

Hocine Cherifi · Sabrina Gaito ·
José Fernando Mendes · Esteban Moro ·
Luis Mateus Rocha *Editors*

Complex Networks and Their Applications VIII

Volume 2 Proceedings of the Eighth
International Conference on Complex
Networks and Their Applications
COMPLEX NETWORKS 2019

Hocine Cherifi · Sabrina Gaito ·
José Fernando Mendes ·
Esteban Moro · Luis Mateus Rocha
Editors

Complex Networks and Their Applications VIII

Volume 2 Proceedings of the Eighth International Conference on Complex Networks and Their Applications
COMPLEX NETWORKS 2019



Springer

Contents

Network Analysis

Characterizing the Hypergraph-of-Entity Representation Model	3
José Devezas and Sérgio Nunes	
Lexical Networks and Lexicon Profiles in Didactical Texts for Science Education	15
Ismo T. Koponen and Maija Nousiainen	
Legal Information as a Complex Network: Improving Topic Modeling Through Homophily	28
Kazuki Ashihara, Chenhui Chu, Benjamin Renoust, Noriko Okubo, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara	
Graph-Based Fraud Detection with the Free Energy Distance	40
Sylvain Courtain, Bertrand Lebichot, Ilkka Kivimäki, and Marco Saerens	
Visualizing Structural Balance in Signed Networks	53
Edoardo Galimberti, Chiara Madeddu, Francesco Bonchi, and Giancarlo Ruffo	
Spheres of Legislation: Polarization and Most Influential Nodes in Behavioral Context	66
Andrew C. Phillips, Mohammad T. Irfan, and Luca Ostertag-Hill	
Why We Need a Process-Driven Network Analysis	81
Mareike Bockholt and Katharina A. Zweig	
Gender's Influence on Academic Collaboration in a University-Wide Network	94
Logan McNichols, Gabriel Medina-Kim, Viet Lien Nguyen, Christian Rapp, and Theresa Migler	
Centrality in Dynamic Competition Networks	105
Anthony Bonato, Nicole Eikmeier, David F. Gleich, and Rehan Malik	

Investigating Saturation in Collaboration and Cohesiveness of Wikipedia Using Motifs Analysis	117
Anita Chandra and Abyayananda Maiti	
ESA-T2N: A Novel Approach to Network-Text Analysis	129
Yassin Taskin, Tobias Hecking, and H. Ulrich Hoppe	
Understanding Dynamics of Truck Co-Driving Networks	140
Gerrit Jan de Bruin, Cor J. Veenman, H. Jaap van den Herik, and Frank W. Takes	
Characterizing Large Scale Land Acquisitions Through Network Analysis	152
Roberto Interdonato, Jeremy Bourgoin, Quentin Grislain, Matteo Zignani, Sabrina Gaito, and Markus Giger	
A Network-Based Approach for Reducing Test Suites While Maintaining Code Coverage	164
Misael Mongiovì, Andrea Fornaia, and Emiliano Tramontana	
Structural Network Measure	
Detection of Conflicts of Interest in Social Networks	179
Saadie Albane, Hachem Slimani, and Hamamache Kheddouci	
Comparing Spectra of Graph Shift Operator Matrices	191
Johannes F. Lutzeyer and Andrew T. Walden	
Induced Edge Samplings and Triangle Count Distributions in Large Networks	203
Nelson Antunes, Tianjian Guo, and Vladas Pipiras	
Spectral Vertex Sampling for Big Complex Graphs	216
Jingming Hu, Seok-Hee Hong, and Peter Eades	
Attributed Graph Pattern Set Selection Under a Distance Constraint	228
Henry Soldano, Guillaume Santini, and Dominique Bouthimon	
On the Relation of Edit Behavior, Link Structure, and Article Quality on Wikipedia	242
Thorsten Ruprechter, Tiago Santos, and Denis Helic	
Establish the Expected Number of Injective Motifs on Unlabeled Graphs Through Analytical Models	255
Emanuele Martorana, Giovanni Micale, Alfredo Ferro, and Alfredo Pulvirenti	
Network Rewiring Dynamics to Form Clustered Strategic Networks . . .	268
Faisal Ghaffar and Neil Hurley	

Measuring Local Assortativity in the Presence of Missing Values	280
Jan van der Laan and Edwin de Jonge	

The Case for Kendall's Assortativity	291
Paolo Boldi and Sebastiano Vigna	

Modeling Human Behavior

Modelling Opinion Dynamics and Language Change: Two Faces of the Same Coin	305
Jérôme Michaud	

Networks of Intergenerational Mobility	316
Tanya Araújo, David Neves, and Francisco Louçã	

Inequality in Learning Outcomes: Unveiling Educational Deprivation Through Complex Network Analysis	325
Harvey Sánchez-Restrepo and Jorge Louçã	

'I Ain't Like You' A Complex Network Model of Digital Narcissism	337
Fakhra Jabeen, Charlotte Gerritsen, and Jan Treur	

Text Sentiment in the Age of Enlightenment	350
Philipp Koncar and Denis Helic	

A Gravity-Based Approach to Connect Food Retailers with Consumers for Traceback Models of Food-Borne Diseases	363
Tim Schlaich, Hanno Friedrich, and Abigail Horn	

The Effect of Social Media on Shaping Individuals Opinion Formation	376
Semra Gündüç	

A Network-Based Analysis of International Refugee Migration Patterns Using GERGMs	387
Katherine Abramski, Natallia Katenka, and Marc Hutchison	

Social Networks

Friendship Formation in the Classroom Among Elementary School Students	403
Raúl Duarte-Barahona, Ezequiel Arceo-May, and Rodrigo Huerta-Quintanilla	

Impact of Natural and Social Events on Mobile Call Data Records – An Estonian Case Study	415
Hendrik Hiir, Rajesh Sharma, Anto Aasa, and Erki Saluveer	

Describing Alt-Right Communities and Their Discourse on Twitter During the 2018 US Mid-term Elections	427
Añgel Panizo-LLedot, Javier Torregrosa, Gema Bello-Orgaz, Joshua Thorburn, and David Camacho	
Social Network Analysis of Sicilian Mafia Interconnections	440
Annamaria Ficara, Lucia Cavallaro, Pasquale De Meo, Giacomo Fiumara, Salvatore Catanese, Ovidiu Bagdasar, and Antonio Liotta	
Who Ties the World Together? Evidence from a Large Online Social Network	451
Guanghua Chi, Bogdan State, Joshua E. Blumenstock, and Lada Adamic	
Temporal Networks	
Comparing Temporal Graphs Using Dynamic Time Warping	469
Vincent Froese, Brijnesh Jain, Rolf Niedermeier, and Malte Renken	
Maximizing the Likelihood of Detecting Outbreaks in Temporal Networks	481
Martin Sterchi, Cristina Sarasua, Rolf Grüter, and Abraham Bernstein	
Efficient Computation of Optimal Temporal Walks Under Waiting-Time Constraints	494
Anne-Sophie Himmel, Matthias Bentert, André Nichterlein, and Rolf Niedermeier	
Roles in Social Interactions: Graphlets in Temporal Networks Applied to Learning Analytics	507
Raphaël Charbey, Laurent Brisson, Cécile Bothorel, Philippe Ruffieux, Serge Garlatti, Jean-Marie Gilliot, and Antoine Mallégol	
Enumerating Isolated Cliques in Temporal Networks	519
Hendrik Molter, Rolf Niedermeier, and Malte Renken	
Networks in Finance and Economics	
A Partially Rational Model for Financial Markets: The Role of Social Interactions on Herding and Market Inefficiency	535
Lorenzo Giannini, Fabio Della Rossa, and Pietro DeLellis	
A Network Structure Analysis of Economic Crises	547
Maximilian Göbel and Tanya Araújo	
A Multiplier Effect Model for Price Stabilization Networks	561
Jun Kiniwa and Hiroaki Sandoh	
Sector Neutral Portfolios: Long Memory Motifs Persistence in Market Structure Dynamics	573
Jeremy D. Turiel and Tomaso Aste	

Beyond Fortune 500: Women in a Global Network of Directors	586
Anna Evtushenko and Michael T. Gastner	
Supplier Impersonation Fraud Detection in Business-To-Business Transaction Networks Using Self-Organizing Maps	599
Rémi Canillas, Omar Hasan, Laurent Sarrat, and Lionel Brunie	
Network Shapley-Shubik Power Index: Measuring Indirect Influence in Shareholding Networks	611
Takayuki Mizuno, Shohei Doi, and Shuhei Kurizaki	
“Learning Hubs” on the Global Innovation Network	620
Michael A. Verba	
Global Transitioning Towards a Green Economy: Analyzing the Evolution of the Green Product Space of the Two Largest World Economies	633
Seyyedmilad Talebzadehhosseini, Steven R. Scheinert, and Ivan Garibay	
Performance of a Multi-layer Commodity Flow Network in the United States Under Disturbance	645
Susana Garcia, Sarah Rajtmajer, Caitlin Grady, Paniz Mohammadpour, and Alfonso Mejia	
Empirical Analysis of a Global Capital-Ownership Network	656
Sammy Khalife, Jesse Read, and Michalis Vazirgiannis	
Multilayer Networks	
Patterns of Multiplex Layer Entanglement Across Real and Synthetic Networks	671
Blaž Škrlj and Benjamin Renoust	
Introducing Multilayer Stream Graphs and Layer Centralities	684
P. Parmentier, T. Viard, B. Renoust, and J.-F. Baffier	
Better Late than Never: A Multilayer Network Model Using Metaplasticity for Emotion Regulation Strategies	697
Nimat Ullah and Jan Treur	
Comparison of Opinion Polarization on Single-Layer and Multiplex Networks	709
Sonoko Kimura, Kimitaka Asatani, and Toshiharu Sugawara	
Learning of Weighted Multi-layer Networks via Dynamic Social Spaces, with Application to Financial Interbank Transactions	722
Chris U. Carmona and Serafin Martinez-Jaramillo	

Influence of Countries in the Global Arms Transfers Network: 1950–2018	736
Sergey Shvydun	
Biological Networks	
Network Entropy Reveals that Cancer Resistance to MEK Inhibitors Is Driven by the Resilience of Proliferative Signaling	751
Joel Maust, Judith Leopold, and Andrej Bugrim	
Computational Modelling of TNFα Pathway in Parkinson's Disease – A Systemic Perspective	762
Hemalatha Sasidharakurup, Lakshmi Nair, Kanishka Bhaskar, and Shyam Diwakar	
Understanding the Progression of Congestive Heart Failure of Type 2 Diabetes Patient Using Disease Network and Hospital Claim Data	774
Md Ekramul Hossain, Arif Khan, and Shahadat Uddin	
Networks of Function and Shared Ancestry Provide Insights into Diversification of Histone Fold Domain in the Plant Kingdom	789
Amish Kumar and Gitanjali Yadav	
In-silico Gene Annotation Prediction Using the Co-expression Network Structure	802
Miguel Romero, Jorge Finke, Mauricio Quimbaya, and Camilo Rocha	
Network Neuroscience	
Linear Graph Convolutional Model for Diagnosing Brain Disorders	815
Zarina Rakhimberdina and Tsuyoshi Murata	
Adaptive Network Modeling for Criterial Causation	827
Jan Treur	
Network Influence Based Classification and Comparison of Neurological Conditions	842
Ruaridh Clark, Niia Nikolova, Malcolm Macdonald, and William McGeown	
Characterization of Functional Brain Networks and Emotional Centers Using the Complex Networks Techniques	854
Richa Tripathi, Dyutiman Mukhopadhyay, Chakresh Kumar Singh, Krishna Prasad Miyapuram, and Shivakumar Jolad	
Topological Properties of Brain Networks Underlying Deception: fMRI Study of Psychophysiological Interactions	868
Irina Knyazeva, Maxim Kireev, Ruslan Masharipov, Maya Zheltyakova, Alexander Korotkov, Makarenko Nikolay, and Medvedev Svyatoslav	

Urban Networks and Mobility

- Functional Community Detection in Power Grids** 883
Xiaoliang Wang, Fei Xue, Shaofeng Lu, Lin Jiang, and Qigang Wu

- Comparing Traditional Methods of Complex Networks
Construction in a Wind Farm Production Analysis Problem** 895
Sara Cornejo-Bueno, Mihaela Ioana Chidean, Antonio J. Caamaño,
Luís Prieto, and Sancho Salcedo-Sanz

- Quantifying Life Quality as Walkability on Urban Networks:
The Case of Budapest** 905
Luis Guillermo Natera Orozco, David Deritei, Anna Vancso,
and Orsolya Vasarhelyi

- A Network Theoretical Approach to Identify Vulnerabilities
of Urban Drainage Networks Against Structural Failures** 919
Paria Hajiamoosha and Christian Urich

- Mining Behavioural Patterns in Urban Mobility Sequences
Using Foursquare Check-in Data from Tokyo** 931
Galina Deeva, Johannes De Smedt, Jochen De Weerdt,
and María Óskarsdóttir

- Temporal Analysis of a Bus Transit Network** 944
Manju Manohar Manjalavil, Gitakrishnan Ramadurai,
and Balaraman Ravindran

- Modeling Urban Mobility Networks Using Constrained
Labeled Sequences** 955
Stephen Eubank, Madhav Marathe, Henning Mortveit, and Anil Vullikanti

Quantifying Success through Social Network Analysis

- A Network Approach to the Formation of Self-assembled Teams** 969
Rustom Ichhaporia, Diego Gómez-Zará, Leslie DeChurch,
and Noshir Contractor

- Predicting Movies' Box Office Result - A Large Scale Study
Across Hollywood and Bollywood** 982
Risko Ruus and Rajesh Sharma

- Using Machine Learning to Predict Links and Improve Steiner
Tree Solutions to Team Formation Problems** 995
Peter Keane, Faisal Ghaffar, and David Malone

- Scientometrics for Success and Influence in the Microsoft
Academic Graph** 1007
George Panagopoulos, Christos Xypolopoulos, Konstantinos Skianis,
Christos Giatsidis, Jie Tang, and Michalis Vazirgiannis

Testing Influence of Network Structure on Team Performance Using STERGM-Based Controls	1018
Brennan Antone, Aryaman Gupta, Suzanne Bell, Leslie DeChurch, and Noshir Contractor	
Author Index	1031

Network Analysis



Characterizing the Hypergraph-of-Entity Representation Model

José Devezas^(✉) and Sérgio Nunes

INESC TEC and Faculty of Engineering,
University of Porto, Rua Dr. Roberto Frias, s/n, 4200-465, Porto, Portugal
{jld,ssn}@fe.up.pt

Abstract. The hypergraph-of-entity is a joint representation model for terms, entities and their relations, used as an indexing approach in entity-oriented search. In this work, we characterize the structure of the hypergraph, from a microscopic and macroscopic scale, as well as over time with an increasing number of documents. We use a random walk based approach to estimate shortest distances and node sampling to estimate clustering coefficients. We also propose the calculation of a general mixed hypergraph density based on the corresponding bipartite mixed graph. We analyze these statistics for the hypergraph-of-entity, finding that hyperedge-based node degrees are distributed as a power law, while node-based node degrees and hyperedge cardinalities are log-normally distributed. We also find that most statistics tend to converge after an initial period of accentuated growth in the number of documents.

Keywords: Hypergraph-of-entity · Combined data · Indexing · Representation model · Hypergraph analysis · Characterization

1 Introduction

Complex networks have frequently been studied as graphs, but only recently has attention been given to the study of complex networks as hypergraphs [11]. The hypergraph-of-entity [10] is a hypergraph-based model used to represent combined data [4, Sect. 2.1.3]. That is, it is a joint representation of corpora and knowledge bases, integrating terms, entities and their relations. It attempts to solve, by design, the issues of representing combined data through inverted indexes and quad indexes. The hypergraph-of-entity, together with its random walk score [10, Sect. 4.2.2], is also an attempt to generalize several tasks of entity-oriented search. This includes ad hoc document retrieval and ad hoc entity retrieval, as well as the recommendation-alike tasks of related entity finding and entity list completion. However, there is a tradeoff. One one side, the random walk score acts as a general ranking function. On the other side, it performs below traditional baselines like TF-IDF. Since ranking is particularly dependent on the structure of the hypergraph, a characterization is a fundamental step towards improving the representation model and, with it, the retrieval performance.

Accordingly, our focus was on studying the structural properties of the hypergraph. This is a task that presents some challenges, both from a practical sense and from a theoretical perspective. While there are many tools [5, 9] and formats [8, 17] for the analysis and transfer of graphs, hypergraphs still lack clear frameworks to perform these functions, making their analysis less trivial. Even formats like GraphML [8] only support undirected hypergraphs. Furthermore, there is still an ongoing study of several aspects of hypergraphs, some of which are trivial in graph theory. For example, the adjacency matrix is a well-established representation of a graph, however recent work is still focusing on defining an adjacency tensor for representing general hypergraphs [21]. As a scientific community, we have been analyzing graphs since 1735 and, even now, innovative ideas in graph theory are still being researched [1]. However, hypergraphs are much younger, dating from 1970 [6], and thus there are still many open challenges and contribution opportunities.

In this work, we take a practical application of hypergraphs, the hypergraph-of-entity, as an opportunity to establish a basic framework for the analysis of hypergraphs. In Sect. 2, we begin by providing an overview on the analysis of the inverted index, knowledge bases and hypergraphs, covering the three main aspects of the hypergraph-of-entity. In Sect. 3, we describe our characterization approach, covering shortest distance estimation with random walks and clustering coefficient estimation with node sampling, as well as proposing a general mixed hypergraph density formula by establishing a parallel with the corresponding bipartite mixed graph. In Sect. 4, we present the results of a characterization experiment of the hypergraph-of-entity for a subset of the INEX 2009 Wikipedia collection and, in Sect. 5, we present the conclusions and future work.

2 Reference Work

The hypergraph-of-entity is a representation model for indexing combined data, jointly modeling unstructured textual data from corpora and structured interconnected data from knowledge bases. As such, before analyzing a hypergraph from this model, we surveyed existing literature on inverted index analysis, as well knowledge base analysis. We then surveyed literature specifically on the analysis of hypergraphs, particularly focusing on statistics like the clustering coefficient, the shortest path lengths and the density.

Analyzing Inverted Indexes. There are several models based on the inverted index that combine documents and entities [3, 7] and that are comparable with the hypergraph-of-entity. There has also been work that analyzed the inverted index, particularly regarding query evaluation speed and space requirements [23, 25].

Voorhees [23] compared the efficiency of the inverted index with the top-down cluster search. She analyzed the storage requirements of four test collections, measuring the total number of documents and terms, as well as the average number of terms per document. She then analyzed the disk usage per collection, measuring the number of bytes for document vectors and the inverted index.

Finally, she measured CPU time in number of instructions and the I/O time in number of data pages accessed at least once, also including the query time in seconds.

Zobel et al. [25] took a similar approach to compare the inverted index and signature files. First, they characterized two test collections, measuring size in megabytes, number of records and distinct words, as well as the record length, and the number of words, distinct words and distinct words without common terms per record. They also analyzed disk space, memory requirements, ease of index construction, ease of update, scalability and extensibility.

For the hypergraph-of-entity characterization, we do not focus on measuring efficiency, but rather on studying the structure and size of the hypergraph.

Analyzing Knowledge Bases. Studies have been made to characterize the entities and triples in knowledge bases. In particular, given RDF’s graph structure, we are interested in understanding which statistics are relevant for instance to discriminate between the typed nodes.

Halpin [16] took advantage of Microsoft’s *Live.com* query log to reissue entity and concept queries over their FALCON-S semantic web search engine. They then studied the results, characterizing their source, triple structure, RDF and OWL classes and properties, and the power-law distributions of the number of URIs, both returned as results and as part of the triples linking to the results. They focused mostly on measuring the frequency of different elements or aggregations (e.g., top-10 domain names for the URIs, most common data types, top vocabulary URIs).

Ge et al. [14] defined an object link graph based on the graph induced by the RDF graph, based on paths linking objects (or entities), as long as they are either direct or established through blank nodes. They then studied this graph for the Falcons Crawl 2008 and 2009 datasets (FC08 and FC09), which included URLs from domains like bio2rdf.org or dbpedia.org. They characterized the object link graph based on density, using the average degree as an indicator, as well as connectivity, analyzing the largest connected component and the diameter. They repeated the study for characterizing the structural evolution of the object link graph, as well its domain-specific structures (according to URL domains). Comparing two snapshots of the same data enabled them to find evidence of the scale-free nature of the network. While the graph almost doubled in size from FC08 to FC09, the average degree remained the same and the diameter actually decreased.

Fernandez et al. [12] focused on studying the structural features of RDF data, identifying redundancy through common structural patterns, proposing several specific metrics for RDF graphs. In particular, they proposed several subject and object degrees, accounting for the number of links from/to a given subject/object (outdegree and indegree), the number of links from a $\langle \text{subject}, \text{predicate} \rangle$ (partial outdegree) and to a $\langle \text{predicate}, \text{object} \rangle$ (partial indegree), the number of distinct predicates from a subject (labeled outdegree) and to an object (labeled indegree), and the number of objects linked from a subject through a single predicate (direct outdegree), as well as the number of subjects linking to an object through a single

predicate (direct indegree). They also measured predicate degree, outdegree and indegree. They proposed common ratios to account for shared structural roles of subjects, predicates and objects (e.g., subject-object ratio). Global metrics were also defined for measuring the maximum and average outdegree of subject and object nodes for the whole graph. Another analysis approach was focused on the predicate lists per subject, measuring the ratio of repeated lists and their degree, as well as the number of lists per predicate. Finally, they also defined several statistics to measure typed subjects and classes, based on the *rdf:type* predicate.

While we study a hypergraph that jointly represents terms, entities and their relations, we focus on a similar characterization approach, that is more based on structure and less based on measuring performance.

Analyzing Hypergraphs. Hypergraphs [6] have been around since the 1970s and, because they are able to capture higher-order relations, there are either conceptually different or multiple counterparts to the equivalent graph statistics. Take for instance the node degree. While graphs only have a node degree, indegree and outdegree, hypergraphs can also have a hyperedge degree, which is the number of nodes in a hyperedge [18]. The hyperedge degree also exists for directed hyperedges, in the form of a tail degree and a head degree¹. The tail degree is based on the cardinality of the source node set and the head degree is based on the cardinality of the target node set. In this work we will rely on the degree, clustering coefficient, average path length, diameter and density to characterize the hypergraph-of-entity.

Ribeiro et al. [22] proposed the use of multiple random walks to find shortest paths in power law networks. They found that random walks had the ability to observe a large fraction of the network and that two random walks, starting from different nodes, would intersect with a high probability. Glabowski et al. [15] contributed with a shortest path computation solution based on ant colony optimization, clearly structuring it as pseudocode, while providing several configuration options. Parameters included the number of ants, the influence of pheromones and other data in determining the next step, the speed of evaporation of the pheromones, the initial, minimum and maximum pheromone levels, the initial vertex and an optional end vertex. Li [19] studied the computation of shortest paths in electric networks based on random walk models and ant colony optimization, proposing a current reinforced random walk model inspired by the previous two. In this work, we also use a random walk based approach to approximate shortest paths and estimate the average path length and diameter of the graph.

Gallagher and Goldberg [13, Eq. 4] provide a comprehensive review on clustering coefficients for hypergraphs. The proposed approach for computing the clustering coefficient in hypergraphs accounted for a pair of nodes, instead of a single node, which is more frequent in graphs. Based on these two-node clustering coefficients, the node cluster coefficient was then calculated. Two-node

¹ Tail and head is used in analogy to an arrow, not a list.

clustering coefficients measured the fraction of common hyperedges between two nodes, through the intersection of the incident hyperedge sets for the two nodes. It then provided different kinds of normalization approaches, either based on the union, the maximum or minimum cardinality, or the square root of the product of the cardinalities of the hyperedge sets. The clustering coefficient for a node was then computed based on the average two-node clustering coefficient for the node and its neighbors.

