee, Tae-Hyung, 1153

Leider, Avery, 299, 357, 547

Leng, Jing, 262

Levinson, Bruce, 541

Li, Chao-Chin, 1187

Li, Dancheng, 162

Li, Jingpeng, 603

Li, Qi, 33

Li, Yuting, 262

Liang, Chi-Hsiu, 1187

Liang, Yu-Pei, 1159

Lien, Yi-Han, 1159

Lim, Hock Chuan, 533

Lim, Teng Joon, 847

Liñan, Ernesto, 1069

Luo, Ling, 162

Luszczek, Piotr, 667

Lyu, Frank, 1227

M

Malik, Saud Ahmed, 131

Maqsood, Muazzam, 131

Martin, Derek, 466

McIntyre, Ryan, 1227

Mehraj, Tehseen, 1106

Metcalf, Lynn, 373

Michael Franklin, D., 312, 466

Mir, Roohie Naaz, 1106

Moctezuma, Luis Alfredo, 830

Molinás, Marta, 830

Moncayo Carreño, Oscar, 1004

Mondrosch, John, 547

Moore, Terry, 667

- Mujahid, Umar, 728
Mwim, Stephen Odirachukwu, 438
- N**
Najam-ul-Islam, Muhammad, 728
Naqvi, Syed Fawad Hussain, 75
Nardelli, Robert, 118
Nasim, Ammara, 75
Nazir, Shah, 964
Ngan, Chew Yee, 326
Nguyen, Thu, 1210
Ni, Tian-Yi, 988
Noor, Waheed, 46
- O**
Oita, Marilena, 235
Olanrewaju, Rashidah Funke, 1106
On-rit, Surajate, 142
Osman, Ziad, 174
Oviedo, Byron, 1004
- P**
Pagidimarri, Venkatesh, 55
Palmer, Xavier-Lewis, 152
Patton, Robert, 98
Peters, Nia, 501
Peterson, Karl, 547
Philabaum, Christopher, 781
Pistorius, Tana, 438
Potok, Thomas, 98
Pramanik, Ankita, 1144
Pranulis, Justin, 326
- Q**
Qian, Kai, 706
- R**
Rahman, Hanif Ur, 964
Raju, Manoj, 55
Rangaswamaiah, Chaitra, 739
Rehman, Ateeq Ur, 964
Rehman, Izaz Ur, 964
Rizzo, Nicola, 453
Robinson, Melvin, 422
Rosenberg, Louis, 373
Rossi, Giovanni, 564, 645
Rusin, Grant, 422
- S**
Sabijon, Caezar Johnlery T., 284
Sammad, Abdul, 46
Sattar, Kashif, 801
Sawant, Nimish, 547
Sayles, Verlyn C., 284
- Schuman, Catherine, 98
Shafiq ur Rehman, 1095
Shafique, Usman, 975
Shakir, Zaenab, 1165
Shashank Karrthikeyaa, A. S., 84
Shehab, Mohamed, 1016
Shen, Chen, 634
Shen, Min-Hong, 1159
Shi, Yong, 706
Shih, Wei-Kuan, 1159
Shimojo, Shinji, 384
Skevoulis, Sotiris, 118
Soltys, Michael, 1227
Sterling, Grigoriy, 1123
Stolmeier, Jacob, 1210
Sun, Zijing, 988
- T**
Tahir, Ghalib Ahmad, 400
Tappert, Charles C., 248, 299, 547
Telesca, Donald A., 781
Temkin, Anatoly, 1220
Thiruvengadam, Raghavendran, 84
Tomar, Anjali, 1200
Tran, Binh, 728
Tsai, Meng-Hsiun, 11
- U**
Uddin Ahmed, Kafil, 46
Uddin, Nizam, 964
Ursell, Steven, 868
- V**
Velasco, Lemuel Clark P., 284
Vijayaraghavan, Vineeth, 84
- W**
Wakatani, Akiyoshi, 1176
Walcott, Kristen R., 713
Weisburd, Ben, 1220
Westfall, Lewis, 357
Willcox, Gregg, 373
Windridge, David, 271
Wu, Hong-Lin, 1187
Wu, Jimmy Ming-Tai, 11
- Y**
Yaacob, Mashkuri, 1106
Yamin, Muhammad Mudassar, 801
Yamiun, Muhammad Mudassar, 1079
Yang, Chane-Yuan, 1187
Yang, Jiayi, 988
Yang, Seung-Won, 25

Yun, Mira, 634
Yunnan, Yu, 1050

Z

Zafar, Saad, 341
Zafar, Zeeshan, 75
Zahur, Shorahbeel Bin, 975
Zambrano-Vega, Cristian, 1004
Zantout, Rached, 174

Zec, Josko, 1165
Zeeshan, Saman, 326
Zhang, Bailu, 988
Zhang, Chaohe, 162
Zhang, Miaomiao, 935
Zhang, Youshan, 33
Zheng, Qingping, 162
Zhipeng, Zhang, 1050