The codegree Turán density [20] can be computed for a family \mathcal{F} of k -uniform hypergraphs, also known as k -graphs. It is calculated based on the codegree Turán number (the extremal number), which takes as parameters the number of nodes n and the family \mathcal{F} of k -graphs. In turn, the codegree Turán number is calculated based on the minimum number of nodes, taken from all sets of $r - 1$ vertices of each hypergraph H that, when united with an additional vertex, will form a hyperedge from H . The codegree density for a family \mathcal{F} of hypergraphs is then computed based on $\limsup_{n \rightarrow \infty} \frac{\text{co-ex}(n, \mathcal{F})}{n}$. Since this was the only concept of density we found associated with hypergraphs or, more specifically, a family of k -uniform hypergraphs, we opted to propose our own density formulation (Sect. 3). The hypergraph-of-entity is a single general mixed hypergraph. In other words, it is not a family of hypergraphs, it contains hyperedges of multiple degrees (it's not k -uniform, but general) and it contains undirected and directed hyperedges (it's mixed). Accordingly, we propose a density calculation based on the counterpart bipartite graph of the hypergraph, where hyperedges are translated to bridge nodes.

3 Characterization Approach

Graphs can be characterized at a microscopic, mesoscopic and macroscopic scale. The microscopic analysis is concerned with statistics at the node-level, such as the degree or clustering coefficient. The mesoscopic analysis is concerned with statistics and patterns at the subgraph-level, such as communities, network motifs or graphlets. The macroscopic analysis is concerned with statistics at the graph-level, such as average clustering coefficient or diameter. In this work, our analysis of the hypergraph is focused on the microscopic and macroscopic scales. We compute several statistics for the whole hypergraph, as well as for snapshot hypergraphs that depict growth over time. Some of these statistics are new to hypergraphs, when compared to traditional graphs. For instance, nodes in directed graphs have an indegree and an outdegree. However, nodes in directed hypergraphs have four degrees, based on incoming and outgoing nodes, as well as on incoming and outgoing hyperedges. While in graphs all edges are binary, leading to only one other node, in hypergraphs hyperedges are n -ary, leading to multiple nodes, and thus different degree statistics. While some authors use ‘degree’ to refer to node and hyperedge degrees [24, Sect. 4][18, Sect. Network Statistics in Hypergraphs], in this work we opted to use the ‘degree’ designation when referring to nodes and the ‘cardinality’ designation when referring to hyperedges. This is to avoid any confusion for instance between an “hyperedge-induced” node degree and a hyperedge cardinality.

For the whole hypergraph, we compute node degree distributions based on nodes and hyperedges, and hyperedge cardinality distributions. For snapshots, we compute average node degrees and hyperedge cardinalities. For both, we compute the estimated clustering coefficient, average path length and diameter, as well as the density and space usage statistics.

Estimating Shortest Distances with Random Walks. Ribeiro et al. [22] found that, in power law networks, there is a high probability that two random walk paths, usually starting from different nodes, will intersect and share a small fraction of nodes. We took advantage of this conclusion, adapting it to a hypergraph, in order to compute a sample of shortest paths and their length, used to estimate the average path length and diameter. We considered two (ordered) sets S_1 and S_2 of nodes sampled uniformly at random, each of size $s = |S_1| = |S_2|$. We then launched r random walks of length ℓ from each pair of nodes S_1^i and S_2^i . For a given pair of random walk paths, we iterated over the nodes in the path starting from S_1^i , until we found a node in common with the path starting from S_2^i . At that point, we merged the two paths based on the common node, discarding the suffix of the first path and the prefix of second path. As the number of iterations r increased, we progressively approximated the shortest path for the pair of nodes. This enabled us to generate a sample of approximated shortest path lengths, which could be used to compute the estimated diameter (its maximum) and the estimated average path length (its mean).

Estimating Clustering Coefficients with Node Sampling. In a graph, the clustering coefficient is usually computed for a single node and averaged over the whole graph. As shown by Gallagher and Goldberg [13, Sect. I.A.], in hypergraphs the clustering coefficient is computed, at the most atomic level, for a pair of nodes. The clustering coefficient for a node is then computed based on the averaged two-node clustering coefficients between the node and each of its neighbors (cf. Gallagher and Goldberg [13, Eq. 4]). Three options were provided for calculating the two-node clustering coefficient, one of them based on the Jaccard index between the neighboring hyperedges of each node [13, Eq. 1], which we use in this work.

As opposed to computing it for all nodes, we estimated the clustering coefficients for a smaller sample S of nodes. Furthermore, for each sampled node $s_i \in S$, we also sampled its neighbors $N_S(s_i)$ for computing the two-node clustering coefficients. We then applied the described equations to obtain the clustering coefficients for each node s_i and a global clustering coefficient based on the overall average.

Computing the Density of General Mixed Hypergraphs. A general mixed hypergraph is general (or non-uniform) in the sense that its hyperedges can contain an arbitrary number of vertices, and it is mixed in the sense that it can contain hyperedges that are either undirected and directed. We compute a hypergraph's density by analogy with its corresponding bipartite graph, which contains all nodes from the hypergraph, along with connector nodes representing the hyperedges.

Consider the hypergraph $H = (V, E)$, with $n = |V|$ nodes and $m = |E|$ hyperedges. Also consider the set of all undirected hyperedges E_U and directed hyperedges E_D , where $E = E_U \cup E_D$. Their subsets E_U^k and $E_D^{k_1, k_2}$ should also be respectively considered, where E_U^k is the subset of undirected hyperedges with k nodes and $E_D^{k_1, k_2}$ is the subset of directed hyperedges with k_1 tail (source) nodes, k_2 head (target) nodes and $k = k_1 + k_2$ nodes, assuming the hypergraph only contains directed hyperedges between disjoint tail and head sets. This means that the union of $E_U = E_U^1 \cup E_U^2 \cup E_U^3 \cup \dots$ and $E_D = E_D^{1,1} \cup E_D^{1,2} \cup E_D^{2,1} \cup E_D^{2,2} \cup \dots$ forms the set of all hyperedges E . We use it as a way to distinguish between hyperedges with different degrees. This is important because, depending on the degree k , the hyperedge will contribute differently to the density, when considering the corresponding bipartite graph. For instance, one undirected hyperedge with degree $k = 4$ will contribute with four edges to the density. Accordingly, we derive the density of a general mixed hypergraph as shown in Eq. 1.

$$D = \frac{2 \sum_k k |E_U^k| + \sum_{k_1, k_2} (k_1 + k_2) |E_D^{k_1, k_2}|}{2(n+m)(n+m-1)} \quad (1)$$

In practice, this is nothing more than a comprehensive combination of the density formulas for undirected and directed graphs. On one side, we consider the density of a mixed graph that should result of the combination of an undirected simple graph and a directed simple graph. That is, each pair of nodes can be connected, at most, by an undirected edge and two directed edges of opposing directions. On the other side, we use hypergraph notation to directly obtain the required statistics from the corresponding mixed bipartite graph, thus calculating the analogous density for a hypergraph.

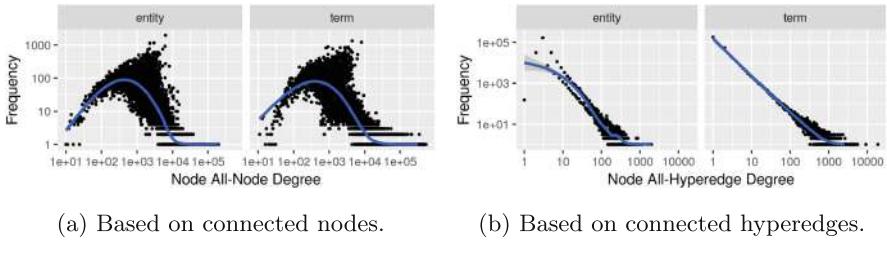
4 Analyzing the Hypergraph-of-Entity

We indexed a subset of the INEX 2009 Wikipedia collection given by the 7,487 documents appearing in the relevance judgments of 10 random topics. We then computed global statistics (macroscale), local statistics (microscale) and temporal statistics. Temporal statistics were based on an increasingly larger number of documents, by creating several snapshots of the index, through a ‘limit’ parameter, until all documents were considered.

Global Statistics. In Table 1, we present several global statistics about the hypergraph-of-entity, in particular the number of nodes and hyperedges, discriminated by type, the average degree, the average clustering coefficient, the average path length, the diameter and the density. The average clustering coefficient was computed based on a sample of 5,000 nodes and a sample of 100,000 neighbors for each of those nodes. The average path length and the diameter were computed based on a sample of shortest distances between 30 random pairs of nodes and the intersections of 1,000 random walks of length 1,000 launched from each element of the pair. Finally, the density was computed based on Eq. 1. As we can see, for the 7,487 documents the hypergraph contains 607,213 nodes and

Table 1. Global statistics

Statistic	Value	Statistic	Value	Statistic	Value
Nodes	607,213	Hyperedges	253,154	Avg. Degree	0.8338
<i>term</i>	323,672	Undirected	14,938	Avg. Clustering Coefficient	0.1148
<i>entity</i>	283,541	<i>document</i>	7,484	Avg. Path Length	8.3667
		<i>related_to</i>	7,454	Diameter	17
		Directed	238,216	Density	3.88e-06
		<i>contained_in</i>	238,216		

**Fig. 1.** Node degree distributions (log-log scale).

253,154 hyperedges of different types, an average degree lower than one (0.83) and a low clustering coefficient (0.11). It is also extremely sparse, with a density of 3.9e–06. Its diameter is 17 and its average path length is 8.4, almost double when compared to a social network like Facebook [2].

Local Statistics. Figure 1 illustrates the node degree distributions. In Fig. 1a, the node degree is based on the number of connected nodes, with the distribution approximating a log-normal behavior. In Fig. 1b, the node degree is based on the number of connected hyperedges, with the distribution approximating a power law. This shows the usefulness of considering both of the node degrees in the hypergraph-of-entity, as they are able to provide different information.

Figure 2 illustrates the hyperedge cardinality distribution. For *document* hyperedges, cardinality is log-normally distributed, while for *related_to* hyperedges the behavior is slightly different, with low cardinalities having a higher frequency than they would in a log-normal distribution. Finally, the cardinality distribution of *contained_in* hyperedges, while still heavy-tailed, presents an initial linear behavior, followed by a power law behavior. The maximum cardinality for this type of hyperedge is also 16, which is a lot lower when compared to *document* hyperedges and *related_to* hyperedges, which have cardinality 8,167 and 3,084, respectively. This is explained by the fact that *contained_in* hyperedges establish a directed connection between a set of terms and an entity that contains those terms, being limited by the maximum number of words in an entity.

Temporal Statistics. In order to compute temporal statistics, we first generated 14 snapshots of the index based on a limit L of documents, for $L \in \{1, 2, 3, 4, 5, 10, 25, 50, 100, 1000, 2000, 3000, 5000, 8000\}$.

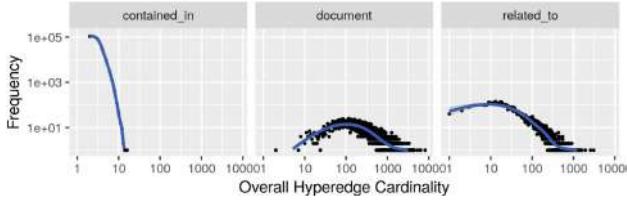


Fig. 2. Hyperedge cardinality based on the total number of nodes (log-log scale).

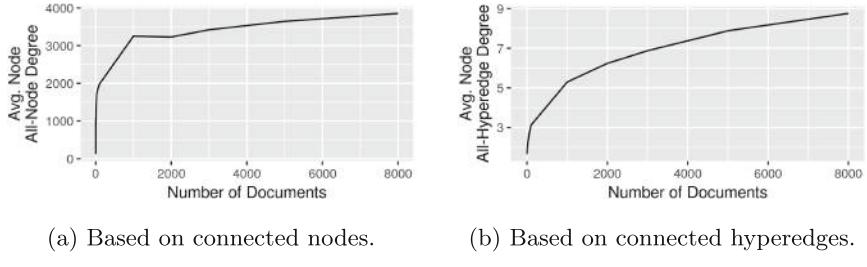


Fig. 3. Average node degree over time.

Figure 3 illustrates the node-based and hyperedge-based average node degrees over time (represented as the number of documents in the index at a given instant). As we can see, both functions tend to converge, however this is clearer for the node-based degree, reaching nearly 4,000 nodes, through only 9 hyperedges, on average. Figure 4 illustrates the average undirected hyperedge cardinality over time, with a convergence behavior that approximates 300 nodes per hyperedge, after rising to an average of 411.88 nodes for $L = 25$ documents.

Figure 5 illustrates the evolution of the average path length and the diameter of the hypergraph over time. For a single document, these values reached 126.1 and 491, respectively, while, for just two documents, they immediately lowered to 3.8 and 10. For higher values of L , both statistics increased slightly, reaching 7.2 and 15 for the maximum number of documents. Notice that these last values are equivalent to those computed in Table 1 (8.4 and 17, respectively), despite resulting in different quantities. This is due to the precision errors in our estimation approach, resulting in a difference of 1.2 and 2, respectively, which is tolerable when computation resources are limited. In Fig. 6, we illustrate the evolution of the clustering coefficient, which rapidly decreases from 0.59 to 0.11. The low average path length and clustering coefficient point towards a weak community structure, possibly due to the coverage of divergent topics. However, we would require a random generation model for hypergraphs, like the Watts–Strogatz model for graphs, in order to properly interpret the statistics.

Figure 7 illustrates the evolution of the density over time. The density is consistently low, starting from $1.37\text{e}{-}03$ and progressively decreasing to $3.91\text{e}{-}06$ as the number of documents increases. This shows that the hypergraph-of-entity

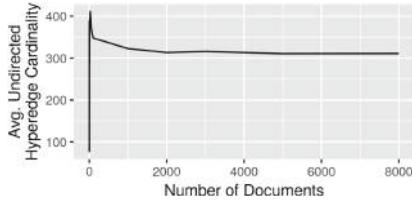


Fig. 4. Average hyperedge cardinality over time.

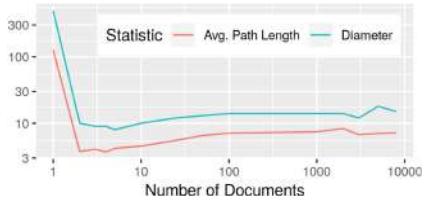


Fig. 5. Average estimated diameter and average shortest path over time.

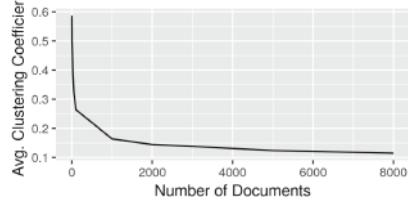


Fig. 6. Average estimated clustering coefficient over time.

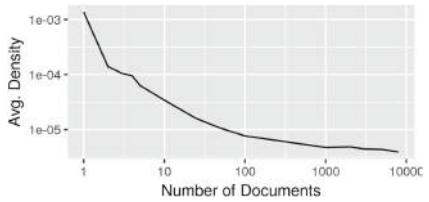


Fig. 7. Average density over time.

is an extremely sparse representation, with limited connectivity, which might benefit precision in a retrieval task.

Finally, we also measured the space usage of the hypergraph, both in disk and in memory. In disk, the smallest snapshot required 43.8 KiB for one document, while the largest snapshot required 181.9 MiB for the whole subset. Average disk space over all snapshots was $37.5 \text{ MiB} \pm 58.9 \text{ MiB}$. In memory, the smallest snapshot used 1.0 GiB for one document, including the overhead of the data structures, and the largest snapshot used 2.3 GiB for the whole subset. Average memory space over all snapshots was $1.3 \text{ GiB} \pm 461.1 \text{ MiB}$. Memory also grew faster for the first 1,000 documents, apparently leading to an expected convergence, although we could not observe it for such a small subset.

5 Conclusions

We have characterized the hypergraph-of-entity representation model, based on the structural properties of the hypergraph. We analyzed the node degree distributions, based on nodes and hyperedges, and the hyperedge cardinality distributions, illustrating their distinctive behavior. We also analyzed the temporal behavior, as documents were added to the index, studying average node degree and hyperedge cardinality, estimated average path length, diameter and clustering coefficient, as well as density and space usage requirements. Our contributions go beyond the characterization of the hypergraph-of-entity, as we show an application of two approximation approaches for computing statistics based on the shortest distance, as well as the clustering coefficient. We also proposed a

simple approach for computing the density of a general mixed hypergraph, based on the corresponding bipartite mixed graph.

In the future, we would like to further explore the computation of density, as the bipartite-based density we proposed, although useful, only accounts for hyperedges already in the hypergraph. We would also like to study the parameterization of the two estimation approaches we proposed, based on random walks and node sampling. Another open challenge in hypergraphs is the definition of random generation model, which would be useful to improve characterization. Finally, several opportunities also exist in the study of the hypergraph at a mesoscale, be it identifying communities, network motifs or graphlet, or exploring unique patterns to hypergraphs.

Acknowledgements. José Devezas is supported by research grant PD/BD/128160/2016, provided by the Portuguese national funding agency for science, research and technology, Fundação para a Ciéncia e a Tecnologia (FCT), within the scope of Operational Program Human Capital (POCH), supported by the European Social Fund and by national funds from MCTES.

References

1. Aparicio, D., Ribeiro, P., Silva, F.: Graphlet-orbit transitions (got): a fingerprint for temporal network comparison. *PLoS One* **13**, e0205497 (2018)
2. Backstrom, L., Boldi, P., Rosa, M., Ugander, J., Vigna, S.: Four degrees of separation. *CoRR* abs/1111.4570 (2011). <http://arxiv.org/abs/1111.4570>
3. Bast, H., Buchhold, B.: An index for efficient semantic full-text search. In: Proceedings of the 22nd ACM International Conference on Conference on Information and Knowledge Management, pp. 369–378 (2013). <https://doi.org/10.1145/2505515.2505689>
4. Bast, H., Buchhold, B., Haussmann, E., et al.: Semantic search on text and knowledge bases. *Found. Trends® Inf. Retrieval* **10**(2–3), 119–271 (2016)
5. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, San Jose, California, USA, 17–20 May 2009 (2009). <http://aaai.org/ocs/index.php/ICWSM/09/paper/view/154>
6. Berge, C.: Graphes et hypergraphes. Dunod, Paris (1970)
7. Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., Petrelli, D.: Hybrid search: Effectively combining keywords and semantic searches. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) European Semantic Web Conference, pp. 554–568. Springer, Berlin (2008)
8. Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., Marshall, M.S.: Graphml progress report structural layer proposal. In: Mutzel, P., Jünger, M., Leipert, S. (eds.) International Symposium on Graph Drawing, pp. 501–512. Springer, Berlin (2001)
9. Csardi, G., Nepusz, T., et al.: The igraph software package for complex network research. *InterJ. Complex Syst.* **1695**(5), 1–9 (2006)
10. Devezas, J., Nunes, S.: Hypergraph-of-entity: a unified representation model for the retrieval of text and knowledge. *Open Comput. Sci.* **9**(1), 103–127 (2019). <https://doi.org/10.1515/comp-2019-0006>

11. Estrada, E., Rodriguez-Velazquez, J.A.: Complex networks as hypergraphs. arXiv preprint physics/0505137 (2005)
12. Fernández, J.D., Martínez-Prieto, M.A., de la Fuente Redondo, P., Gutiérrez, C.: Characterizing RDF datasets. *J. Inf. Sci.* **1**, 1–27 (2016)
13. Gallagher, S.R., Goldberg, D.S.: Clustering coefficients in protein interaction hypernetworks. In: ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics, ACM-BCB 2013, Washington, DC, USA, 22-25 September 2013, p. 552 (2013). <https://doi.org/10.1145/2506583.2506635>
14. Ge, W., Chen, J., Hu, W., Qu, Y.: Object link structure in the semantic web. In: The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3 2010, Proceedings, Part II, pp. 257–271 (2010). https://doi.org/10.1007/978-3-642-13489-0_18
15. Gąłkowski, M., Musznicki, B., Nowak, P., Zwierzykowski, P.: Shortest path problem solving based on ant colony optimization metaheuristic. *Image Process. Commun.* **17**(1–2), 7–17 (2012)
16. Halpin, H.: A query-driven characterization of linked data. In: Proceedings of the WWW2009 Workshop on Linked Data on the Web, LDOW 2009, Madrid, Spain, 20 April 2009. http://ceur-ws.org/Vol-538/lдов2009_paper16.pdf
17. Himsolt, M.: GML: A portable graph file format. Technical report, Universität Passau (1997)
18. Klamt, S., Haus, U., Theis, F.J.: Hypergraphs and cellular networks. *PLoS Comput. Biol.* **5**(5), e1000385 (2009). <https://doi.org/10.1371/journal.pcbi.1000385>
19. Li, D.: Shortest paths through a reinforced random walk. Tech. rep., University of Uppsala (2011)
20. Mubayi, D., Zhao, Y.: Co-degree density of hypergraphs. *J. Comb. Theory, Ser. A* **114**(6), 1118–1132 (2007). <https://doi.org/10.1016/j.jcta.2006.11.006>
21. Ouvrard, X., Goff, J.L., Marchand-Maillet, S.: Adjacency and tensor representation in general hypergraphs part 1: e-adjacency tensor uniformisation using homogeneous polynomials. *CoRR* abs/1712.08189 (2017). <http://arxiv.org/abs/1712.08189>
22. Ribeiro, B.F., Basu, P., Towsley, D.: Multiple random walks to uncover short paths in power law networks. In: 2012 Proceedings IEEE INFOCOM Workshops, Orlando, FL, USA, 25-30 March 2012, pp. 250–255 (2012). <https://doi.org/10.1109/INFCOMW.2012.6193500>
23. Voorhees, E.M.: The efficiency of inverted index and cluster searches. In: SIGIR 1986, Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 8-10 September 1986, pp. 164–174 (1986). <https://doi.org/10.1145/253168.253203>
24. Yu, W., Sun, N.: Establishment and analysis of the supernetwork model for nanjing metro transportation system. *Complexity* **2018**, 4860531:1–4860531:11 (2018). <https://doi.org/10.1155/2018/4860531>
25. Zobel, J., Moffat, A., Ramamohanarao, K.: Inverted files versus signature files for text indexing. *ACM Trans. Database Syst.* **23**(4), 453–490 (1998). <https://doi.org/10.1145/296854.277632>



Lexical Networks and Lexicon Profiles in Didactical Texts for Science Education

Ismo T. Koponen^(✉) and Maija Nousiainen

Department of Physics, University of Helsinki, P.O. Box 64, 00014 Helsinki, Finland
ismo.koponen@helsinki.fi

Abstract. The lexical structure of language of science as it appears in teaching and teaching materials plays a crucial role in learning the language of science. We inspect here the lexical structure of two texts, written for didactic purposes and discussing the topic of wave-particle dualism as it is addressed in science education. The texts are analyzed as lexical networks of terms. The analysis is based on construction of stratified lexical networks, which allows us to analyze the lexical connections from the level of cotext (sentences) to context. Based on lexical networks, we construct lexicon profiles as they appear in two texts addressing the wave-particle dualism of electrons and photons. We demonstrate that the lexicon profiles of the two texts, although they discuss the same topic with similar didactic goals, nevertheless exhibit remarkable variation and differences. The consequences of such variation of lexicon profiles for practical teaching are discussed.

Keywords: Lexical network · Lexicon learning · Science education

1 Introduction

The structure of the language of science as it appears in science teaching and instruction is a widely researched topic owing to its importance in introducing students to the proper ways in which to use the vocabulary and syntax of scientific language [1–4]. The structure of scientific language in learning has often been approached by examining the structure of networks formed by terms that stand for concepts. Most often, this approach is referred as analysis of semantic networks, where meaning of words and terms is assumed to emerge from their mutual connections [2,5–8]. Such an approach to study role of scientific language in learning finds support from two directions: from analysis of scientific knowledge and from cognitive linguistics.

Analysis of the structure of scientific knowledge suggests, according to Kuhn [9], that it is built essentially in the form of lexical networks, as a lexicon of terms, where connections between the terms derive from contextualized instances of how terms are used. Such a notion of lexical networks and lexicons was central to the conception of scientific knowledge that Kuhn developed later in his research [10].

For Kuhn, lexicons also define scientific communities, because the individual members of the communities must share substantial parts of the lexicons to be able to communicate with and provide identity to the community [9, 10].

Focus on lexical networks also finds support from the research on how the meaning of words is learned [11, 12]. It has been pointed out that learning the meaning of words involves lexical networks made of names and words and the semantic connections between them, which build upon the conceptual system but are different from it. Thus, the conceptual system is not directly accessible in communication, in the form of written or spoken knowledge, but lexical networks and semantic connections provide access points to it [11]. In this view, the three levels - lexical, semantic and conceptual - are understood as distinct but related levels.

Some researchers maintain that representations encoded by language can be equated with semantic meanings [12], while some others see these linguistic and semantic structures as different [11]. From the viewpoint of lexicons and the role of lexical networks as adopted here, several studies, which have framed their targets as students' semantic networks, have in fact focused on the lexical networks rather than on semantic networks [2, 5–8, 13, 14]. The difference between the lexical and semantic networks, however, is not crucial here but for reasons of clarity, we have chosen to retain the distinction and frame our target as the analysis of lexical rather than semantic networks.

The lexical network approach to analyzing written knowledge provides practical tools to develop effective operationalizations to study the structure of lexical networks and lexicons in learning. We focus here on didactical texts meant for science education, and investigate how different lexicons appear in them. The decision to pay attention on didactical texts derives from the notion that the vocabulary of didactical texts like textbooks may have a crucial role in learning (see e.g. [2] and references therein). As a target of the study, we have chosen two well-known texts about a topic that has recently raised interest in science education: the wave-particle dualism of electrons and photons as quantum entities, which is known to raise much confusion not only in learning the topic but also how different interpretations of dualism infiltrate the process of teaching it [15, 16]. The texts we have chosen to study are: (A) a text that introduces electrons and photons as field quanta, discusses them both quite similarly as field excitations and approaches wave-particle dualism from this viewpoint [17]; (B) another text, which takes a viewpoint emphasizing the role of statistical interpretation of measurements and statistical (or ensemble) understanding of the quantum nature of measurement events [18]. The differences between these views and their consequences for didactics are well recognized [15]. Therefore they are good candidates to test the method of analyzing the lexical structure in didactical texts which have similar similar goal, scope and intended audience, but supposedly differences in their vocabulary.

2 Method

The two didactical texts A and B studied here are analysed stepwise: first, the texts are pruned to create the text-corpus to be analyzed; second; a stratified lexical network is constructed to reveal the connectivity of words and terms in different levels; third and finally, a lexicon profile that condenses the important terms is formed using a network-based analysis.

2.1 Construction of Text Corpus

The parts of texts in A and B to be analyzed have a common theme of wave-particle dualism (WPD). The text excerpts consist of about 1200 words in A and 1000 in B. In both texts, the double-slit experiment for weak intensity light and electron beams play a crucial role, thus setting the broad context in which to discuss WPD. This body of text is divided into 11 narrower, roughly similar, contexts (denoted by KK in what follows), including the preliminary introduction of the topic (KK₁), discussion of experimental apparatus (KK₂), discussions of details of the experiments and observations (KK₃ - KK₅), qualitative interpretations (KK₆ - KK₈), theoretical explanations (KK₉ - KK₁₀) and summary of main conclusion (KK₁₁). Each context KK contains from one to four cotexts (K), in which sentences form units discussing some sub-topic within the context KK. Within the cotexts K, the text is pruned by reducing the sentences to main clauses *v*. The nouns *n* appearing in the clauses are then found and listed. For clearly synonymous nouns, a common noun is used. In addition, the clauses are roughly classified according to their modality: Questions (Q), assertions (A) and conclusions (C).

The lexicons of interest here are related to terms T = [electron] and T = [photon], where brackets indicate that we treat them as lexical expressions. The lexicon is formed by all those terms and words that are linked to T through the lexical network. Instead of a complete lexicon, the goal is to form a condensed representation of it, in the form of a lexicon profile. The lexicon profiles are formed in four steps by:

1. Constructing the stratified lexical network from clauses *v* to contexts KK.
2. Finding the most relevant terms through counting weighted walks.
3. Constructing lexical proximity network on the basis of walks.
4. Extracting the key terms which form the lexicon profile for T.

2.2 Construction of Lexical Networks

The construction of stratified lexical networks is performed so that we can differentiate levels from single clauses to the cotext of several clauses and finally up to context. To accomplish this, the pruned text consisting of main clauses is transformed into a network in which nodes corresponding to relevant terms and words (nouns) *n* are first connected to nodes representing a root verb *v*. The root verb nodes *v* are next connected to nodes V_X that represent the modality

of clause $X \in \{Q, A, C\}$, and are then connected to cotext K. These connections are clarified in scheme in Table 1. Note that V_Q and V_C are connected to V_A as shown in Table 1 since modalities Q and C also require assertion-type clauses A to be meaningful. After making these connections, the cotexts are connected to contexts KK and finally, connections between contexts KK are made to form the complete structure of the text. The part of the network which consists of connections from set of nouns $\{n\}$ up to cotext K and context KK is shown in Table 1 with length of walks L needed to reach the term T from given noun n . Note that T is part of set $\{n\}$ but reversed link $v \rightarrow T$.

Table 1. The construction of lexical networks. The form of links are shown in column “Link”. The length L of walk that connects a given noun n to term T of interest is given in column “Walk”. Note that the term T is not shown, but it would be a link $v \rightarrow T$.

Scheme	Link	Walk
Nouns n $n \rightarrow v$	L=2	
Clauses v $v \leftrightarrow V$	L=4	
Modality V $V \leftrightarrow K$	L=5-6	
Cotext K $K \leftrightarrow KK$	L=7-8	
Context KK $KK \leftrightarrow KK'$	L=9-10	

2.3 Analysis of the Network

The better the global connectivity between terms and words (nouns) in the network based on counting the walks between them, the more relevant we consider the connections between them. This decision to attach relevance to terms is different from the standard closeness criteria used to construct lexical (or semantic) networks [2, 5, 6]. However, a measure which takes into account the global connectivity of nodes in the stratified network has more resolving power with regard to the level of connections. We quantify such connectivity here by using communicability [19–21]. Closeness, however, will be relevant later on when we have constructed the lexical proximity network.

The lexical network can be described by a $N \times N$ adjacency matrix \mathbf{A} with elements $[A]_{pq} = a_{pq}$, where $a_{pq} = 1$ when nodes are connected and $a_{pq} = 0$

when they are not connected. The powers k of adjacency matrix \mathbf{A} can be used to obtain the number of walks of length k connecting two nodes within the network. In a connected network, however, the number of long walks increases rapidly, nearly factorially with the length of the walk. Therefore, the number of walks is usually divided by the factorial, to obtain the communicability [19–21]. For the walk counting, we use the the communicability matrix \mathbf{G} with elements G_{pq} between each pair of nodes p and q . The communicability describes roughly how (e.g. information) content of node p flows to node q [19–21]. Here, we use slightly modified communicability \mathbf{G} where the first $M-1$ walks are removed and define it as

$$\mathbf{G}(\beta; M) = \sum_{k=M}^{\infty} \frac{\beta^k \mathbf{A}^k}{k!} = \text{Exp}[\beta \mathbf{A}] - \left(\mathbf{1} + \sum_{k=1}^{M-1} \frac{\beta^k \mathbf{A}^k}{k!} \right) \quad (1)$$

We use here $M = 5$ which excludes the simplest levels $L \leq 4$ from counting of walks. In practice, the analysis is not sensitive to choice of M provided it remains below $M = 5$, since in these lower levels the walks do not include the connection that are established and level of walks longer than 6, through the contexts and contexts. In the lower levels and with $M \leq 5$ the analysis essentially coincides with the word-frequency analysis. The parameter β is used to tune how extensive apart of the network is included in counting the walks. By varying the parameter β we get information how meaning attached to terms changes when only sentence-level connections are included (β low) in comparison to case when contexts level is taken into account (β high). It should be noted that due to finite size of the networks, after certain high enough value of β the contribution from context-level does not change anymore. An optimum value of parameter β is such that all paths that increase the diversity of key terms and words that contribute to the total communicability are included with the lowest possible value of β . This corresponds roughly to the maximum value of the Frechet-derivative [19] of \mathbf{G} occurring with values $1.5 < \beta < 3.0$ in the cases studied here. In what follows, the communicability is normalized to a maximum value of one. By using the (normalized) communicability $[\mathbf{G}]_{pq} = G_{pq}$ between nodes p and q we can now obtain the total lexical support of node q from all other nodes p , which are taken to be relevant in providing the lexical meaning of it.

The communicability of nodes within the lexical network is next used as the basis to form the lexical proximity network. In the lexical proximity network we retain only those connections between nodes p and q that exceed a certain threshold G^* of communicability. The lexical proximity network thus contains the most important lexical connections.

2.4 Construction of Lexicon Profile

The terms in the lexical proximity network are the basis to construct the lexicon profiles for the terms $T = [\text{electron}]$ or $T = [\text{photon}]$. The key terms that are connected to T in the proximity network are classified in N categories, which condense the information of the lexical connections. Each category of the key

terms describes a given descriptive property $P = 1, 2, \dots, N$ of interest. This classification is made for the practical purpose of condensing the relevant lexical information, because lexical networks with complete interrelationships between terms are too rich to easily yield the relevant information of the lexical structure. The condensed representation of the lexicon in the form of a descriptive property of the words in a given category P and with information on the relative importance of that category is referred to as a lexicon profile in what follows.

The lexicon profile is formed by defining the lexical support $\Pi(P)$ the term T receives from the lexical proximity network of key terms $p(P)$ attached to a given feature P . Such support is operationalized as the sum of closeness centrality $CC_p(P; T)$ of node $p(P)$ in proximity network of term T , in the form

$$\Pi_P(T) = \sum_{p \in P} CC_p(P; T) \quad (2)$$

In Eq. (2) the closeness centrality CC_p of node p is defined in the standard way [19]. The components Π_P form the lexical profile as an N -dimensional vector of lexical supports $\bar{\Pi}(T) = (\Pi_1, \Pi_2, \dots, \Pi_N)/\text{Max}\{\{\Pi_P\}\}$ where the normalization make the lexicon profiles of different texts comparable.

3 Results

The lexical proximity networks of texts A and B are shown in Figs. 1 and 2, respectively. In the lexical proximity networks, we have denoted (with symbols explained in Table 2) the nodes representing terms and words, which are relevant for lexicons of $T = [\text{electron}]$ and $T = [\text{photon}]$. Note that the lexical proximity network are always specific to the term T of interest. The links in the network correspond to communicability between nodes that exceeds the threshold value, which is chosen to be approximately $G^* = 0.1 \beta$, with $\beta \geq 1$ (note that with increasing β the values of communicabilities increase). The closeness centrality of nodes in the proximity network is shown as the size of the node.

The parameter β controls the extent of walks included in the analysis. The highest value $\beta \approx 2.5$ corresponds to the maximum of the Frechet-derivative of the total communicability, and with increasing β no essentially new connections are available. However, up to this value, with increased values of β , the closeness of the nodes change, revealing that more remote connections start to contribute to the communicability. In the case of the stratified network construction (as explained in Table 1) this means that for $\beta = 1$ mainly the lexical at level L from 4 up to 5 is included, while for $\beta = 2.5$ levels $L \geq 7$ also contribute to the communicability of nodes. These levels $L \geq 7$ bring in the semantic, context-related connections.

The lexicons for speaking about electrons and photons are here taken to consist of terms connected to nine ($N = 9$) different types of properties which are relevant in the contexts of the terms electron and photon. Examples of key words related to these properties are listed in Table 2. The nine properties P of interest here are:

1. (W) Field and/or wave-properties and interference
2. (P) Particle properties, particle models, particle-like existence
3. (D) Dualistic nature of entities and wave-particle dualism
4. (e) Epiphenomenal nature of entities (i.e. exists only in localization/measurement)
5. (Q) Quanta and quantization, quantum nature of excitations
6. (m) Classical mechanics related properties (e.g energy and linear momentum)
7. (S) Stochastic and probabilistic properties and indeterminacy
8. (N) Nomic (law-like) properties that refer to theory, laws and principles
9. (X) Experiment-related properties (in double-slit experiments)

The results in Fig. 1 show that the lexical proximity networks of text A attached to $T = [\text{electron}]$ and $T = [\text{photon}]$ have many similarities and their lexical networks overlap because many nodes play similar roles in them. However there are also several terms attached to $[\text{electron}]$ and to $[\text{photon}]$ which are not shared but which exceed the threshold G^* for communicability. Such terms and words are specific to given T and thus important, although their closeness to T in the proximity network is not very high. In general, text A uses rather symmetric vocabulary for $[\text{electron}]$ and $[\text{photon}]$. Some of the properties like Q and X appear to be very dominant in their lexicon profiles.

The results in Fig. 2 for text B show that in B the lexical networks for $[\text{electron}]$ and $[\text{photon}]$ are somewhat more limited than in A. In addition, the differences between lexical proximity networks for $[\text{electron}]$ and $[\text{photon}]$ appear to be larger in B than in A. The vocabulary in B does not show dominance of properties Q and D, instead, property S is strongly present.

On the basis of the lexical proximity network we form the lexicon profiles that the texts A and B attach to $[\text{electron}]$ and $[\text{photon}]$ and reduce them to nine-dimensional vectors, where each dimension is denoted by one of the tags $P \in \{F, P, D, e, Q, m, S, N, X\}$ as they are indicated by the set p of key-words (for some examples see Table 2). The choice of key words is specific to the text, some of them identical, but generally different vocabularies are used. Therefore, the choice of key words contain an element of interpretation about the significance of the word.

The Fig. 3 shows the lexicon profiles corresponding to the lexical proximity networks in Figs. 1 and 2. The differences between the lexicon profiles corresponding to texts A and B are now clearly visible. Lexicon profiles for $[\text{electron}]$ and $[\text{photon}]$ in A are dominated by terms related to quanta and quantization (Q), and also words and terms related to dualism (D) and stochasticity (S) are in center of the vocabulary. The lexicon profiles are also very symmetric, revealing that $[\text{electron}]$ and $[\text{photon}]$ are addressed using very similar vocabulary. These features of the lexicon profiles are in line with the goal of text A, which is to discuss electrons and photons from a unified perspective and with emphasis on field quantization.

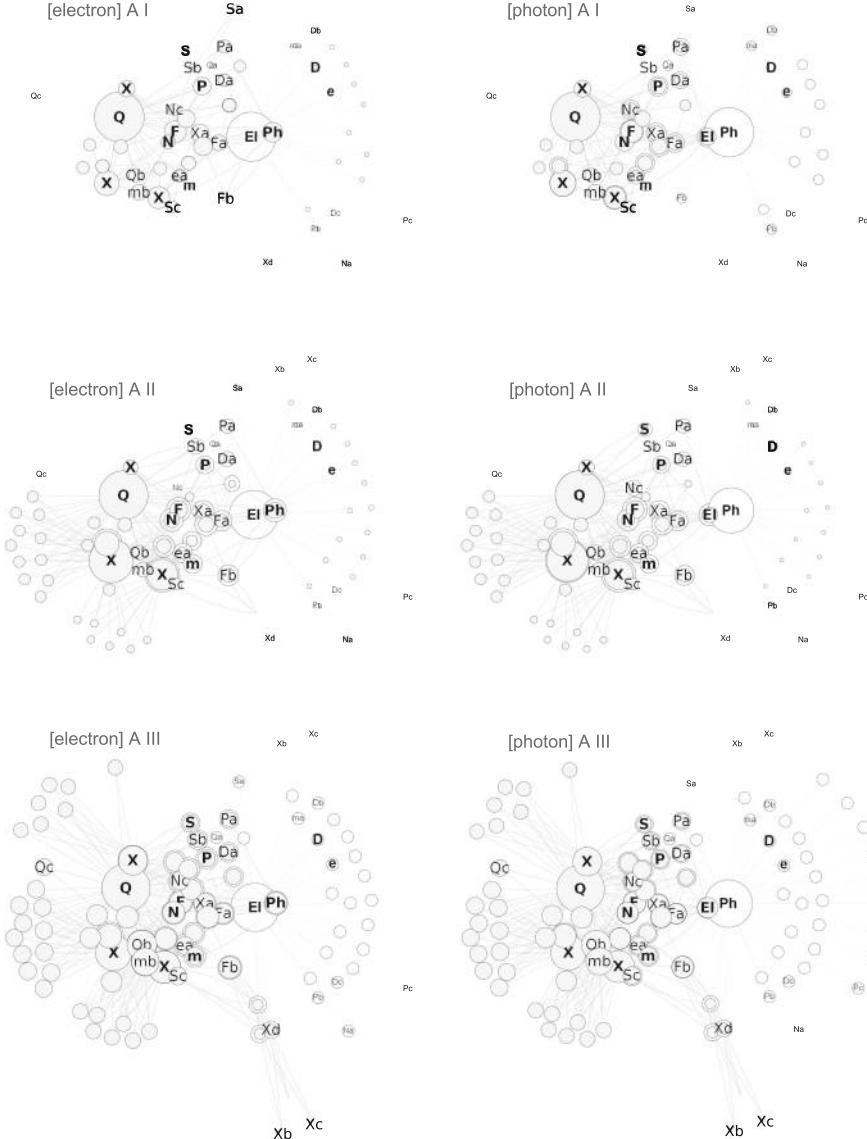


Fig. 1. The lexical proximity network for terms $T = [\text{electron}]$ and $T = [\text{photon}]$ as it appears in text A. The nodes corresponding to $[\text{electron}]$ and $[\text{photon}]$ are denoted by El and Ph, respectively. In the left column, lexical proximity network for $[\text{electron}]$ is shown in light gray and for the $[\text{photon}]$ with white nodes. In the right column, $[\text{electron}]$ is shown with white and $[\text{photon}]$ with light gray nodes. The sizes of the nodes correspond to the closeness centrality of the node. The results are shown for $\beta = 1.0$ (I), 2.0 (II) and 2.5 (III). The threshold value for the communicability between the nodes, which is the basis of forming the proximity network, is $G^* = 0.1 \beta$.

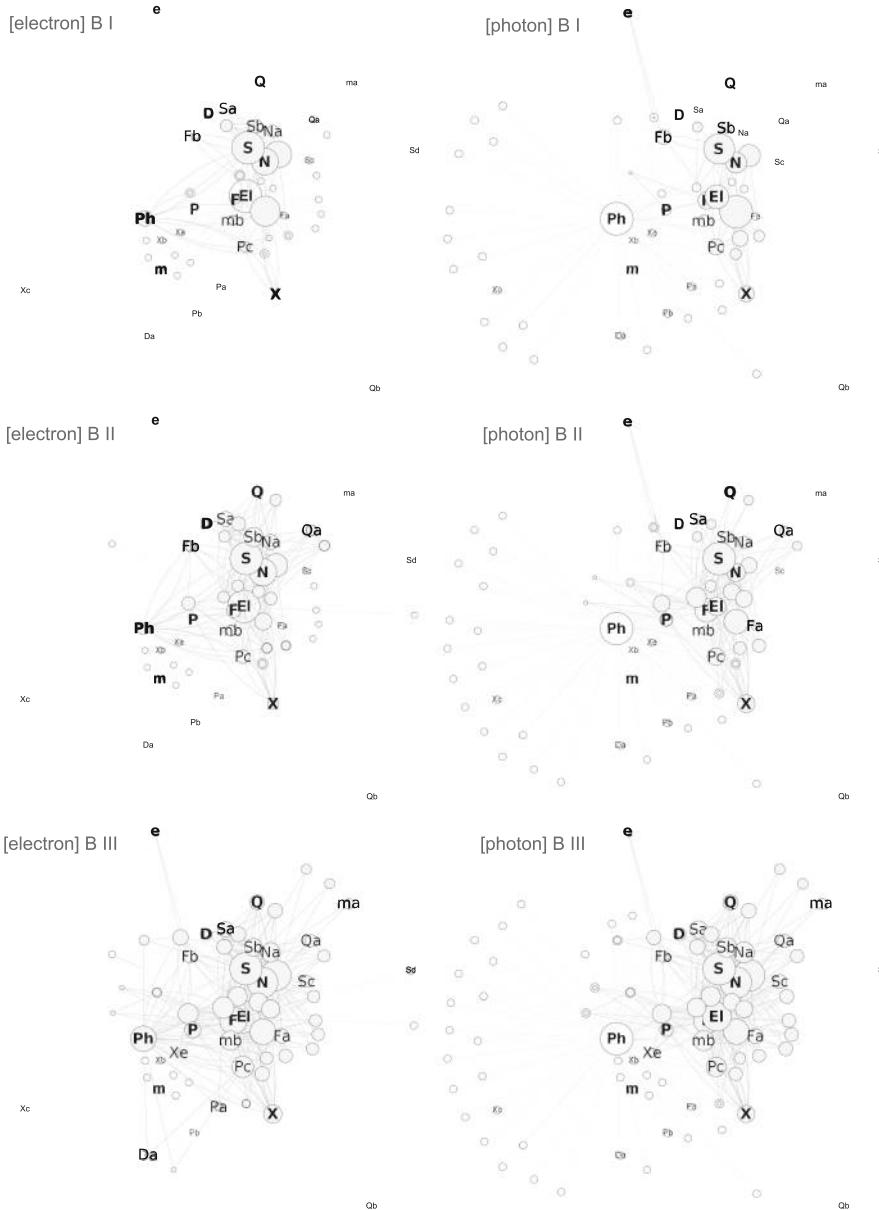


Fig. 2. The lexical proximity network for terms $T = [\text{electron}]$ and $T = [\text{photon}]$ as it appears in text B. The nodes corresponding to [electron] and [photon] are denoted by El and Ph, respectively. In the left column, lexical proximity network for [electron] is shown in light gray and for [photon] with white nodes. In the right column, [electron] is shown with white and [photon] with light gray nodes. The sizes of the nodes correspond to the closeness centrality of the node. The results are shown for $\beta = 1.0$ (I), 2.0 (II) and 2.5 (III). The threshold value for the communicability between the nodes, which is the basis of forming the proximity network, is $G^* = 0.1 \beta$.

Table 2. The $N=9$ properties P defining the lexicon profiles and examples of the key words and terms attached to the properties. The most central key-word/term of the given property in texts A and B are denoted by subscripts, while the second and third most important additional key-words are denoted by subscript ADD. The symbols from F to X correspond to those in Figs. 1, 2 and 3.

Property P	Key words and terms p for P		
	PA	PB	PADD
Field/Wave	F field	wave	interference, interf/extend. pattern
Particle	P particle	particle	particle track, atom, object
Dualism	D field-partc. dual	wave-partc. dual	paradox, partc.-like
Epiphenom	e excitation	emergence	apparent particle, localization
Quantization	Q quantum	quantumobject	quantumstate, quantized field
Mechanics	m energy	monoenergetic	linear momentum, superposition
Stochastics	S probability	(prob. of) impacts	prob. density/distribution/amplit.
Nomic	N wavefunction	wavefunction	quantum mech./phys./theory.
Experiments	X detector/screen	interferometer	two-slit exp, electron/light beam

The lexicon profiles in B, on the other hand, are clearly different from the profiles in A. Apart the fact that in B the stochastic (S) dimension of the lexicon profile is pronounced for both [electron] and [photon] as well, the lexicon profiles in B are quite different from those in A. It is noteworthy that in B vocabulary for speaking about quanta and quantization is rather weak, as well as for dualism. In addition, the role of experiments for electrons and photons is asymmetric in B; dominant for [photon] but weak for [electron]. These notions are in agreement with the general tone of text B, which emphasizes the statistical interpretation of measurement events, a viewpoint which closely resembles the ensemble interpretation of quantum physics. The asymmetry of the role of experiments is clearly a consequence of B discussing interferometric experiments thoroughly for photons but not for electrons; for electrons no experimental results are discussed at all.

The lexicon profiles are shown in Fig. 3 for different values of β , corresponding to inclusion of different stratified levels of network. With increasing values from $\beta = 1$ up to $\beta = 2.5$ the analysis gradually covers levels from L from 4 and 5 up to levels $L \leq 7$, where more remote connections start to contribute. These more remote connections increasingly provide the semantic, context related connections to the lexical terms. The more remote connections are supposedly important in providing the semantic content and also the different contextual ways to understand the meaning of terms; i.e. they reveal the context-relatedness and dependence of lexicons.

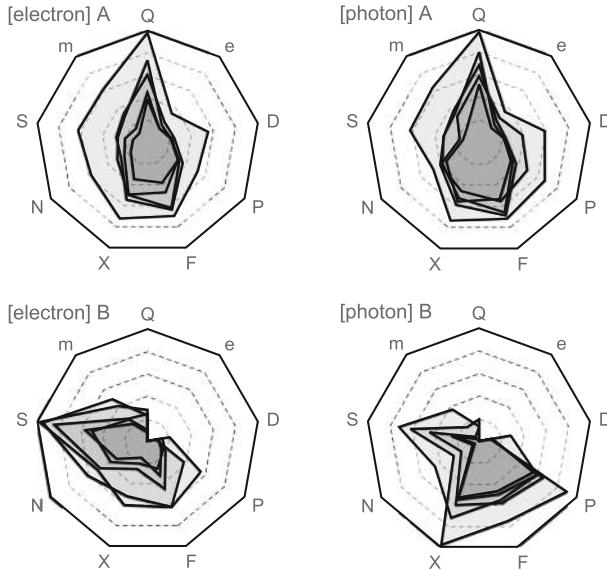


Fig. 3. The lexicon profiles for $T = [\text{electrons}]$ and $T = [\text{photons}]$ in texts A (the upper row) and B (the lower row). The symbols for the properties $P \in \{F, P, D, e, Q, m, S, N, X\}$ and how they are related to key words/terms is as in Table 2. Polygons show results for $\beta = 1.0, 1.5, 2.0$ and 2.5 , inner polygon corresponding to the lowest and the outmost to the highest value of β .

4 Discussion and Conclusions

The analysis of lexical networks and lexicon profiles is here used to reveal how didactical texts about the same topic may nevertheless use very different vocabularies. We used network-based methods to analyze two didactical texts about wave-particle dualism and nature of electrons and photons as quantum entities. Although the topic is the same, the texts analyzed here approach the topic differently. Therefore, they are suitable to test the method of analysis, which attempts to first construct the stratified lexical structure and then find the lexical proximity of terms in that structure.

The analysis presented here resembles more traditional analyses of semantic networks (compare with e.g. [2, 5, 6]). The difference, however, is that here we have performed stratified analysis, which is sensitive to features on the lexical level of cotext up to the level of contexts, where the lexical level meets the semantic level. In the lowest level, where connections are on sentence level, the results are expected to coincide with simple word-frequency counts. The results of the study show that the stratified analysis of lexical structure is able to reveal how the lexicons become augmented when the deeper contextual levels, by inclusion of more remote connections between the terms, are included in the analysis. The advantage of the method is thus the control it provides over the different level of connections.

Regarding teaching and learning, an analysis of the lexical structures is an important starting point to better understand how lexicons affect what kinds of conceptions are conveyed in teaching and instruction, and how lexicons may either facilitate or hinder discussions of certain aspects of the topics to be learned. The didactical texts A and B discussed here clearly differ in how richly they cover vocabulary to discuss different aspects of the wave-particle duality. The text A has richer vocabulary, which is enriched when context level is taken into account so that the stochastic, dualistic and quantization dimensions S, D and Q are strengthened. This is as desired, since these dimensions are related to explanatory aspects, quite naturally a desirable feature of a didactical text. Still the symmetry of vocabulary needed to convey the symmetry in wave-particle dualism is maintained. The text B, on the other hand, has more limited vocabulary, which also changes when context level is taken into account, but now mostly in dimensions particle (P), field/wave (F) and experiment (X). These dimensions, however, are descriptive rather than explanatory. In addition, the vocabularies for electrons and photons are somewhat asymmetric. There is no indication in text B that such an asymmetric use of vocabulary was an intended features, given that the goal is wave-particle dualism. For a teacher who uses the didactical texts, either in teaching or for study, the ability to recognize and master rich enough terminology for many-faceted discussion is a very important component of professional competency. Therefore, it is also important to have tools to analyze didactical texts to become aware of such differences and to be able to detect them. To monitor the richness of lexicons used for didactical purposes, we need methods of research that are sensitive enough to features of different levels, from lexical to semantic levels, and which are controlled and reliable. For such purposes network-based methods provide new tools that complement and augment more traditional approaches.

Funding. This research was funded by the Academy of Finland, Grant 311449.

References

1. Darian, S.G.: *Understanding the Language of Science*. University of Texas Press, Austin (2003)
2. Yun, E., Park, Y.: Extraction of scientific semantic networks from science textbooks and comparison with science teachers' spoken language by text network analysis. *Int. J. Sci. Educ.* **40**, 2118–2136 (2018)
3. Brookes, D.T., Etkina, E.: The importance of language in students' reasoning about heat in thermodynamic processes. *Int. J. Sci. Educ.* **37**, 759–779 (2015)
4. Rincke, K.: It's rather like learning a language: development of talk and conceptual understanding in mechanics lessons. *Int. J. Sci. Educ.* **33**, 229–258 (2011)
5. Clariana, R.B., Wolfe, M.M., Kim, K.: The influence of narrative and expository lesson text structures on knowledge structures: alternate measures of knowledge structure. *Educ. Tech. Res. Dev.* **62**, 601–616 (2014)
6. Clariana, R.B., Wallace, P.E., Godshalk, V.M.: Deriving and measuring group knowledge structure from essays: the effects of anaphoric reference. *Educ. Tech. Res. Dev.* **57**, 725–737 (2009)

7. Derman, A., Eilks, I.: Using a word association test for the assessment of high school students' cognitive structures on dissolution. *Chem. Educ. Res. Pract.* **17**, 902–913 (2016)
8. Neiles, K.Y., Todd, I., Bunce, D.M.: Establishing the validity of using network analysis software for measuring students' mental storage of chemistry concepts. *J. Chem. Educ.* **93**, 821–831 (2016)
9. Kuhn, T.S.: *The Road Since Structure*. University of Chicago Press, Chicago (2000)
10. Hoyningen-Huene, P.: *Reconstructing Scientific Revolutions*. University of Chicago Press, Chicago (1993)
11. Evans, V.: *How Words Mean: Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford University Press, Oxford (2009)
12. Langacker, R.W.: *Grammar and Conceptualization*. Mouton de Gruyter, Berlin (1999)
13. Koponen, M., Asikainen, M.A., Viholainen, A., Hirvonen, P.E.: Using network analysis methods to investigate how future teachers conceptualize the links between the domains of teacher knowledge. *Teach. Teach. Educ.* **79**, 137–152 (2019)
14. Kubisch, M., Nordine, J., Neumann, K., Fortus, D., Krajcik, J.: Probing the relation between students' integrated knowledge and knowledge-in-use about energy using network analysis. *Eur. J. Math. Sci. Tech. Educ.* **15**, 1728 (2019)
15. Cheong, Y.W., Song, J.: Different levels of the meaning of wave-particle dualism and a suspensive perspective on the interpretation of quantum theory. *Sci. Educ.* **23**, 1011–1030 (2014)
16. Ayene, M., Krick, J., Damitie, B., Ingerman, A., Thacker, B.: A holistic picture of physics student conceptions of energy quantization, the photon concept, and light quanta interference. *Int. J. Sci. Math. Educ.* **17**, 1049–1070 (2019)
17. Hobson, A.: There are no particles, there are only fields. *Am. J. Phys.* **81**, 211–223 (2013)
18. Müller, R., Wiesner, H.: Teaching quantum mechanics on an introductory level. *Am. J. Phys.* **70**, 200–209 (2002)
19. Estrada, E.: *The Structure of Complex Networks*. Oxford University Press, Oxford (2012)
20. Koponen, I.T., Nousiainen, M.: Concept networks in learning: finding key concepts in learners' representations of the interlinked structure of scientific knowledge. *J. Complex Netw.* **2**, 187–202 (2014)
21. Koponen, I.T., Nousiainen, M.: Concept networks of students' knowledge of relationships between physics concepts: finding key concepts and their epistemic support. *Appl. Netw. Sci.* **3**, 14 (2018)



Legal Information as a Complex Network: Improving Topic Modeling Through Homophily

Kazuki Ashihara^(✉), Chenhui Chu, Benjamin Renoust, Noriko Okubo,
Noriko Takemura, Yuta Nakashima, and Hajime Nagahara

Graduate of Information Science and Technology, Institute for Datability Science,
Graduate School of Law and Politics, Osaka University, Osaka, Japan
ashihara.kazuki@ist.osaka-u.ac.jp, greenaccess@law.osaka-u.ac.jp,
{chu,renoust,takemura,n-yuta,nagahara}@ids.osaka-u.ac.jp

Abstract. Topic modeling is a key component to computational legal science. Network analysis is also very important to further understand the structure of references in legal documents. In this paper, we improve topic modeling for legal case documents by using homophily networks derived from two families of references: prior cases and statute laws. We perform a detailed analysis on a rich legal case dataset in order to create these networks. The use of the reference-induced homophily topic modeling improves on prior methods.

Keywords: Network of legal documents · Topic modeling · Homophily

1 Introduction

Computational legal science is a growing field that is the intersection of many disciplines, applying quantitative computational to laws, bringing a pluridisciplinary effort of, among others, sociophysics, network science, natural language processing, machine learning and statistical methods [12]. The effort on studying law from a computational social science perspective has steadily increased in the past decade [19]. The search for patterns in citation analysis of legal documents has been proven to be an efficient strategy over no matter the country, always concluding in a complex network analysis [8, 14, 15, 17, 18, 20, 26].

Citation analysis has been popularly used for academic evaluation [10] although it has shown controversies even recently [28]. The structure of legal cases may not be as deep as of citation networks (deep in the sense of depth in directed acyclic graphs). It rather forms a flat organization with an emphasized importance attached precedent cases [26], which makes the topology of a legal citation network slightly different to an academic citation network.

Network analysis has also been relying on the investigation of the co-citation patterns [14, 15]. This is sometimes referred to as homophily [6, 22], which is the implied similarity of two entities, and the property of entities to agglomerate when being similar. Homophily often corresponds to a bipartite structure which may be projected into single type networks. Topic modeling has been a strategy

to explore homophily in these projections [30]. Topic modeling is the key to numerous different process of information retrieval and clustering as part of legal information analysis [13]. It gives rise to the COLIEE challenge [32].

In this paper, we contribute by constituting networks from a rich collection of cases, based on references to prior cases and statute laws. We use these references to explore homophily relationships between cases, and improve topic modeling for legal information using case homophily. We apply our analysis on a public dataset of the Canadian Federal Court (as part of the COLIEE challenge [32]) and compare different weighing strategies.

After presenting the related work in Sect. 2, we first introduce the COLIEE dataset [32] and its characteristics in Sect. 3, and then propose a description of this dataset as a network in Sect. 4. We describe our improvements of topic modeling for legal cases in Sect. 5 before concluding.

2 Related Work

There are related work in both fields of legislation networks and topic modeling.

Legislation Networks. The interest of network analysis in the context of legal documents is steadily growing since the last decade. Many studies have shown the relevance of analyzing legal networks as complex networks [8, 14, 15, 17, 18, 20, 26]. Fowler *et al.* [8] have developed an interesting centrality measure based on *Hubs* and *Authorities* [16] dedicated to case citations in the US Supreme Court from a citation network of 26k+ cases. Kim [15] has later explored the dynamic structure of 747 treaties and 1k citations to find homophily behavior and complex network properties in the network formed by the treaties. Pelc [26] investigates the fundamental concept of precedent, *i.e.*, previous deliberations cited in a case, in the international commercial cases also through a centrality study of *Hubs* and *Authorities*, confirming the relevance of the network structure in predicting output of cases. Koniaris [17] builds a network of reference to laws from the Official Journal of the European Union, showing it has properties of a multilayer complex network. Khanam *et al.* [14] also propose a citation analysis of judgements of Indian courts using betweenness centrality. We should also underline the relevance of the EUCaseNet project [20], which combines centrality-based visualization and network analysis for the exploration of the whole corpus of EU case laws. Most recently, Lee *et al.* [18] have also explored patterns of constitution articles *vs.* court decision in Korea, including topical analysis of the main clusters. Our application context is close to these as we are investigating the network formed by the Federal Court of Canada case laws. We are very interested in investigating the potential of homophily for our own analysis, as illustrated by all of the work above. In contrast, we propose topic modeling that is tapping into the co-citation structure of cases, and feeding back on homophily of documents in terms of topical proximity.

Topic Modeling. Topic modeling is first introduced by Blei *et al.* [4] as the latent Dirichlet allocation (LDA). LDA is a graphical model that can infer hidden document-topic and word-topic distributions from observed documents.

Section 5 gives a detailed description of LDA. Blei and Lafferty [1] propose the correlated topic model, which uses the logistic normal for LDA to model topic occurrences. Blei *et al.* propose the dynamic topic model to model temporal information in sequence data [2]. Supervised topic modeling also has been studied. Supervised LDA [3] models topics of documents and responses, which is fit for data such as product reviews that has both descriptions and corresponding evaluation scores of products. Ideal point topic models [25] further assume that the responses are hidden. Relational topic model (RTM) [7] models the topic of a document pair that has some links such as citations of papers between them. A detailed description of RTM is given in Sect. 5. Collaborative topic models [31] use user data to make user preferences for recommendation. In this paper, we improve the RTM through the use of homophily in co-citation networks and apply it for legal case analysis.

3 Data

Our data is extracted from COLIEE 2018 [32],¹ which is the Competition on Legal Information Extraction/Entailment. We use the *Case Law Competition Data Corpus* for our task, which has been used in *Task 1* and *Task 2*. The data contains 6,154 cases of the Federal Court of Canada during roughly 40 years between 1974 and 2016, while the most of the cases are reported since 1986. This data is extremely rich. Each case is semi-structured under a text form. Each text contains (but non exhaustively) a summary of the content of the case, court, ruler, legal topics of interest, cases and status notices, counsels, solicitors, important facts, and miscellaneous information.

As we want to study the reference network, we will focus on two types of citations, which are utilized in making the decision. They are separated by the following paragraph titles from the text input:

- **Cases Noticed**, they correspond to the past concluded trials in connection to the trial.
- **Statutes Noticed**, they correspond the laws being referred to in order to give the verdict of this trial.

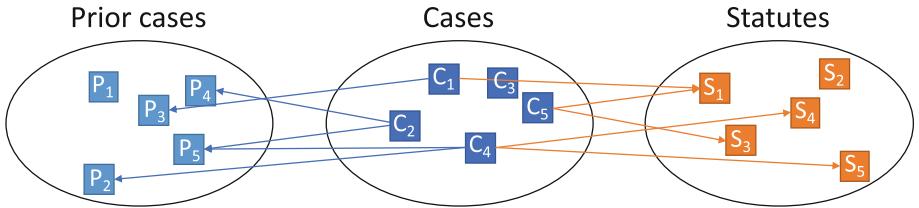
There is then one reference per consecutive line. Note that because it is Canada, these may contain not only English, but also French. Among all the cases, only 5,576 cases make mention to prior cases and statutes noticed.

The coarsening of reference destination is usually very detailed and references may be divided into chapters or paragraphs. If we use these as units of analysis in our network modeling, the resulting network would be very sparse and only a small amount of references would be redundant across the cases. Instead, we consider the references to the full case or statute articles. Each case or statute is identified by a title, year, and some references. We base our parsing on finding this year structure, giving titles at the higher level of granularity which becomes

¹ <https://sites.ualberta.ca/~miyoung2/COLIEE2018/>.

Table 1. Identifying references within cases.

Prior case	Singh v. Minister of Employment and Immigration, [1985] 1 S.C.R. 177; 58 N.R. 1; 17 D.L.R.(4th) 422; 14 C.R.R. 13; 12 Admin. L.R. 137, refd to. [para. 8]
→	Singh v. Minister of Employment and Immigration (1985)
Statute	Canadian Bill of Rights, R.S.C.1970, App.I, sect.2(e) [para. 5, footnote 2]
→	Canadian Bill of Rights, R.S.C. (1970)

**Fig. 1.** Structure of the COLIEE dataset.

our nodes (see Table 1 for an example). We discard the 39 citing cases for which we could not retrieve years. We additionally keep the year information attached to the nodes.

4 Network Modeling

4.1 Network Structure

We may now create a network in which each court case refers to its multiple prior cases and statute laws noticed. In our network model $G = (V, E)$, the document of each court case is taken as a node $v_1 \in V$, a noticed case or statute is also represented as a node v_2 , and we regard each reference relationship as a link $(v_1, v_2) = e \in E$. Figure 1 roughly shows our modeling.

Each court case refers to multiple prior cases and statute laws noticed. In total, we consider our initial set of $|C| = 5,539$ studied cases. Each case $c \in C$ may have reference to some prior cases $p \in P$, in total $|P| = 25,112$ prior cases. They may also refer to a statute $s \in S$, with a total $|S| = 1,288$ statutes. The references to cases may be as early as the 18th century (we cannot guarantee the reliability of the year information to cases prior to the 18th century, which concerns only a handful 78 cases).

The resulting network is very sparse and contains 31,976 nodes for 53,554 links. Note that we have a reliable year information only for 29,319 nodes in total (and 1,224 status nodes do not have any year information). We may divide this network into two sub-networks, a first one G_P (29,952 nodes with 53,554 edges) with only the cases and the prior cases they are citing, and a second one

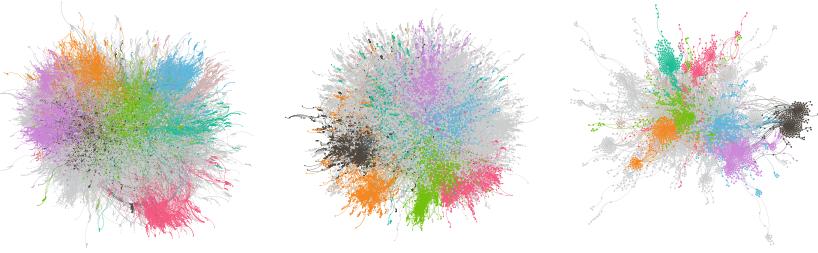


Fig. 2. Main connected components with community structure (left: G , middle: G_P , right: G_S).

G_S (4,441 nodes with 6,453 edges) only considering the initial cases the statutes they are citing.

All these networks present one main connected component covering most nodes and edges, in G the main connected component of 30,456 nodes for 52,453 links, against 27,353/44,871 for G_P , and 4,125/6,150 for G_S . We now investigate the potential for finding communities of cases in these networks using modularity [24] and the Louvain clustering [5]. While the main component of G shows a modularity $Q_G = 0.739$ with 34 communities, the main component of G_P shows a modularity $Q_{G_P} = 0.762$ for 45 communities, and the main component of G_S shows a modularity $Q_{G_S} = 0.747$ for 27 communities. These results encourage us in searching for community structure through citations (see Fig. 2).

4.2 Homophily Network

Cases and statutes citations imply a double bipartite structure in our network: from G_P *case–prior case* relationships, and from G_S , *case–statute* relationships. Bipartite projections into one-mode networks imply a complex network structure [9]. We further investigate homophily, by deriving three one-mode networks: $G'_P = (C, E'_P)$, $G'_S = (C, E'_S)$, and $G' = (C, E')$.

To do so, we project onto *case–case* relationships all other bipartite relationships. In other words, let $u \in C$, $v \in C$ be two cases of the initial set of cases we are investigating. Each of these cases may be assigned a set of references $R_u = \{r_1, r_2, r_3, \dots\}$ where r_x may be either a prior case or a statute law. In a projected network G' , the original cases $\{u, v\} \subseteq C$ become the nodes, and there exists a link $(u, v) = e \in E'$ if and only if the intersection of their respective reference sets is non empty: $R_u \cap R_v \neq \emptyset$. In the network induced by prior cases G'_P , each reference $r_x \in P$ is a prior case. In the network induced by statutes G'_S , each reference $r_x \in S$ is a statute law. In the general projected network G' , a reference $r_x \in S \cup P$ can be of any type.

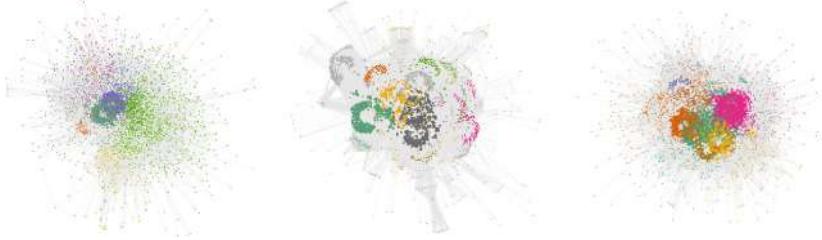


Fig. 3. Projected networks and their communities (left: G'_P , middle: G'_S , right: G').

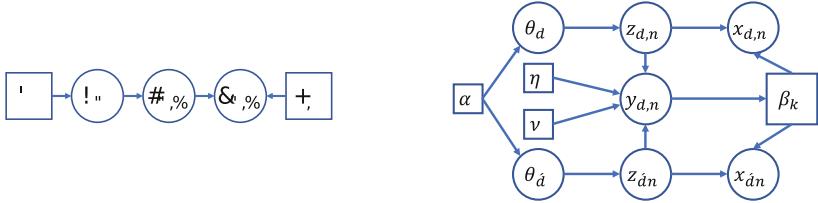


Fig. 4. Statistical model of LDA (left) and RTM (right).

The weight of each resulting link can be obtained by the following two methods, either w_1 the number of shared citations between two cases, or w_2 the Jaccard index of these [11].

$$w_1 = |R_u \cap R_v| \quad \text{and} \quad w_2 = \frac{w_1}{|R_u \cup R_v|} \quad (1)$$

Unsurprisingly, the resulting networks are very dense, we count for G'_P 4,803 nodes and 286,435 edges, for G'_S 3,138/379,447, and for G' a size of 5,576/643,729. We may note that prior cases and statutes induced links overlap only a little. We may further investigate the main components of these networks, with a size of 4,244/286,403 for G'_P with modularity $Q_{G'_P} = 0.428$ for 14 communities, 3,033/379,426 for G' and G'_S with modularity $Q_{G'_S} = 0.542$ for 13 communities, and 4,870/643,725 with $Q_{G'} = 0.502$ and 7 communities for G' (see Fig. 3).

5 Complex Network for Relational Topic Model

We first introduce the conventional topic model of LDA [4] (as illustrated in the left part of Fig. 4), then present the RTM [7] (as illustrated in the right part of Fig. 4), to which we apply the weights that are derived from the homophily relationships for legal document analysis.

LDA [4] is a generative probabilistic model that uses a set of “topics,” distributions over a fixed vocabulary, to describe a corpus of documents. The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. It is assumed that the words in documents are generated by the probability of the topic. From the parameter α , θ_d , the appearance probability of a topic in document d is generated. Next, the topic probability of each word $z_{d,n}$ in document d is generated. Finally, the word $x_{d,n}$ of n -th word in document d is generated from the occurrence probabilities of the vocabulary in the topic k of β_k and $z_{d,n}$. The LDA model focuses only on the content of documents, not the relationships between documents.

In the RTM model [7], the link relationship between documents is also considered. Documents are first generated from topic distributions in the same way as in LDA. The links between documents are then modeled as binary variables, one for each pair of documents. For each pair of documents d , \acute{d} , a binary link indicator is drawn by the following equation:

$$y|z_d, z_{\acute{d}} \sim \psi(\cdot|z_d, z_{\acute{d}}), \quad (2)$$

where ψ is the distributed link probability function between two documents. Two methods have been proposed for ψ , sigmoid or exponential. We adopt the exponential, which shows higher accuracy in the experiments [7]. This function is dependent on the topic assignments that generate their words, z_d and $z_{\acute{d}}$, which is denoted as the following equation:

$$\psi = \exp(\eta^T (\bar{z}_d \circ \bar{z}_{\acute{d}}) + \nu), \quad (3)$$

which is parameterized by coefficients η and intercept ν .

To improve on RTM, we rely on the structure of the homophily networks described in Sect. 4.2. To reinforce the influence of co-citation patterns, we use the links with either the weight w_1 or w_2 as document links in the RTM model. We investigate different thresholds in our experiments.

6 Experiments

We conducted topic modeling experiments on the Canadian law corpus as described in Sect. 3.

6.1 Settings

We performed the following preprocessing for topic modeling. We lemmatized each word using NLTK [21] and lower-cased all words. In addition, we removed the words that are not composed of alphabets. We also excluded stop words by using NLTK’s list of English stop words.²

² <https://gist.github.com/sebleier/554280>.

We used each of the homophily networks G'_P , G'_S , and G' . From the edge weights obtained from Eq. 1, we decided which edge to use or not to use based on a threshold. We used this RTM implementation³ for our experiments. We trained the parameters of RTM⁴ using the obtained network information and documents. We set the number of topics to 200 and max iterator to 10.

We compared our model against the LDA model [4],⁵ which does not consider information between links. We also compared against RTM ($w \rightarrow \infty$), which is an RTM without using any link information.

6.2 Evaluation Metric

Output topics were evaluated using coherence [23]. Coherence is used to evaluate the performance of topic modeling, by measuring the similarity of the output words. Coherence is obtained by the following equation:

$$\text{coherence} = \sum_{i=1} \sum_{j=i+1} \text{sim}(w_i, w_j), \quad (4)$$

where w_i and w_j are i -th and j -th topic word that a topic model outputs, and $\text{sim}(\cdot, \cdot)$ calculates the cosine similarity between two words obtained as word representations by GloVe840B [27].⁶

7 Results

Table 2. Coherence of topics from G'_P and G'_S against w_1 and w_2 .

	G'_P, w_1		G'_P, w_2		G'_S, w_1		G'_S, w_2	
LDA		0.131		0.131		0.131		0.131
RTM	$w_1 \rightarrow \infty$	0.159	$w_2 \rightarrow \infty$	0.159	$w_1 \rightarrow \infty$	0.159	$w_2 \rightarrow \infty$	0.159
RTM	$w_1 \geq 100$	0.166	$w_2 \geq 0.75$	0.160	$w_1 \geq 10$	0.167	$w_2 \geq 0.75$	0.164
RTM	$w_1 \geq 50$	0.165	$w_2 \geq 0.50$	0.166	$w_1 \geq 5$	0.159	$w_2 \geq 0.50$	0.164
RTM	$w_1 \geq 5$	0.167	$w_2 \geq 0.25$	0.161	$w_1 \geq 0$	0.164	$w_2 \geq 0.25$	0.165
RTM	$w_1 \geq 0$	0.166	$w_2 \geq 0$	0.166			$w_2 \geq 0$	0.164

Table 2 shows the coherence scores using case G'_P and statute G'_S networks, with both weighting methods. In the table, w_x with $x \in \{1, 2\}$ denotes the weight threshold between two nodes from Eq. 1. An edge is considered only when

³ https://github.com/dongwookim-ml/python-topic-model/blob/master/notebook/RelationalTopicModel_example.ipynb.

⁴ <https://github.com/dongwookim-ml/python-topic-model>.

⁵ <https://radimrehurek.com/gensim/models/ldamodel.html>.

⁶ <https://nlp.stanford.edu/projects/glove/>. We focused on the top ten words output by a topic model and evaluated its performance.

Table 3. Coherence of topics on G'

Model	coherence
LDA	0.131
RTM ($w \rightarrow \infty$)	0.159
RTM ($w_1^P \geq 0, w_1^S \geq 0$)	0.163
RTM ($w_1^P \geq 5, w_1^S \geq 10$)	0.163
RTM ($w_2^P \geq 0.50, w_2^S \geq 0.25$)	0.167

its weight $w_e \geq w_x$, then within the RTM model, an edge is created between two nodes and trained by RTM. $w_x \geq 0$ means that all given edges regardless of their weight are considered, $w_x \rightarrow \infty$ means no link is input in the model (corresponding to the RTM baseline).

In all cases, we can see that our RTM model given link information shows higher performance than the baselines. In addition, giving a weight threshold also improves in comparison to the inclusion of all links $w_x \geq 0$. This indicates that some links may contain some noise information. When comparing the case of creating a node only from the information of prior cases G'_P and statute laws G'_S , no noticeable difference is obtained, indicating that both are good sources to improve topic modeling. In addition, there is no significant difference when comparing between w_1 , which is the number of common references between two cases, with w_2 measures homophily as a similarity. When only prior cases references are used, in G'_P , $w_1 \geq 5$ shows the highest performance, and when only statutes references are used, in G'_S , $w_1 \geq 10$ shows the highest performance. Also, $w_2 \geq 0.50$ and $w_2 \geq 0.25$ shows highest scores for G'_P and G'_S respectively.

We also tested the right balance between the influence of G'_P and G'_S in the combined links of G' . To avoid the dominance of the prior cases over statutes or *vice versa*, we balance the two weights w_x^P and w_x^S on links induced by cases or statutes by using their best thresholds in Table 2. Results are shown in Table 3. Although its performance is higher than the no link model $w \rightarrow \infty$, the performance is almost the same as the separated ones. It indicates that even if the best parameter is used as each citation information, the performance is not necessarily improved by using both link types.

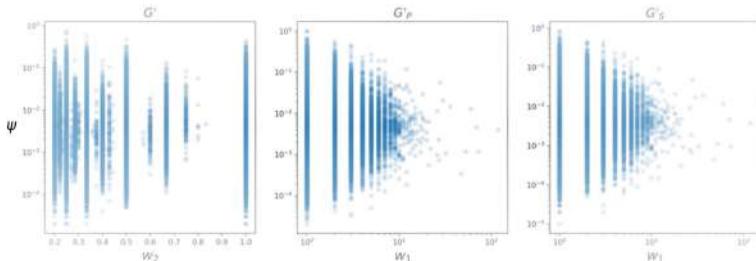
In addition, we investigated the impact of topic numbers for the topic models. Table 4 shows the coherence for different number of topics $|T|$ with their best weight setting for G'_P , G'_S and G' , respectively. It shows the best at $|T| = 10$ when using G'_S and G' , $|T| = 50$ when using G'_P . Although it depends on the case, it shows that coherence tends to increase as the number of topic decreases.

We further explored the difference between homophily (w_x) and topic similarity (ψ) as an output of our trained models. Figure 5 shows the difference we have in the resulting weights. Although the shapes of the topic similarity look very similar with regards to their best weight used w_1 , if we analyze the most similar cases, we obtain different results. With the cases-based topical similarity, with $\psi = 0.65$, $w_1 \geq 1$, and $w_2 \geq 0.05$ for G'_P the closest cases are *Bargig v. Can. (M.C.I.) (2015) case #4127* and *Barak v. Can. (M.C.I.) (2008) #4984*.

Table 4. Coherence over different number of topics for G'_P , G'_S and G' .

$G'_P, w_1 \geq 5$		$G'_S, w_1 \geq 10$		$G', (w_2^P \geq 0.50, w_2^S \geq 0.25)$	
LDA	0.131	LDA	0.131	LDA	0.131
RTM ($w \rightarrow \infty$)	0.159	RTM ($w \rightarrow \infty$)	0.159	RTM ($w \rightarrow \infty$)	0.159
RTM $ T = 200$	0.167	RTM $ T = 200$	0.167	RTM $ T = 200$	0.167
RTM $ T = 100$	0.166	RTM $ T = 100$	0.174	RTM $ T = 100$	0.173
RTM $ T = 50$	0.171	RTM $ T = 50$	0.179	RTM $ T = 50$	0.167
RTM $ T = 10$	0.159	RTM $ T = 10$	0.182	RTM $ T = 10$	0.180

These cases talk about *immigration*. More specifically, both cases are investigating exception requests under the *Immigration and Refugee Protection Act*, and both were rejected under *insufficient humanitarian and compassionate grounds*. With the cases-based topical similarity, with $\psi = 0.48$, $w_1 \geq 1$, and $w_2 \geq 0.5$ for G'_S the closest cases are *Can-Am Realty Ltd. v. MNR (1993) case #4580* and *Deconinck v. MNR (1988) case #475*. These cases talk about *tax*. In both cases, a taxpayer is contesting a *tax assessment*, but one case accepted the plaintiff's appeal (#475) while rejected the other (#4580). With the statutes and case-based topical similarity, with $\psi = 0.019$ for G' the closest cases are *Diabate v. Can. (M.C.I.) (2013) case #3451* and *De Araujo v. Can. (M.C.I.) (2007) case #1276*. These cases talk about *immigration*. In both cases, the applicants asked a judicial review for a *humanitarian and compassionate relief*. One application was accepted (#3451) and the other one rejected (#1276).

**Fig. 5.** Comparison of the best w_1 and w_2 against the topic similarity ψ .

8 Conclusion

We proposed a new approach to a dataset of the legal article and extracted networks of thousands of cases and references, across two main classes of citations: prior cases, and statutes laws. We used these two types of references to investigate citation homophily among cases. These allowed us to improve upon topic modeling thanks to these references. Data is available online.⁷

⁷ https://figshare.com/articles/Legal_Information_as_a_Complex_Network_Improving_Topic_Modeling_through_Homophily/9724070.

In the future, we will further investigate the similarity and content of topics by visualizing the overlapping of topics in a multilayer network model with Detangler [29]. We also wish to combine all the heterogeneous matter contained in the data, such as counselors and so on. Using our topic modeling, we wish to investigate further and find new links in the dataset, based on topic similarity, so that we may explore homophily between the cited cases and laws.

References

1. Blei, D., Lafferty, J.: A correlated topic model of science. *Ann. Appl. Stat.* **1**, 17–35 (2007)
2. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning, ICML 2006, pp. 113–120. ACM, New York (2006)
3. Blei, D.M., McAuliffe, J.D.: Supervised topic models. In: Proceedings of the 20th International Conference on Neural Information Processing Systems, NIPS 2007, pp. 121–128. Curran Associates Inc., USA (2007). <http://dl.acm.org/citation.cfm?id=2981562.2981578>
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
5. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**(10), P10008 (2008)
6. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. *Science* **323**(5916), 892–895 (2009)
7. Chang, J., Blei, D.M.: Relational Topic Models for Document Networks. In: International Conference on Artificial Intelligence and Statistics, pp. 81–88 (2009)
8. Fowler, J.H., Johnson, T.R., Spriggs, J.F., Jeon, S., Wahlbeck, P.J.: Network analysis and the law: measuring the legal importance of precedents at the us supreme court. *Polit. Anal.* **15**(3), 324–346 (2007)
9. Guillaume, J.L., Latapy, M.: Bipartite graphs as models of complex networks. *Physica A* **371**(2), 795–813 (2006)
10. Hirsch, J.E.: An index to quantify an individual's scientific research output. *Proc. Nat. acad. Sci.* **102**(46), 16569–16572 (2005)
11. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull. Soc. Vaudoise Sci. Nat.* **37**, 547–579 (1901)
12. Katz, D.M.: What is computation legal studies? (2011)
13. Katz, D.M., Bommarito, M.J., Seaman, J., Candeub, A., Agichtein, E.: Legal n-grams? a simple approach to track the ‘evolution’ of legal language. In: Proceedings of JURIX (2011)
14. Khanam, N., Wagh, R.S.: Application of network analysis for finding relatedness among legal documents by using case citation data. *i-Manager's J. Inf. Technol.* **6**(4), 23 (2017)
15. Kim, R.E.: The emergent network structure of the multilateral environmental agreement system. *Global Environ. Change* **23**(5), 980–991 (2013)
16. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *J. ACM (JACM)* **46**(5), 604–632 (1999)
17. Koniaris, M., Anagnostopoulos, I., Vassiliou, Y.: Network analysis in the legal domain: a complex model for european union legal sources. *J. Complex Networks* **6**(2), 243–268 (2017)

18. Lee, B., Lee, K.M., Yang, J.S.: Network structure reveals patterns of legal complexity in human society: the case of the constitutional legal network. *PloS one* **14**(1), e0209844 (2019)
19. Lettieri, N., Faro, S.: Computational social science and its potential impact upon law. *Eur. J. Law Technol.* **3**(3) (2012)
20. Lettieri, N., Faro, S., Malandrino, D., Faggiano, A., Vestoso, M.: Network, visualization, analytics. a tool allowing legal scholars to experimentally investigate EU case law. In: Pagallo, U., Palmirani, M., Casanovas, P., Sartor, G., Villata, S. (eds.) *AI Approaches to the Complexity of Legal Systems*, pp. 543–555. Springer, Cham (2018)
21. Loper, E., Bird, S.: NLTK: the natural language toolkit. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 69–72 (2006)
22. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* **27**(1), 415–444 (2001)
23. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108 (2010)
24. Newman, M.E.: Modularity and community structure in networks. *Proc. Nat. Acad. Sci.* **103**(23), 8577–8582 (2006)
25. Nguyen, V.A., Boyd-Graber, J., Resnik, P., Miler, K.: Tea party in the house: a hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1438–1448. Association for Computational Linguistics, Beijing, July 2015
26. Pelc, K.J.: The politics of precedent in international law: a social network application. *Am. Polit. Sci. Rev.* **108**(3), 547–564 (2014)
27. Pennington, J., Socher, R., Manning, C.: GloVe: global Vectors for Word Representation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543 (2014)
28. Renoust, B., Claver, V., Baffier, J.F.: Multiplex flows in citation networks. *Appl. netw. sci.* **2**(1), 23 (2017)
29. Renoust, B., Melançon, G., Munzner, T.: Detangler: visual analytics for multiplex networks. In: *Computer Graphics Forum*, vol. 34, pp. 321–330. Wiley Online Library, Hoboken (2015)
30. Renoust, B., Melançon, G., Viaud, M.L.: Entanglement in multiplex networks: understanding group cohesion in homophily networks. In: Missaoui, R., Sarr, I. (eds.) *Social Network Analysis-Community Detection and Evolution*, pp. 89–117. Springer, Cham (2014)
31. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011*, pp. 448–456. ACM, New York (2011)
32. Yoshioka, M., Kano, Y., Kiyota, N., Satoh, K.: Overview of japanese statute law retrieval and entailment task at coliee-2018. In: *Twelfth International Workshop on Juris-informatics (JURISIN 2018)* (2018)



Graph-Based Fraud Detection with the Free Energy Distance

Sylvain Courtain^{1(✉)}, Bertrand Lebichot^{1,3}, Ilkka Kivimäki⁴,
and Marco Saerens^{1,2}

¹ LOURIM, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium
SylvainCourtain@ucloco.vin.be

² ICTTEAM, Université catholique de Louvain, Ottignies-Louvain-la-Neuve, Belgium
³ MLG, Université Libre de Bruxelles, Brussels, Belgium

⁴ Department of Computer Science, Aalto University, Espoo, Finland

Abstract. This paper investigates a real-world application of the *free energy distance* between nodes of a graph [14, 20] by proposing an improved extension of the existing *Fraud Detection System* named APATE [36]. It relies on a new way of computing the free energy distance based on paths of increasing length, and scaling on large, sparse, graphs. This new approach is assessed on a real-world large-scale e-commerce payment transactions dataset obtained from a major Belgian credit card issuer. Our results show that the free-energy based approach reduces the computation time by one half while maintaining state-of-the art performance in term of Precision@100 on fraudulent card prediction.

Keywords: Credit card fraud detection · Network science · Network data analysis · Free energy distance · Semi-supervised learning

1 Introduction

With the emergence of e-commerce systems, the number of online credit card transactions has skyrocketed. However, not all of these transactions are legitimate – worldwide card fraud losses in 2017 reached 24.26 billion US dollars, an increase of 6.4% from 2016, and the forecast for the following years goes in the same direction [8]. This huge amount of loss has led to the development of a series of countermeasures to limit the number of frauds. Among these countermeasures, Fraud Detection System (FDS) aims to identify perpetrated fraud as soon as possible [4].

The credit card fraud detection domain presents a number of challenging issues [1, 7]. Firstly, there are millions of credit card transactions processed each day, creating a massive stream of data. That is why data mining and machine learning play an important role in FDS, as they are often applied to extract and uncover the hidden truth behind very large quantities of data [26]. Secondly, the data are unbalanced: there is (fortunately) a more prevalent number of genuine transactions than fraudulent ones. The main risk with unbalanced classes is that the classifier tends to be overwhelmed by the majority class and to ignore the minority class [22]. Thirdly, the data are exposed to a concept drift as the habits

and behaviors of the consumers and fraudsters change over time [9]. Finally, the FDS needs to process the acceptance check of an online credit-card transaction within a few seconds to decide whether to pursue the transaction or not [36].

This work focuses on automatically detecting fraudulent e-commerce transactions using network-related features and free energy distance [20]. Our work is based on a recent paper [21] which introduced several improvements to an existing collective inference algorithm called APATE [36]. More precisely, this algorithm starts from a defined number of known frauds and propagates the fraudulent influence through a graph to obtain a risk score, quantifying the fraudulent behavior for each transaction, cardholder, and merchant [36]. In short, the main contributions of this paper are:

- a new way of computing the free energy distance [20] scaling on large, sparse, graphs;
- an application of this method to the fraud detection field through the adaptation of the existing FDS APATE [36];
- an experimental comparison between this method and others on a large real-life e-commerce credit card transaction dataset obtained from Worldline SA/NV.

The remainder of the paper is organised as follows. Section 2 contains the related work. Section 3 introduces the proposed contributions. In Sect. 4, we present the experimental comparisons and analyse the results. Finally, Sect. 5 concludes the paper.

2 Related Work

Over the past few years, credit card fraud detection has generated a lot of interest and a wide range of techniques has been suggested. However, the number of publications available is only the tip of the iceberg. Indeed, credit card issuers protect the sharing of data and most algorithms are produced in-house, concealing the details of the models [36].

Credit card fraud detection techniques can be seen as a classification problem. Therefore, it can be categorized into three broad types of learning: supervised (SL), unsupervised (USL) and semi-supervised (SSL). The most widespread approach is the supervised one which uses the information content in the labels, i.e. ‘fraud’ and ‘genuine’ in our case, to build a classification model. Common supervised methods are logistic regression [30], decision trees [30], Bayes minimum risk [2], support vector machines [10], meta-learning [7], case-based reasoning [38], genetic algorithms [12], hidden Markov models [3, 33], association rules [29], random forest [10] and, the most prevalent one for the moment, artificial neural networks [10, 18, 40]. Unlike supervised techniques, unsupervised learning does not use the class label to build the model, but simply extracts clusters of similar observations while maximizing the difference between these clusters. Common unsupervised methods include standard clustering methods, self-organizing map [39] and peer group analysis [37]. The interested reader is advised to consult [10, 26, 41] for more information.

The last category of classification techniques is semi-supervised learning which lies between supervised and unsupervised techniques, since it constructs predictive models using labeled samples together with a usually larger amount of unlabeled samples [13]. Some common semi-supervised methods are graph-based approaches, which consist in the creation of a graph model that reflects the relations included in the data and then transfers the labels on the graph to build a classification model [41]. Compared to the two other categories presented before, there are few publications about semi-supervised methods applied to card fraud detection. Ramaki et al. [28] proposed a model on a semantic connection between data stored for every transaction fulfilled by a user, then represent it by ontology graph and finally store them in patterns databases. Cao et al. [6] suggested Hit-Fraud, a collective fraud detection algorithm that captures the inter-transaction dependency based on a heterogeneous information network. Molloy et al. [24] presented a new approach to cross channel fraud detection based on feature extraction techniques applied to a graph of transactions. Finally, Lebichot et al. [21] proposed several improvements to an existing FDS, APATE, which spreads fraudulence influence through a graph by using a limited set of confirmed fraudulent transactions [36]. Our FDS is of this type.

As mentioned earlier, we base our approach on the previous work of Lebichot et al. [21] and on the methodology of APATE [36]. The rest of this section summarizes these two works to make this paper self-contained.

APATE starts by building a real tripartite symmetric transaction/cardholder/merchant adjacency matrix $\mathbf{A}^{\text{tri}} = (a_{ij})$ based on a list of time stamped, labeled transactions where each cardholder and merchant is known,

$$\mathbf{A}^{\text{tri}} = \begin{bmatrix} \mathbf{0}_{t \times t} & \mathbf{A}_{t \times c} & \mathbf{A}_{t \times m} \\ \mathbf{A}_{c \times t} & \mathbf{0}_{c \times c} & \mathbf{0}_{c \times m} \\ \mathbf{A}_{m \times t} & \mathbf{0}_{m \times c} & \mathbf{0}_{m \times m} \end{bmatrix}$$

where $\mathbf{A}_{c \times t}$ is a biadjacency matrix where cardholders are linked with their corresponding transactions, $\mathbf{A}_{m \times t}$ is a biadjacency matrix where merchants are linked with their corresponding transactions and $\mathbf{0}_{\dots \times \dots}$ is a correctly sized matrix full of zeros. Moreover, a column vector $\mathbf{z}_0^{\text{tri}} = [\mathbf{z}_0^{\text{Trx}}; \mathbf{z}_0^{\text{CH}}; \mathbf{z}_0^{\text{Mer}}]$, containing the risk score of each transaction (Trx), cardholder (CH) and merchant (Mer) is created and initialized with zeroes, except for known fraudulent transactions (Trx) which are set to one.

APATE integrates also a time decay factor in order to address the dynamic behavior of fraud. Interested readers can consult the original APATE paper [36] for more information as it is not crucial to understand the framework in detail here. At the end, we obtain four pairs of \mathbf{A}^{tri} and $\mathbf{z}_0^{\text{tri}}$ corresponding to four different time windows: no decay, day decay, or short term (ST), week decay, or medium term (MT) and monthly decay, or long term (LT).

Then for each of the four time window, in order to spread fraudulence influence through the tripartite graph, the vector $\mathbf{z}_0^{\text{tri}}$ is updated following an iterative procedure similar to the PageRank algorithm [27], namely the random walk with restart (RWWR) [35]:

$$\mathbf{z}_k^{\text{tri}} = \alpha \mathbf{P}^T \mathbf{z}_{k-1}^{\text{tri}} + (1 - \alpha) \mathbf{z}_0^{\text{tri}} \quad (1)$$

where k is the iteration number, $\mathbf{P} = (p_{ij}) = (\frac{a_{ij}}{a_{i\bullet}})$ is the transition probability matrix [13] associated to \mathbf{A}^{tri} , α is the probability to continue the walk, and symmetrically $(1 - \alpha)$ is the probability to restart the walk from a fraudulent transaction. Equation 1 is iterated until convergence, to reach $\mathbf{z}_{k^*}^{\text{tri}}$ (where k^* stands for k at convergence) from which three new feature vectors can be extracted, $\mathbf{z}_{k^*}^{\text{Trx}}$, $\mathbf{z}_{k^*}^{\text{CH}}$ and $\mathbf{z}_{k^*}^{\text{Mer}}$. These three features correspond to a risk measure for each transaction, cardholder and merchant respectively. Therefore, for each transaction, there are 12 new graph based features created.

As this procedure cannot be computed in a few seconds, the scores for each transaction, cardholder and merchant are only re-estimated once a day or once per hour, in order to analyse transactions that will occur during the day. In cases where new merchants or cardholders appear, their scores are set to zero as nothing can be inferred from the past graph data. The risk score of a new transaction that did not yet occur in the past can be approximated using an update formula presented in [36].

Finally, APATE combines those 12 graph-based features with the transaction-related features initially present in the dataset (see [36] for details), and use these as input of a random forest classifier.

While APATE is showing good performance, according to Lebichot et al. [21], it can be improved in three ways by: dealing with hubs, introducing a time gap and including investigators feedback.

The first way of improvement consists in dealing with hubs, which are nodes having a high degree, i.e. a large number of links with other nodes. Due to their connections to a lot of transactions, hubs tend to accumulate a high risk score. A simple solution to counterbalance this accumulation is to divide the risk score by the node degree after convergence of Eq. 1. Lebichot et al. [21] make the link between this solution and the Regularized Commute Time Kernel (RCTK) [23] in the sense that the elements of this kernel have the same interpretation as for the RWWR used in APATE. Therefore, they recommended the use of RCTK to deal with the problem of hubs.

Their second proposal is to introduce a time gap between the training set and the test set. Lebichot et al. [21] explain that, in most real FDS, the model cannot be based on the past few days, as is proposed in APATE, for two reasons. The first reason is that, in a real setting, the fraudulent transaction tags cannot be known without human investigator feedback. However, this feedback usually takes several days, mainly because it is often the cardholders that report undetected fraud. The second reason is that the strategy of the fraudsters changes over time and so it is less reliable to build the model on old data.

The third way of improvement consists in including feedback from the investigators on the predictions of the previous days. Even if it appears clear, in view of the second proposal, that it is impossible to know all fraud tags for the gap set, it is still conceivable that a fraction of previous alerts have been confirmed or overturned by human investigators (which is indeed the case in practice). We will refer later to this option as FB (for feedback) in Sect. 4.

3 The Free Energy Distance

As discussed previously, the main contributions of this work are three-fold, (1) to propose a way of computing the free energy distance scaling on large, sparse, graphs and (2) to incorporate the free energy framework into a FDS and (3) evaluate its performance on a real-world large-scale fraud detection problem. In this section, we start by providing a short account of the free energy distance and its properties, before discussing its implementation in the FDS. Note that this distance measure between nodes obtained very good results in a number of semi-supervised classification and clustering tasks [14, 16, 31, 32].

3.1 Background

The free energy distance [20], also known as the bag-of-paths potential distance [14], is a distance measure between nodes of a directed, strongly connected, graph based on the bag-of-paths framework. It is usually introduced by considering a statistical physics framework where it corresponds to the minimized free energy¹ of the bag-of-paths system connecting two nodes, but we will consider here a more intuitive explanation.

We already introduced the random walk on the graph whose transition probability matrix is \mathbf{P} , see Eq. 1. Recall that its elements are nonnegative and each of its rows sums to one. The free energy distance will be computed to some nodes of interest \mathcal{A} (in our application, the fraudulent nodes), called *target nodes*. These nodes are made killing and absorbing by setting the corresponding rows in the transition probability matrix to zero. Note that if the original graph is not strongly connected, we used a common trick, namely to add a new absorbing, killing, (sink) node connected to the set of target nodes \mathcal{A} with a directed link. In addition, we also assume that there is a nonnegative cost $c_{ij} \geq 0$ associated to each edge (i, j) of the graph with $\mathbf{C} = (c_{ij})$ being the cost matrix. The cost on an edge is assigned depending on the application and quantifies, in the model, the difficulty of following this edge in the random walk [13]. In our application, we fixed the cost to $c_{ij} = 1/a_{ij}$.

Then, a new matrix $\mathbf{W} = \mathbf{P} \circ \exp[-\theta \mathbf{C}]$ is introduced, where \circ is the elementwise (Hadamard) product and θ is a positive parameter (the inverse temperature). This matrix is substochastic because each of its row sums is less or equal to 1 and at least one row sum is strictly less than 1 (for example the killing, absorbing, nodes whose row sum is equal to zero). In fact, this matrix defines a transition probability matrix of a *killed random walk* on the graph, because at each time step, when visiting a node i , the random walker has a probability $0 \leq (1 - \sum_{j \in \mathcal{Succ}(i)} w_{ij}) \leq 1$, where $\mathcal{Succ}(i)$ is the set of successor nodes of node i , of giving up the walk – we then say that the walker is killed. The larger the cost to successors, the larger the probability of being killed.

In this context, it can be shown that the *directed free energy dissimilarity* $\phi_{i,\mathcal{A}}$ between any node $s = i$ (starting node) and the absorbing target nodes

¹ Expected total cost of the paths plus scaled relative entropy of the probability distribution of following these paths (see [20] for details).

in \mathcal{A} is simply $-\frac{1}{\theta} \log P(\text{reaching}(\mathcal{A})|s = i)$, that is, minus the logarithm of the probability of surviving during the killed random walk, i.e., of reaching an absorbing node without being killed during the walk [14]. Let us now explain how it can be computed.

The free energy distance between two nodes i and j is obtained by $\Delta_{ij} = \frac{\phi_{ij} + \phi_{ji}}{2}$. Besides being a distance measure, it has many interesting properties. One of those is the fact that it interpolates between two widely used distances, the shortest path distance and the commute cost distance (which is proportional to the effective resistance also called resistance distance [13]). Indeed, if the parameter θ approaches ∞ , the free energy distance converges to the shortest path distance [13]. Conversely, if θ approaches 0^+ , we recover the commute cost distance [13]. The details and proofs of these properties are available in [14].

The free energy distance between all pairs of nodes can be computed by performing a matrix inversion [20]. However, Fran oisse et al. [14] showed that the directed free energy distance to a unique, fixed, target node t (the set of absorbing nodes reduces to node t) can also be computed thanks to an extension of the Bellman-Ford formula:

$$\phi_{it}(\tau + 1) = \begin{cases} -\frac{1}{\theta} \log \left[\sum_{j \in \mathcal{S}_{\text{succ}}(i)} p_{ij} \exp[-\theta(c_{ij} + \phi_{jt}(\tau))] \right] & \text{if } i \neq t \\ 0 & \text{if } i = t \end{cases} \quad (2)$$

where τ is the iteration number². This expression uses the softmin, or log-sum-exp, operator [25], $\text{softmin}_{\theta, \mathbf{q}}(\mathbf{x}) = -\frac{1}{\theta} \log \sum_{j=1}^n q_j \exp[-\theta x_j]$ (with $q_j \geq 0$ and $\sum_{j=1}^n q_j = 1$) where, in the present context of Eq. 2, $q_j = p_{ij}$ and $x_j = (c_{ij} + \phi_{jt}(\tau))$. This operator interpolates between the minimum and the weighted average of the values x_j with weights q_j . In fact, Eq. 2 is nothing else than the Bellman-Ford formula (based on dynamic programming; see, e.g., [15]) where the minimum operator is replaced by the soft minimum operator. Equation 2 can be iterated until convergence to the directed free energy distances. The main advantage of this formulation is that it can be applied on large, sparse, graphs thanks to some specific techniques which are explained below. After convergence, ϕ_{it} contains $-\frac{1}{\theta} \log$ of the probability of surviving during a killed random walk from i to t with transition matrix \mathbf{W} [14].

3.2 Computing the Directed Free Energy Distance on Large Graphs

In order to scale the computation of the free energy, we will use Eq. 2 and rely on two different ideas: (i) the *log-sum-exp* trick and (ii) to *bound the length* of the set of paths on which the distance is computed. This second point brings another benefit: it allows to tune the length of the walks, which has been shown to improve the performance in some situations (see, e.g., [5, 23]).

² Notice that the usual free energy distance (not directed) is defined by symmetrization of ϕ_{ij} (Eq. 2, so that the resulting distance is symmetric [14, 20]), but this quantity will not be used in this work.

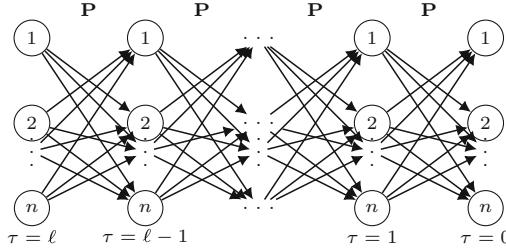


Fig. 1. The directed lattice derived from the original graph. It only considers walks of length up to ℓ .

The log-sum-exp trick [17, 19, 25] aims at pre-computing $x^* = \min_{j \in \{1 \dots n\}} (x_j)$, leading to the form $\text{softmax}_{\theta, \mathbf{q}}(\mathbf{x}) = x^* - \frac{1}{\theta} \log \sum_{j=1}^n q_j \exp[-\theta(x_j - x^*)]$ and then neglecting the terms in the summation for which $\theta(x_j - x^*)$ is too large (exceeds a certain threshold). This is a kind of pruning and has two benefits: it reduces significantly the number of terms to be computed and it avoids numerical underflow problems.

Whereas the standard bag-of-paths framework is based on paths of unbounded length (from 0 to ∞), our second technique considers only walks bounded by a given length ℓ [5, 23]. This is done by defining a directed lattice L unfolding the original graph G in terms of increasing walk lengths. More precisely, this lattice is made of the graph nodes repeated at walk lengths $\tau = 0, 1, \dots, \ell$, usually with $\ell \ll n$ [23] (see Fig. 1). Then, transitions are only allowed from nodes at walk length τ to successor nodes at length $\tau - 1$ by means of the transition matrix \mathbf{P} of the killed random walk associated to the graph. This lattice represents a bounded random walk on the graph G where walks' lengths are in the interval $[0, \ell]$. The combination of these two tricks allows to scale on large, sparse, graphs – many edges are pruned and the computation occurs on a (usually small) grid.

In this context, on lattice L and considering now a set of target nodes \mathcal{A} , Eq. 2 becomes, for the initialization of the distances at $\tau = 0$, corresponding to zero-length walks,

$$\phi_{i,\mathcal{A}}(0) = \begin{cases} \infty & \text{if } i \notin \mathcal{A} \\ 0 & \text{if } i \in \mathcal{A} \end{cases} \quad (3)$$

Moreover, $\phi_{i,\mathcal{A}}(\tau)$ contains the directed free energy distance from node i to the absorbing nodes in \mathcal{A} when considering walks up to length τ . A resulting distance of ∞ means that no walk of length up to τ exists between node i and a node in \mathcal{A} – absorbing nodes cannot be reached in τ steps. Then, for walk length $\tau > 0$,

$$\phi_{i,\mathcal{A}}(\tau + 1) = \begin{cases} -\frac{1}{\theta} \log \left[\sum_{j \in \text{succ}(i)} p_{ij} \exp [-\theta(c_{ij} + \phi_{j,\mathcal{A}}(\tau))] \right] & \text{if } i \notin \mathcal{A} \\ 0 & \text{if } i \in \mathcal{A} \end{cases} \quad (4)$$

This recurrence relation defines the directed free energy distance to target nodes \mathcal{A} that will be used in our fraud detection application.

3.3 Application to the Fraud Detection Problem

In order to incorporate the free energy (FE) framework into FDS APATE to create the FDS called FraudsFree (FF), we introduce some other modifications.

The first modification is related to the computation of the risk score vector $\mathbf{z}_{k*}^{\text{tri}} = \phi_{k*}^{\text{tri}}$, presented in Eq. 1. For that, we use Eq. 4 by considering each known fraud (Trx) as an absorbing node $a \in \mathcal{A}$. We iterate this equation until convergence to obtain the distance between all the nodes i and the set \mathcal{A} which corresponds to the risk score of each transaction, cardholder and merchant. At this point, a score near 0 represents a fraud and a high value represents a genuine transaction. In order to keep the same interpretation of the risk score as in APATE³, we apply the following transformation

$$\phi_{k*}^{\text{tri}} = \max(\phi_{k*}^{\text{tri}}) - \phi_{k*}^{\text{tri}} \quad (5)$$

As for APATE [36] (see Sect. 2), we set the score of a new transaction j that did not yet occur in the past between a cardholder k and a merchant i via Eq. 2, which provides the following expression

$$\begin{aligned} \text{score}(Trx_j) = & -\frac{1}{\theta} \log \left[p_{ji} \exp \left[-\theta(c_{ji} + \text{score}(Mer_i)) \right] \right] \\ & -\frac{1}{\theta} \log \left[p_{jk} \exp \left[-\theta(c_{jk} + \text{score}(CH_k)) \right] \right] \end{aligned} \quad (6)$$

where $p_{ji} = p_{jk} = 0.5$ because a transaction is linked by construction with one merchant and one cardholder and we fixed $c_{ji} = c_{jk} = 1$ as the transaction appends now (no decay), but this is still a degree of freedom of our method that we left for further work.

4 Experimental Comparisons and Discussion

To evaluate our approach following the methodology of [21], we perform a comparison between the different versions of our FraudsFree (FF) model and the other variations of APATE (Random Walk With Restart (RWWR) and Regularized Commute Time Kernel (RCTK)), in supervised (SL) and semi-supervised learning (SSL) with feedback (FB), on the same real-life e-commerce credit card transaction dataset as [21]. The dataset contains 16 socio-demographic features on 25,445,744 e-commerce transactions gathered during 139 days. The data are highly imbalanced, with only 78,119 frauds among the transactions (<0.31%). The average size of \mathbf{A}^{tri} is 3,910,783. This dataset does not focus on a certain type of card fraud but contains all reported fraudulent transactions in the investigated time period [21]. Besides the 16 original features, a set of 12 graph-based features per node, as described in Sect. 2 and 3, is created for each method. A small sample of this dataset is available on www.kaggle.com/mlg-ulb/creditcardfraud

³ So that a score near 0 represents a genuine transaction and a high value represents a fraud. The higher the score, the higher the risk.

but the data are anonymised and the transactions are not presented day by day. Finally, all these features are fed into a class-rebalanced random forest with 400 trees. Each tree is built based on a random selection of 4 features of the original dataset and 4 graph-based features.

In order to assess the performance of each method, we select two measures. In accordance with field experts, we chose the Precision@100 in terms of card (Card Pr@100) [21,34] (which is the most realistic setting). More precisely, we select the more fraudulent transactions according to the model until we screen 100 cards. As a second measure, we use the average time in seconds required to create the tripartite graph and extract the 12 graph features between Day 41 and Day 139.

Table 1. Mean Card Pr@100 between Day 41 and Day 139 for each of the 8 methods (see Sects. 2, 3 and 4 for acronyms). The average time in seconds required to create the tripartite graph and extract the 12 graph features between Day 41 and Day 139 is also reported for each of the 8 methods.

Classifier	Hubs	Learning	Feedback	Card Pr@100	Time	Best parameter
RWWR SL = APATE	No	SL	No	18.19	155.40	0.1
RWWR SSL+FB	No	SSL	Yes	20.90	273.16	0.9
RCTK SL	Yes	SL	No	21.48	195.43	0.9
RCTK SSL+FB	Yes	SSL	Yes	24.03	294.07	0.7
RCTK SSL+FB 5 ITER	Yes	SSL	Yes	22.82	122.60	0.9
FF SL	No	SL	No	16.13	204.48	5
FF SSL+FB	No	SSL	Yes	24.20	489.43	0.5
FF SSL+FB 5 ITER	No	SSL	Yes	24.01	155.40	0.5

Concerning the hyperparameters, the RWWR and RCTK methods consider tuning values of $\alpha = \{0.1, 0.3, 0.5, 0.7, 0.85, 0.9\}$ and, for the FF methods $\theta = \{0.1, 0.5, 1, 5, 10\}$. For each method, we tuned its parameter based on the mean Card Pr@100 between Day 30 and Day 40. The results for the best parameter is presented in Table 1⁴. We obtained these results by applying a sliding window technique in accordance with expert knowledge [21]. We set 15 days in the training set, 7 days in the gapset (see Sect. 2 and [21]), 1 day in the test set and we shifted the sliding window day by day. Furthermore, in order to exploit all the properties of the bounded FE, we select the number of iterations (the walk length ℓ) for the best FF-based method based on a reasonable trade-off Card Pr@100/time. For the sake of comparisons, we also limited the number of RCTK iterations to the same number as in FF-based method.

⁴ Numerical values differ from [21] because the dataset was further curated: some obvious fraud cases were removed.

To analyse the results with Card Pr@100 of Table 1, we use a nonparametric Friedman-Nemenyi statistical test and a Wilcoxon signed-ranks tests [11]. We perform all statistical tests at a level of confidence of 95%, which amounts to taking an α of 0.05. From the results of the Nemenyi test illustrated in Fig. 2, four methods perform equivalently, in terms of Card Pr@100, to the best one: the FE SSL+FB, the FF SSL+FB 5 ITER, the RCTK SSL+FB and the RCTK SSL+FB 5 ITER. However, the Wilcoxon tests report that RCTK SSL+FB 5 ITER is significantly inferior to the three other methods in one-to-one comparisons (with respective p -values of 0.0055, 0.0252 and 0.0030). Even if we cannot ensure a statistical difference between the top three, we still observe that there are some differences in terms of their computation times. The FF SSL+FB 5 ITER method is the fastest of the top three with a reduction of 47.16% in computation time compared to the best method proposed by Lebichot et al. [21], RCTK SSL+FB. All results were obtained with Matlab (version R2017a) running on an Intel Xeon with 2×8 3.6 GHz processors and 128 GB of RAM.

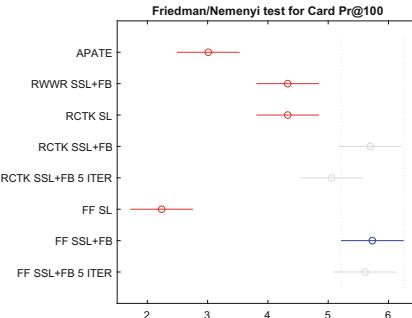


Fig. 2. Mean ranks and 95% Nemenyi confidence intervals for the 8 methods (see Table 1) based on the Card Pr@100. Two methods are considered as significantly different if their confidence intervals do not overlap.

5 Conclusion

In this paper, we investigate a version of the free energy distance that scales on large, sparse graphs. It is used in the existing Fraud Detection System APATE in order to extract features from the graph of transactions. Thanks to the properties of the free energy, we manage to reduce the computational time and improve the scalability of the Fraud Detection System. The Fraud Detection System based on the free energy distance, FraudsFree, is competitive as it obtains a Pr@100 score on fraudulent card prediction comparable to the previous work of Lebichot et al. [21] with a significant speed-up in computation. This shows that the free energy distance can be used on real-word applications involving large graphs. One considered further work is to deal with the hub nodes by modifying directly the cost matrix, as it has shown good results in other studies [21].

Another avenue that could be explored is to determine if our approach is complementary, or just redundant, to existing fraud defence lines of our industrial partner.

Acknowledgements. This work was partially supported by the Immediate funded by Wallon Region project and by the Defeatfrauds project funded by Innoviris. We thank these institutions for giving us the opportunity to conduct both fundamental and applied research. We also thank Worldline SA/NV, R&D, for providing us the data and expertise.

References

1. Abdallah, A., Maarof, M.A., Zainal, A.: Fraud detection system: a survey. *J. Network Comput. Appl.* **68**, 90–113 (2016)
2. Bahnsen, A.C., Stojanovic, A., Aouada, D., Ottersten, B.: Cost sensitive credit card fraud detection using bayes minimum risk. In: 2013 12th International Conference on Machine Learning and Applications, vol. 1, pp. 333–338. IEEE (2013)
3. Bhusari, V., Patil, S.: Study of hidden markov model in credit card fraudulent detection. *Int. J. Comput. Appl.* **20**(5), 33–36 (2011)
4. Bolton, R.J., Hand, D.J.: Statistical fraud detection: a review. *Stat. Sci.* **1**, 235–249 (2002)
5. Callut, J., Francoisse, K., Saerens, M., Dupont, P.: Semi-supervised classification from discriminative random walks. In: W. Daelemans, K. Morik (eds.) *Proceedings of the 19th European Conference on Machine Learning (ECML 2008)*, Lecture Notes in Artificial Intelligence, vol. 5211, pp. 162–177. Springer, Berlin (2008)
6. Cao, B., Mao, M., Viidu, S., Yu, P.: Collective fraud detection capturing inter-transaction dependency. In: KDD 2017 Workshop on Anomaly Detection in Finance, pp. 66–75 (2018)
7. Chan, P.K., Fan, W., Prodromidis, A.L., Stolfo, S.J.: Distributed data mining in credit card fraud detection. *IEEE Intell. Syst.* **14**(6), 67–74 (1999)
8. Consultants, H.: The nilson report issue 1142 (2018). <https://nilsonreport.com>
9. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., Bontempi, G.: Credit card fraud detection: a realistic modeling and a novel learning strategy. *IEEE Trans. Neural Networks Learn. Syst.* **29**(8), 3784–3797 (2018)
10. Dal Pozzolo, A., Caelen, O., Le Borgne, Y.A., Waterschoot, S., Bontempi, G.: Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Syst. Appl.* **41**(10), 4915–4928 (2014)
11. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. *J. Mach. learn. res.* **7**, 1–30 (2006)
12. Duman, E., Elikucuk, I.: Solving credit card fraud detection problem by the new metaheuristics migrating birds optimization. In: International Work-Conference on Artificial Neural Networks, pp. 62–71. Springer, Berlin (2013)
13. Fouss, F., Saerens, M., Shimbo, M.: Algorithms and Models for Network Data and Link Analysis. Cambridge University Press, Cambridge (2016)
14. Françoisse, K., Kivimäki, I., Mantrach, A., Rossi, F., Saerens, M.: A bag-of-paths framework for network data analysis. *Neural Networks* **90**, 90–111 (2017)
15. Gondran, M., Minoux, M.: Graphs and Algorithms. Wiley, Hoboken (1984)
16. Guex, G., Courtain, S., Saerens, M.: Covariance and correlation kernels on a graph in the generalized bag-of-paths formalism. arXiv preprint [arXiv:1902.03002](https://arxiv.org/abs/1902.03002) (2019)

17. Huang, X., Ariki, Y., Jack, M.: Hidden Markov Models for Speech Recognition. Edinburgh University Press, Edinburgh (1990)
18. Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.E., He-Guelton, L., Caelen, O.: Sequence classification for credit-card fraud detection. *Expert Syst. Appl.* **100**, 234–245 (2018)
19. Kivimäki, I.: Distances, centralities and model estimation methods based on randomized shortest paths for network data analysis. Ph.D. thesis, UCL-Université Catholique de Louvain (2018)
20. Kivimäki, I., Shimbo, M., Saerens, M.: Developments in the theory of randomized shortest paths with a comparison of graph node distances. *Physica A* **393**, 600–616 (2014)
21. Lebichot, B., Braun, F., Caelen, O., Saerens, M.: A graph-based, semi-supervised, credit card fraud detection system. In: Cherifi, H., Gaito, S., Quattrociocchi, W., Sala, A. (eds.) International Workshop on Complex Networks and their Applications, pp. 721–733. Springer, Cham (2016)
22. Liu, Q., Wu, Y.: Supervised learning. Encyclopedia of the Sciences of Learning, pp. 3243–3245 (2012)
23. Mantrach, A., Van Zeebroeck, N., Francq, P., Shimbo, M., Bersini, H., Saerens, M.: Semi-supervised classification and betweenness computation on large, sparse, directed graphs. *Pattern Recognit.* **44**(6), 1212–1224 (2011)
24. Molloy, I., Chari, S., Finkler, U., Wiggerman, M., Jonker, C., Habeck, T., Park, Y., Jordens, F., van Schaik, R.: Graph analytics for real-time scoring of cross-channel transactional fraud. In: Grossklags, J., Preneel, B. (eds.) International Conference on Financial Cryptography and Data Security, vol. 9603, pp. 22–40. Springer, Berlin (2016)
25. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT press, Cambridge (2012)
26. Ngai, E.W., Hu, Y., Wong, Y.H., Chen, Y., Sun, X.: The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature. *Decision Support Systems* **50**(3), 559–569 (2011)
27. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)
28. Ramaki, A.A., Asgari, R., Atani, R.E.: Credit card fraud detection based on ontology graph. *Int. J. Secur. Priv. Trust Manag. (IJSPTM)* **1**(5), 1–12 (2012)
29. Sánchez, D., Vila, M., Cerda, L., Serrano, J.M.: Association rules applied to credit card fraud detection. *Expert syst. appl.* **36**(2), 3630–3640 (2009)
30. Shen, A., Tong, R., Deng, Y.: Application of classification models on credit card fraud detection. In: 2007 International conference on service systems and service management, pp. 1–4. IEEE (2007)
31. Sommer, F., Fouss, F., Saerens, M.: Comparison of graph node distances on clustering tasks. In: Artificial Neural Networks and Machine Learning – Proceedings of ICANN 2016. Lecture Notes in Computer Science, vol. 9886, 192–201. Springer Cham (2016)
32. Sommer, F., Fouss, F., Saerens, M.: Modularity-driven kernel k-means for community detection. Artificial Neural Networks and Machine Learning (Proceedings of ICANN 2016). Lecture Notes in Computer Science, vol. 10614, pp. 423–433. Springer, Cham (2017)
33. Srivastava, A., Kundu, A., Sural, S., Majumdar, A.: Credit card fraud detection using hidden markov model. *IEEE Trans. dependable secure comput.* **5**(1), 37–48 (2008)

34. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press Inc., Cambridge (2008)
35. Tong, H., Faloutsos, C., Pan, J.Y.: Fast random walk with restart and its applications. In: Sixth International Conference on Data Mining (ICDM 2006), pp. 613–622. IEEE (2006)
36. Van Lasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B.: Apate: a novel approach for automated credit card transaction fraud detection using network-based extensions. *Decis. Support Syst.* **75**, 38–48 (2015)
37. Weston, D.J., Hand, D.J., Adams, N.M., Whitrow, C., Juszczak, P.: Plastic card fraud detection using peer group analysis. *Adv. Data Anal. Classif.* **2**(1), 45–62 (2008)
38. Wheeler, R., Aitken, S.: Multiple algorithms for fraud detection. In: Ellis, R., Moulton, M., Coenen, F. (eds.) Applications and Innovations in Intelligent Systems VII, pp. 219–231. Springer, London (2000)
39. Zaslavsky, V., Strizhak, A.: Credit card fraud detection using self-organizing maps. *Inf. Secur.* **18**, 48 (2006)
40. Zhang, Z., Zhou, X., Zhang, X., Wang, L., Wang, P.: A model based on convolutional neural network for online transaction fraud detection. *Secur. Commun. Networks* **2018**, 9 (2018)
41. Zhou, X., Cheng, S., Zhu, M., Guo, C., Zhou, S., Xu, P., Xue, Z., Zhang, W.: A state of the art survey of data mining-based fraud detection and credit scoring. In: MATEC Web of Conferences, vol. 189. EDP Sciences, Les Ulis (2018)



Visualizing Structural Balance in Signed Networks

Edoardo Galimberti^{1,2(✉)}, Chiara Madeddu¹, Francesco Bonchi²,
and Giancarlo Ruffo¹

¹ University of Turin, Turin, Italy

edoardo.galimberti@isi.it

² ISI Foundation, Turin, Italy

Abstract. *Network visualization* has established as a key complement to network analysis since the large variety of existing network layouts are able to graphically highlight different properties of networks. However, *signed networks*, i.e., networks whose edges are labeled as friendly (positive) or antagonistic (negative), are target of few of such layouts and none, to our knowledge, is able to show *structural balance*, i.e., the tendency of cycles towards including an even number of negative edges, which is a well-known theory for studying friction and polarization.

In this work we present **Structural-balance-viz**: a novel visualization method showing whether a connected signed network is balanced or not and, in the latter case, how close the network is to be balanced. **Structural-balance-viz** exploits spectral computations of the signed Laplacian matrix to place network's nodes in a Cartesian coordinate system resembling a balance (a scale). Moreover, it uses edge coloring and bundling to distinguish positive and negative interactions. The proposed visualization method has characteristics desirable in a variety of network analysis tasks: **Structural-balance-viz** is able to provide indications of balance/polarization of the whole network and of each node, to identify two factions of nodes on the basis of their polarization, and to show their cumulative characteristics. Moreover, the layout is reproducible and easy to compare. **Structural-balance-viz** is validated over synthetic-generated networks and applied to a real-world dataset about political debates confirming that it is able to provide meaningful interpretations.

Keywords: Network visualization · Signed networks · Structural balance · Spectral theory

1 Introduction

Signed networks are simple yet informative network representations in which edges are annotated as positive or negative [11]. They are applied in a large variety of domains in which interactions between entities are either friendly or antagonistic, e.g., international relations [9], and online social media and social networks [21]. The theory of *structural balance* has established as the standard

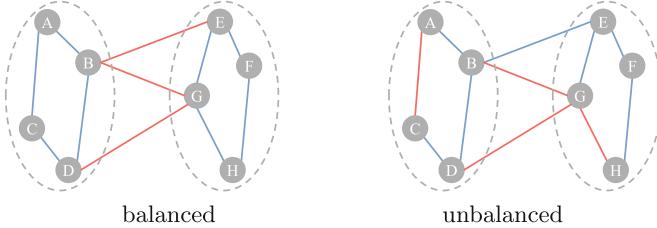


Fig. 1. Examples of balanced (left) and unbalanced (right) networks. Positive edges are reported in blue, while negative edges in red.

for studying, from a theoretical standpoint in sociology and psychology, the formation of opinions in both individuals and social groups. Structural balance is widely applied to signed networks, e.g., for the analysis of social media [18], and the study of opinion separation [22]. A signed network has been proved to be *structurally balanced* or *balanced* if and only if all cycles are balanced, i.e., include an even number of negative edges [6]. As a consequence, network's nodes can be assigned to two different sets such that we find only positive ties between nodes in the same set and all negative ones between nodes of different sets [10]. Figure 1 shows two simple examples of balanced and unbalanced networks. The network on the left is balanced and has the two properties discussed above, i.e., all cycles are balanced and a clustering can be found in agreement to all edges' signs. On the other hand, the network on the right is not balanced: there are unbalanced cycles (e.g., the one composed by the node sequence $[A, B, D, C, A]$) and there are edges disagreeing with the clustering (e.g., edge (B, E)). Even if a balanced network represents the most natural configuration, structural balance is not necessarily a “positive” configuration, e.g., it is observed in the alliance network between European nations just before World War I [20]. Moreover, most of the large real-world networks are expected to be unbalanced since a single unbalanced cycle makes the whole network unbalanced. Therefore, it has also been shown the importance of measuring to what extent an unbalanced signed network is close to be balanced [17]. Structural balance is also linked with *group polarization*, i.e., the division of a group of entities (e.g., nodes of a network) into two subgroups each reaching consensus and having opposite opinions [5].

Network visualization has emerged as a key complement to standard network analysis techniques to fill the gap between computation and interpretation, communicate findings, and deepen insight [16]. A large variety of network layouts exists in literature [15] and, also, implemented for visualization applications, as, e.g., Gephi and Cytoscape. Surprisingly, little attention has been paid to the visualization of signed networks [17] and, to our knowledge, none of the existing layouts highlights structural balance properties of signed networks.

In this work we tackle the task of identifying, through a visualization, whether a connected signed network is balanced or unbalanced and, in the latter case, how much the network is unbalanced. The proposed visualization method, **Structural-balance-viz**, places nodes in a Cartesian coordinate system exploiting spectral

properties of the signed Laplacian matrix. Edges are colored and bundled to make positive and negative signs distinguishable and to ease the understanding of the global balance/polarization of the network. At a glance, it is possible to catch if a network is balanced: no positive edges cross the y -axis and no negative edges have both endpoints in the same quadrant, namely, the y -axis finds a partition of the nodes as explained in [10]. The visual perception of the portion of edges “disagreeing” with the partitioning, i.e., the fraction of positive edges crossing the y -axis and negative edges internal to a quadrant, gives an indication of the level of balance of a network. Moreover, we utilize the x -axis as a scale to show cumulative characteristics of the sets of nodes identified by the y -axis, and include a textual indication of the level of balance of the network under analysis in order to improve the comparability between different visualizations.

The layout produced by `Structural-balance-viz` has the following characteristics that are useful in a variety of network analysis tasks: (i) it shows whether the input network is balanced or not and, in the second case, how close the network is to be balanced; (ii) by nodes’ x -coordinate, it provides an indication of the contribute to the balance structure of the network and, also, of the individual balance/polarization of each node (such information might be exploited, e.g., for the task of finding non-polarized representatives [19]); (iii) it identifies two factions of nodes on the basis of their polarization which finds applications in clustering problems, e.g., 2-correlation-clustering [2,7]; (iv) the scale represented by the x -axis shows cumulative characteristics of the identified factions, e.g., size or internal clustering coefficient; and, (v) the resulting visualization are reproducible (desirable feature but not common to all network layouts, e.g., force based) and easy to compare in terms of balance structure. We verify such characteristics by running `Structural-balance-viz` on synthetic networks and a real-world dataset representing political debates.

The rest of the paper is structured as follows. Section 2 covers the related work, in Sect. 3 we describe our visualization procedure, while Sect. 4 shows the experimental validation and the real-world application. Finally, Sect. 5 concludes the paper.

2 Related Work

Structural Balance and Signed Networks. The concept of *structural balance* first appears as psychological theory of balance in triangles of sentiments. *Signed networks* are later introduced in the seminal work by Harary [11], who also generalizes the balance theory to signed networks [6]. Harary and Kabell develop a simple algorithm to test whether a given signed network is balanced [12] by enumerating the cycles in the network containing an even number of negative edges. A complete signed network is balanced if and only if all its triangles are balanced [10]. Akiyama et al. [1] study how to estimate the minimum number of sign changes required so that a signed network satisfies the balance property. Recent works link spectral properties of signed networks to the balance theory. Hou et al. [14] prove that a signed network is balanced if and only if the smallest

eigenvalue of the signed Laplacian is 0. Moreover, [13] investigates the relationship between the smallest eigenvalue of the signed Laplacian and the level of balance of a signed network.

A fundamental problem studied in signed networks is correlation clustering [4], i.e., partition the nodes into clusters so as to maximize (minimize) the number of edges that “agree” (“disagree”) with the partitioning. The 2-correlation-clustering problem [7], also known as the frustration-index problem [2], is also widely studied. Finally, a more recent line of work introduces the problem of discovering antagonistic communities in signed networks [5].

Network Visualization. Many *network visualizations* have been proposed in literature in order to graphically express specific characteristics, properties, and patterns of networks. Force-based visualizations map an energy function to the desired layout and minimize it [15]. Hive plots [16] place nodes on radially oriented linear axes according to a coordinate system defined by nodes characteristics and/or network properties. Eigenvectors are exploited for visualizing networks in different works. In particular, [17] studies the application of clustering, prediction, and visualization methods to signed networks by using the signed Laplacian and its eigenvalue decomposition. Despite using eigenvectors to place nodes in a Cartesian coordinate system, the visualization algorithm of [17] has different purposes and strongly differs from ours: (i) it wants to highlight clustering properties and not structural balance; (ii) it does not provide information about the contribute of each node to the balance/polarization structure; (iii) it does not cluster nodes into two factions and, therefore, it cannot show factions’ cumulative properties; and, (iv) the resulting layouts are hardly comparable between each other.

3 Visualizing Structural Balance

In this section we describe the details of **Structural-balance-viz**, the proposed visualization method whose main objective is to show whether a connected signed network is balanced or unbalanced and, in the latter case, how much the network is unbalanced.

First, we provide preliminary notations and definitions. We denote a signed undirected network as $G = (V, E_+, E_-)$, where V is a set of nodes, E_+ is a set of positive edges, and E_- is a set of negative edges. In this work, we require G to be connected. Let A be the signed adjacency matrix of G , i.e., for each pair of nodes $u, v \in V$, $A[u, v] = 1$ if $(u, v) \in E_+$, $A[u, v] = -1$ if $(u, v) \in E_-$, $A[u, v] = 0$ otherwise. Let also $\bar{D} = \text{diag}(\bar{d}_{u_1}, \dots, \bar{d}_{u_{|V|}})$ be the signed degree matrix of G , where $\bar{d}_u = \sum_{v \in V} |A[u, v]|$ represents the signed degree, i.e., the number of neighbors disregarding the sign, of a node $u \in V$. Finally, we define the signed Laplacian matrix of G as:

$$\bar{L} = \bar{D} - A.$$

Algorithm 1: Structural-balance-viz

Input: A signed network $G = (V, E_+, E_-)$ and a network measure μ (optional).
Output: A visualization of G .

```

/* Eigenvalue decomposition */
```

- 1 compute the signed Laplacian \bar{L} of G
- 2 compute the smallest eigenvalue λ_m of \bar{L} and its corresponding eigenvector \mathbf{v}_m
 /* Nodes coordinates */

- 3 $\mathbf{X} \leftarrow \emptyset; \mathbf{Y} \leftarrow \emptyset$
- 4 **forall** $u \in V$ **do**
 - 5 $\mathbf{X}[u] = \mathbf{v}_m[u]$
 - 6 $\mathbf{Y}[u] = |\{v \in V \mid \mathbf{v}_m[v] = \mathbf{v}_m[u] \wedge v < u\}|$
- 7 /* Edge partitioning */
- 8 $E_+^i = \{e = (u, v) \in E_+ \mid \mathbf{X}[u] = \mathbf{X}[v]\}$
- 9 $E_-^i = \{e = (u, v) \in E_- \mid \mathbf{X}[u] = \mathbf{X}[v]\}$
- 10 $E_+^e = E_+ \setminus E_+^i$
- 11 $E_-^e = E_- \setminus E_-^i$
 /* Drawing */
- 12 draw the Cartesian axes
- 13 draw the nodes in V according to \mathbf{X} and \mathbf{Y}
- 14 draw the edges in E_+^i in **blue** with **horizontal-external** bundling
- 15 draw the edges in E_-^i in **red** with **horizontal-internal** bundling
- 16 draw the edges in E_+^e in **blue** with **vertical-upper** bundling
- 17 draw the edges in E_-^e in **red** with **vertical-lower** bundling
 /* Additional features */
- 18 **if** $\mu \neq \text{NULL}$ **then**
 - 19 $C_l = \{u \in V \mid \mathbf{X}[u] < 0\}; C_r = \{u \in V \mid \mathbf{X}[u] \geq 0\}$
 - 20 let $\gamma = \mu(C_l) - \mu(C_r)$ be the angular coefficient of the x -axis
- 21 draw the label “ $y = \lambda_m$ ”

We now describe our algorithm for visualizing structural balance in signed networks, which is outlined as Algorithm 1. As mentioned beforehand, Structural-balance-viz makes use of the signed Laplacian of the input network G . In fact, it starts by computing the signed Laplacian together with its smallest eigenvalue λ_m and the corresponding eigenvector \mathbf{v}_m (Line 2). At this point, we already have all the information required for the visualization handy. At first, we identify the coordinates of the nodes in V and store them in \mathbf{X} and \mathbf{Y} (cycle starting at Line 4). The x -coordinate of each node u is directly obtained by the element of \mathbf{v}_m corresponding to u . Since more than a node might have the same abscissa and we want to avoid nodes to overlap, the y -coordinates are computed in order to distribute nodes having the same x -coordinate vertically. Next (Lines 7–10), edges are divided into four sets since, on the basis of the coordinates of their endpoints and of their sign, different layouts are applied:

- E_+^i contains the positive edges having two endpoint with the same x -coordinate;

- E_-^i contains the negative edges having two endpoint with the same x -coordinate;
- E_+^e contains the positive edges having two endpoint with different x -coordinate;
- E_-^e contains the negative edges having two endpoint with different x -coordinate.

Structural-balance-viz is then ready to draw the visualization (Lines 11–16). At first, the Cartesian axes and the nodes are positioned. Then, the edges are drawn exploiting coloring and bundling to highlight their sign. In particular, positive edges are depicted in blue, while negative edges in red. A positive edge $e_+ \in E_+$ is bundled towards the top of the visualization, if $e_+ \in E_+^e$, or externally, if $e_+ \in E_+^i$; while a negative edge $e_- \in E_-$ is bundled towards the bottom, if $e_- \in E_-^e$, or internally, if $e_- \in E_-^i$.

In order to improve the informativeness of our layout, we include two additional features in **Structural-balance-viz** (from Line 17): one wants to provide information about the two sets of nodes identified by the y -axis, while the latter has the aim of making different visualizations more comparable.

Any eigenvector \mathbf{v} of the signed Laplacian can be used to derive a partition of network's nodes into two sets on the basis of the sign of the corresponding elements in \mathbf{v} . Such partitioning is at the basis of spectral-clustering methods [8] and it can identify polarized structures, i.e., two sets of nodes showing high internal consensus and warring between each other [5]. In the proposed visualization, the two sets are identified by the nodes in the left and in the right quadrants, i.e., C_l and C_r computed at Line 18 of **Structural-balance-viz**, respectively. In practical applications, it is often of interest to know (and visualize) network measures of the two polarized sets, e.g., size, internal clustering coefficient, internal density of positive edges, ratio of positive edges, etc. We provide a simple visual expedient based on the angular coefficient of the x -axis that resembles the behavior of a scale. Let μ be the network measure of interest. Note that μ is an optional input parameter of **Structural-balance-viz** and the lines corresponding to this additional feature are executed if μ is actually provided in input. We define the angular coefficient of the x -axis as

$$\gamma = \mu(C_l) - \mu(C_r).$$

The work enclosed in [13, 14] proves theoretical bounds on the smallest eigenvalue of the Laplacian of a signed network and investigates its relationship with respect to the level of balance in the network. It is shown that a connected signed network is structurally balanced if and only if $\lambda_m = 0$, i.e., the smallest eigenvalue of the Laplacian is zero, and that the higher λ_m , the lower the level of balance of the network is. Therefore, λ_m is the simplest indicator to take into account for comparing structural balance in different networks (of equal densities). More complex indicators of balance could also be employed [3]. Ideally, the y -coordinate where the x -axis crosses the y -axis would be a simple manner to graphically show λ_m . Unfortunately, we devoted consistent effort to visualize such information in this way, but all attempts worsened the clarity of the layout

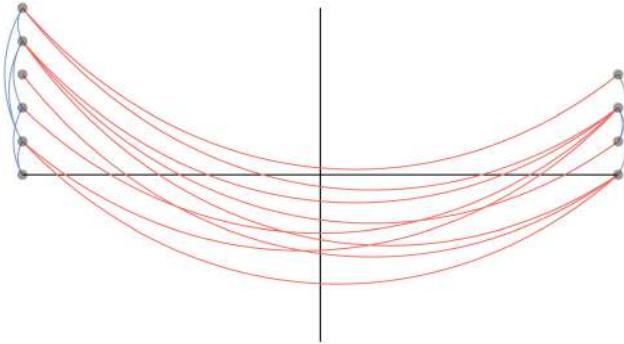


Fig. 2. Visualization by Structural-balance-viz of a balanced network: all the cycles are balanced.

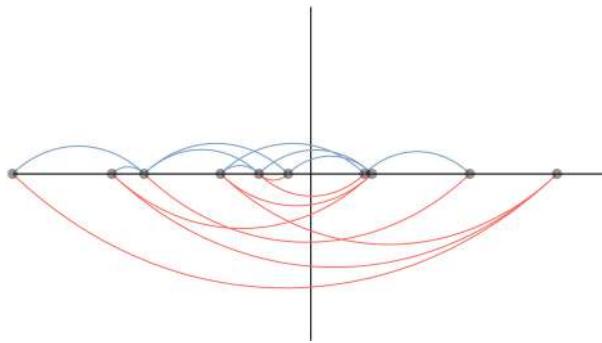


Fig. 3. Visualization by Structural-balance-viz of an unbalanced network: not all the cycles are balanced.

(e.g., cut off edges). To this extent, we include in **Structural-balance-viz** a label reporting the value of λ_m on the top of the y -axis and leave the visualization of λ_m without the label as future work.

The time complexity of **Structural-balance-viz** is governed by the time required by the eigenvalue decomposition of \bar{L} , while the space complexity is $\mathcal{O}(|V|^2)$, again imposed by \bar{L} . Note that computational-intensive network measures μ might considerably extend the running time when drawing large networks.

Figures 2 and 3 show two examples of visualizations generated by **Structural-balance-viz** for a balanced and an unbalanced network, respectively. For such visualizations, we remove the label reporting λ_m to prove how obvious the difference between the two networks is even without textual information. Also, as for all other examples in this paper, edge bundling is not applied; however, it will be available in the tools we plan to publicly release. It is immediate to note that the network represented in Fig. 2 is balanced: all the nodes are at the extremes of the x -axis and no blue (red) edge crosses the y -axis (lays in the same quadrant). This configuration highlights the fact that all the cycles of the represented

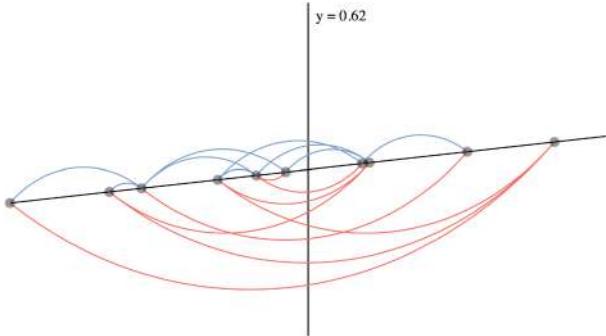


Fig. 4. Visualization by Structural-balance-viz of an unbalanced network with the the additional features. The x -axis scale compares the size of the two factions of nodes.

network are balanced. On the other hand, Fig. 3 shows an unbalanced network since there are positive edges in-between the two factions of nodes and a negative edge within two nodes in the left quadrant; therefore, we easily find the presence of unbalanced cycles.

Figure 4 shows the same network of Fig. 3 with both the additional features of Structural-balance-viz; in this case, the x -axis scale compares the size of the two factions of nodes, i.e., μ counts the number of nodes in the sets. At a glance, it is possible to understand that the left faction is slightly larger than the right one (six and four nodes, respectively) and that the smallest eigenvalue of the signed Laplacian is not far from zero; this means that the network is not far from being balanced (i.e., there are not many unbalanced cycles).

4 Validation and Application

In this section we validate the proposed network layout by visualizing synthetic networks. Also, we apply Structural-balance-viz to derive concrete insights from a dataset representing political debates.

We develop Structural-balance-viz by using D3.js with a Java back-end. The visualization is made available by a web interface that allows the selection of the input dataset and of μ (i.e., the network measure that defines the angular coefficient of the x -axis)¹. The current implementation can consider only the size of the sets of nodes as μ , but the code is easily extendable to consider other characteristics. The time required by our implementation to produce each visualization has always been less than a few seconds.

Validation: Synthetic Networks

We first focus our attention on synthetic-generated networks with the aim of proving that the visualizations produced by Structural-balance-viz are easily comparable. The generative process for signed networks we follow requires in input

¹ Code available at github.com/egalimberti/balance_visualization.

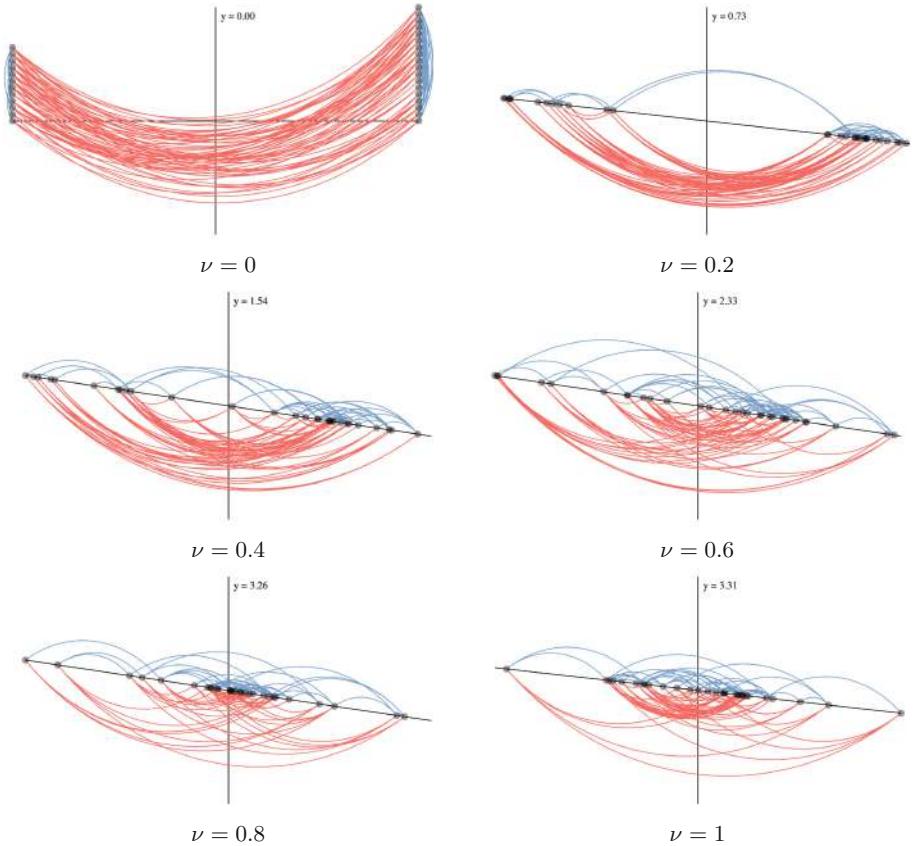


Fig. 5. Visualization by Structural-balance-viz of synthetic networks for increasing values of ν ($n = 30$, $\delta = 0.3$).

three parameters: n indicates the number of nodes, δ defines the edge density, while ν is the ratio of unbalanced triangles in the network (which is another indicator of how much a network is balanced [10]). The procedure works as follows:

- generate a complete balanced network of n nodes (this can be achieved by partitioning the n nodes into two and by assigning negative sign to the edges connecting nodes in different sets while positive sign to all others edges);
- randomly remove edges that do not disconnect the network until the edge density is less or equal than δ ;
- randomly change signs of edges appearing in balanced triangles until the ratio of unbalanced triangles is less or equal than ν .

In Fig. 5 we report our visualization for six networks generated by the described procedure by progressively increasing ν ($\nu \in [0, 0.2, 0.4, 0.6, 0.8, 1]$)

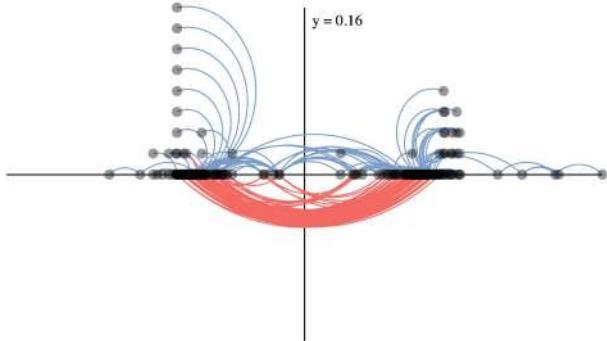


Fig. 6. Visualization by Structural-balance-viz of the United States Congress network.

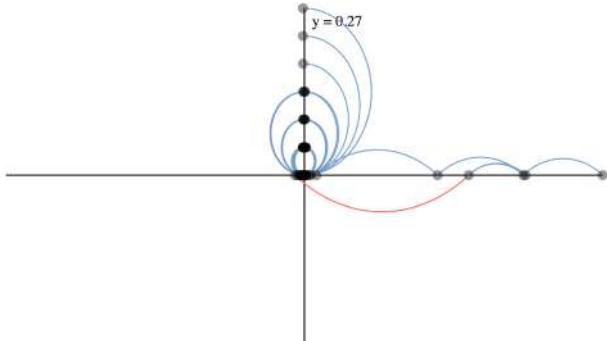


Fig. 7. Visualization by Structural-balance-viz of the United States Congress network after sign reshuffling.

while keeping n and δ fixed ($n = 30$, $\delta = 0.3$). Therefore, we have the full range of networks in terms of structural balance: on one extreme ($\nu = 0$) the network is perfectly balanced, on the other ($\nu = 1$) the network has no balanced triangles. When $\nu = 0$, as expected, we obtain the perfectly distinguishable configuration of balanced networks, where all nodes are in either extremes of the x -axis, no positive edge crosses the y -axis, and no negative edge entirely lies in the same quadrant. Note that, for the balanced case, we do not provide in input to Structural-balance-viz any network function μ since the number of nodes in the sets can be inferred by the height of the two stacks. As ν grows, the most of the nodes gradually moves from the extreme ordinates to the center of the plot; nonetheless, even for $\nu = 1$, we note a few highly-polarized nodes at the margins of the horizontal domain. In addition, more and more both positive edges cross the y -axis and negative edges are within one of the two quadrants. The additional features result to be extremely useful in these cases. At first, the scale gives a precise indication that the right faction is larger than the left one for all values of ν . Also, the smallest eigenvalue of the signed Laplacian, which grows coherently

with ν , eases the comparison of visualizations that might appear similar (e.g., $\nu = 0.8$ and $\nu = 1$) and provides a definitive indication about the structural balance of the visualized networks.

A Case Study: The United States Congress Network

Next, we apply Structural-balance-viz to the analysis of a real-world network obtained from data of the United States Congress modeling a political debate². Nodes ($|V| = 219$) are politicians speaking in the Congress, edges ($|E_+ \cup E_-| = 521$) denote that a speaker mentions another speaker, while signs report whether mentions are in support (positive) or opposition (negative).

Figure 6 shows the visualization of the original Congress network. It is easy to notice that the members are divided into two (almost) equally-sized factions that are close to be balanced; in fact, there is only one negative edge within the left faction and a relatively few positive edges crossing the y -axis. The x -axis can be seen as the left-right political spectrum: the most of the politicians are quite moderate, while there are some polarized members especially in the right, and a few nodes close to $x = 0$ (probably the mediators between the two factions).

To have a better understanding of the structural balance of the Congress network, we compare it to a null model. In particular, we maintain the same network structure while reshuffling the edge signs, leaving the number of positive and negative edges unchanged. The visualization of the resulting reshuffled network is reported in Fig. 7. In this case, the balance/polarization structure of the network is destroyed since the majority of the nodes collapse close to the origin. All the negative edges (except one) lay between such nodes and are no more visible in the layout. Only five members maintain their polarization in the right. Moreover, the smallest eigenvalue of the signed Laplacian is greater than in the original network. All this indications suggests that, the United States Congress network is more balanced/polarized than what is expected by chance, according to a reshuffled null model. The Congress is instead quite polarized, very close to being structurally balanced, due to the political parties and alliances.

5 Conclusions

In this paper we introduce Structural-balance-viz: a novel algorithm that places nodes in a Cartesian coordinate system, that resembles the behavior of a scale, and exploits edge coloring and bundling for showing whether a connected signed network is balance or unbalanced and, in the latter case, how far it is from being balanced. Structural-balance-viz is validated by the analysis of synthetic networks: it is proved to provide an indication of balance/polarization of the whole network and individually of each node, to identify two factions of nodes on the basis of their polarization and show their cumulative characteristics, and to produce reproducible and easily comparable visualizations. An application to a real-world dataset about political debates confirms that Structural-balance-viz can provide meaningful insights about the polarization structure of the network.

² Dataset available at konect.cc.

As future work, we plan to devote more effort in embedding the value of the smallest eigenvalue of the signed Laplacian in **Structural-balance-viz** without textual supplement. Moreover, we want to deploy our implementation, including edge bundling, to a public web interface and make it available for network visualization tools, e.g., Cytoscape. Finally, we will employ **Structural-balance-viz** for future analysis of real-world signed networks.

References

1. Akiyama, J., Avis, D., Chvátal, V., Era, H.: Balancing signed graphs. *Discrete Appl. Math.* **3**(4), 227–233 (1981)
2. Aref, S., Mason, A.J., Wilson, M.C.: Computing the line index of balance using integer programming optimisation. In: Goldengorin, B. (ed.) *Optimization Problems in Graph Theory*, pp. 65–84. Springer, Cham (2018)
3. Aref, S., Wilson, M.C.: Measuring partial balance in signed networks. *J. Complex Netw.* **6**(4), 566–595 (2017)
4. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. *Mach. Learn.* **56**(1–3), 89–113 (2004)
5. Bonchi, F., Galimberti, E., Gionis, A., Ordozgoiti, B., Ruffo, G.: Discovering polarized communities in signed networks. In: *Proceedings of the 2019 ACM on Conference on Information and Knowledge Management*. ACM (2019)
6. Cartwright, D., Harary, F.: Structural balance: a generalization of heider's theory. *Psychol. Rev.* **63**(5), 277 (1956)
7. Coleman, T., Saunderson, J., Wirth, A.: A local-search 2-approximation for 2-correlation-clustering. In: *European Symposium on Algorithms*, pp. 308–319 (2008)
8. Coleman, T., Saunderson, J., Wirth, A.: Spectral clustering with inconsistent advice. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 152–159. ACM (2008)
9. Doreian, P., Mrvar, A.: Structural balance and signed international relations. *J. Soc. Struct.* **16**, 1 (2015)
10. Easley, D., Kleinberg, J.: Positive and negative relationships. In: *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, Cambridge University Press (2010)
11. Harary, F.: On the notion of balance of a signed graph. *Mich. Math. J.* **2**(2), 143–146 (1953)
12. Harary, F., Kabell, J.A.: A simple algorithm to detect balance in signed graphs. *Math. Soc. Sci.* **1**(1), 131–136 (1980)
13. Hou, Y.P.: Bounds for the least Laplacian eigenvalue of a signed graph. *Acta Mathematica Sinica* **21**(4), 955–960 (2005)
14. Hou, Y., Li, J., Pan, Y.: On the Laplacian eigenvalues of signed graphs. *Linear and Multilinear Algebra* **51**(1), 21–30 (2003)
15. Kaufmann, M., Wagner, D.: Drawing graphs: methods and models, vol. 2025. Springer (2003)
16. Krzywinski, M., Birol, I., Jones, S.J.M., Marra, M.A.: Hive plots—rational approach to visualizing networks. *Briefings in Bioinform.* **13**(5), 627–644 (2011)
17. Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., De Luca, E.W., Albayrak, S.: Spectral analysis of signed graphs for clustering, prediction and visualization. In: *Proceedings of the 2010 SIAM International Conference on Data Mining*, pp. 559–570. SIAM (2010)

18. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Signed networks in social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1361–1370. ACM (2010)
19. Ordozgoiti, B., Gionis, A.: Reconciliation k-median: clustering with non-polarized representatives. In: The World Wide Web Conference, pp. 1387–1397. ACM (2019)
20. Redner, S.: Social balance on networks: the dynamics of friendship and hatred. In: APS Meeting Abstracts (2006)
21. Tang, J., Chang, Y., Aggarwal, C., Liu, H.: A survey of signed network mining in social media. ACM Comput. Surv. (CSUR) **49**(3), 42 (2016)
22. Xia, W., Cao, M., Johansson, K.H.: Structural balance and opinion separation in trust-mistrust social networks. IEEE Trans. Control Netw. Syst. **3**(1), 46–56 (2015)



Spheres of Legislation: Polarization and Most Influential Nodes in Behavioral Context

Andrew C. Phillips, Mohammad T. Irfan^(✉), and Luca Ostertag-Hill

Bowdoin College, Brunswick, ME 04011, USA

andrewphillips182@gmail.com, {mirfan,ldostert}@bowdoin.edu
<http://www.bowdoin.edu/~mirfan>

Abstract. Game-theoretic models of influence in networks often assume the network structure to be static. In this paper, we allow the network structure to vary according to the underlying behavioral context. This leads to several interesting questions on two fronts. First, how do we identify different contexts and learn the corresponding network structures using real-world data? We focus on the U.S. Senate and apply unsupervised machine learning techniques, such as fuzzy clustering algorithms and generative models, to identify different spheres of legislation as context and learn an influence network for each sphere. Second, how do we analyze these networks in order to gain an insight into the role played by the spheres of legislation in various interesting constructs like polarization and most influential nodes? To this end, we apply both game-theoretic and social network analysis techniques. In particular, we show that game-theoretic notion of most influential nodes brings out the strategic aspects of interactions like bipartisan grouping, which typical centrality measures fail to capture. We also show that for the same set of senators, some spheres of legislation are more polarizing than others.

Keywords: Influence in networks · Machine learning and networks · Computational game theory · Graphical games

1 Introduction

In recent times, the study of social influence has extended beyond mathematical sociology [12, 31] and has entered the realm of computation [1, 3–5, 16, 18, 21–24]. A computational study of “influence”—however we define it—is key to understanding the behavior of individuals embedded in networks. In this paper, we model and analyze social influence in a strategic setting where one’s behavior depends on others’ behavior. Since game theory reliably captures such interdependence of behavior in a population, we ground our computational approach

ACP worked on this research as an undergraduate student at Bowdoin. MTI is the corresponding author. LOH is currently an undergraduate student at Bowdoin.

in game theory. The strategic setting of our interest here is the U.S. Senate. We model the influence structure among the senators by taking into account the relevant context, which we call the spheres of legislation. We learn these models of influence from the real-world behavioral data on Senate bills and voting records. Our particular focus is on analyzing machine learned influence networks to answer various questions on polarization and most influential nodes.

Interestingly, most computational models of influence assume a fixed network structure among individuals. We relax this simplifying assumption, allowing the network of influence to vary according to the spheres of legislation. For example, bills on finance may induce a very different influence network among senators than bills on defense, which may in turn have different impacts on inference problems like polarization and most influential nodes. One central question in this regard is: How do we identify different spheres of legislation that may have different implications on these inference problems? We address this in Sect. 2.

After identifying spheres of legislation, we can learn an influence network among the senators for each sphere by adopting game-theoretic models of strategic behavior. Broadly speaking, modeling and analyzing congressional voting behavior has been a trending topic in both political science and computer science [6, 10, 16, 18, 29], in part due to the availability of data. In particular, we use the linear influence game (LIG) model of strategic behavior proposed by Irfan and Ortiz [17, 18] and its recent extension [16]. We learn these models using data from the spheres of legislation. In LIG, each senator exerts influence upon (and is subject to influences from) other senators in a network-structured way. The model focuses on interdependence among the senators and adopts the game-theoretic solution concept of *Nash equilibrium* to predict stable outcomes from a complex system of influences. This notion of Nash equilibrium leads to a definition of the most influential senators, where a group of senators is called most influential with respect to a desirable outcome if their support for that outcome influences enough other individuals to achieve that outcome. The LIG model will be elaborated in Sect. 3 and machine learning of this model using the spheres of legislation will be detailed in Sect. 4.

While game-theoretic *prediction* of congressional votes has been well studied using the LIG model and its extensions [16–18], an *analysis* of the machine learned networks of influence did not get much attention, which we address here. Similarly, algorithms for computing most influential nodes in a strategic setting have been studied before (e.g., [18]), but their structural analogs like centrality measures have not been explored in a comparative fashion. In other words, what do we gain by using a game-theoretic definition of most influential nodes as opposed to a structural definition? We address questions like this.

Furthermore, polarization in social networks has been well studied [1, 13, 27], especially in the political arena [7, 9, 26, 32, 33]. Three salient points distinguish our approach from the rich body of literature: (1) Ours is a model-based approach, where networks are central to predicting collective outcomes, (2) we learn the networks using behavioral data, because the networks are not observable, and (3) we seek to show that polarization in Senate varies according to the spheres of legislation. We do not touch on the rising polarization in Senate over time, which by now is a well-settled matter [8].

The past two terms of Congress are especially interesting for analyzing network behavior and polarization. The 114th Congress ran from January 2015 to January 2017 and the 115th Congress ran from January 2017 to January 2019. In both terms, Republicans controlled the Senate, but the executive power was different. In the 114th Congress, Barack Obama (D) held the presidency; in the 115th, Donald Trump (R) held the presidency. Despite the two opposing parties holding presidency, both terms are perceived to be deeply polarized. Interestingly, when we study different influence networks *among the same group of senators* arising from different spheres of legislation we find that polarization is not really equally applicable. It very much depends on the sphere under consideration. Our aim is to put polarization and other inference questions like most influential nodes in context. Before continuing, we note that all supplementary materials, including detailed literature reviews, visualizations, and technical details, are included in the Appendix.

2 Spheres of Legislation

We use an unsupervised machine learning technique, namely fuzzy clustering, to assign bills to different spheres of legislation based on the bill subjects. The clustering algorithm uses data obtained from the @unitedstates project (<https://github.com/unitedstates/congress>). In particular, we use bill data and roll-call data. The latter contains each senator’s “yea,” “nay,” or abstaining votes, while bill data includes a list of subjects incident to the bill, among other attributes. There are 820 subjects ranging from “Abortion” to “Zimbabwe,” and multiple subjects describe each bill. Additionally, each bill is assigned a single “top term,” the broad subject which best describes the bill out of 23 possible top-level subjects. We use the roll-call data to learn strategic interactions and bill data to extract bill topics. For the 114th and 115th Senates, we have a total of 103 senators and 722 bills (details are in the Appendix).

We seek to split the bills into a small number of broad categories, each of which encompasses many bills. On their own, the “top terms” are too specific to be used as clusters of their own. Making each top term its own cluster would result in some clusters containing only one bill and others containing a hundred. Due to the exponential “outcome space” of LIGs, learning LIGs requires a relatively large amount of data. Therefore, small clusters would be unusable.

Rather than manually re-categorizing bills, we took a statistical clustering approach to grouping, based on a bill’s assigned “top term” in addition to all subjects it contains. For each data point, we assigned each possible subject a weight: 0 if missing, 1 if present, or 10 if it is the “top term.” By including both measures of subjects (top and regular), we produce more meaningful categories than using top terms or bill subjects lists alone.

In data science, K-Means (KM) is often used as a simple yet effective clustering algorithm [25]. Cluster membership in KM is crisp, meaning each data point belongs to one and only one cluster. While effective at producing distinct clusters, KM is not ideal for our purposes because bills often belong to multiple

clusters. For example, a bill about increasing defense spending is about national security as well as economics. The Fuzzy C-Means (FCM) clustering algorithm addresses this problem. FCM is an extension of KM which allows for overlaps in clusters [2, 30]. The objective function in FCM is largely the same as in KM, with the addition of membership values w_{ij} and a fuzzifier m . Membership values describe how closely each data point i belongs to cluster j . The fuzzifier changes membership values: $m = 1$ results in crisp clusters ($w_{ij} \in \{0, 1\}$), and higher values of m result in fuzzier clusters.

Iterating over a range of values, we found that number of clusters, $c = 4$ and $m = 1.3$ resulted in clusters which were relatively distinct, had intuitive descriptions and also contained an adequate number of bills for machine learning. Additionally, we experimented with the threshold values for cluster membership and settled on 0.15. That is, a bill is considered a member of a cluster if its membership value is above 0.15. Table 1 describes the results of our chosen FCM parameters. Each cluster is assigned a shorthand name describing its contents and is called a sphere of legislation in this paper. We next describe the model.

Table 1. Shorthand names and descriptions for each of the spheres of legislation identified by the FCM algorithm. Spheres 1 and 2 are relatively distinct from the rest, while Spheres 3 and 4 share a large number of bills.

Sphere#	Size	Name of sphere	Sampling of bill subjects	Ovlp. 1	Ovlp. 2	Ovlp. 3	Ovlp. 4
1	105	Security & Armed Forces	Armed forces and national security (77), Emergency management (11), Transportation and public works (10)	7%	20%	20%	
2	263	Economics & Finance	Economics and public finance (263)	3%	0%	0%	
3	284	Energy & Infrastructure	Energy (69), Education (31), Taxation (28), Transportation and public works (27)	7%	0%		76%
4	313	Public Welfare	Health (52), Crime and law enforcement (43), Taxation (38), Education (31)	7%	0%	69%	

3 The LIG Model

We represent the senate influence network as a linear influence game (LIG) [17, 18], one type of 2-action graphical game [20]. Nodes represent senators, or *players*, and are connected by directed edges. Edge weights represent the influence exerted by the source node upon the target. Influence weights can be negative, positive, or zero. The directed edges are allowed to be asymmetric, meaning nodes A and B may exert different levels of influences on each other. Additionally, nodes have a threshold level, which represents “stubbornness.” Nodes with thresholds further from zero are more resistant to change. Absent influences, a node with negative threshold is predisposed to adopting action +1 (*yea* vote),

and a node with positive threshold is predisposed to -1 (*nay* vote). The matrix of influence weights $\mathbf{W} \in \mathbf{R}^{n \times n}$ and the threshold vector $\mathbf{b} \in \mathbf{R}^n$ constitute the LIG model. The action $x_i \in \{+1, -1\}$ chosen by each node i is the outcome of the model, as described below in game-theoretic terms.

Each node's *best response* to other nodes' actions depends on the net incoming influence and the node's threshold. When the total incoming influence from nodes playing $+1$ minus the total incoming influence from nodes playing -1 exceeds the node's threshold level, that node's best response is $+1$. If below, it is -1 ; in the case of a tie, the node is indifferent and can play either. Note that the best responses of the nodes are interdependent. A vector of *mutual best responses* of all the nodes is a stable outcome of the model, formally known as a *Nash equilibrium*. It is stable because no node has any incentive to deviate from it. The LIG model adopts Nash equilibria to represent stable collective outcomes from a complex network of influence. Below is a formal description, using the same notation as [18]. An example is provided in the Appendix.

Definition 1 (Linear Influence Game (LIG) [18]). *In LIG, the influence function of each individual i , given others' actions \mathbf{x}_{-i} , is defined as $f_i(\mathbf{x}_{-i}) \equiv \sum_{j \neq i} w_{ij}x_j - b_i$ where for any other individual j , $w_{ij} \in \mathbb{R}$ is a weight parameter quantifying the "influence factor" that j has on i , and $b_i \in \mathbb{R}$ is a threshold parameter for i 's level of "tolerance."*

Here, individuals receive influences from other players and have an influence threshold of their own, which accounts for their own resistance to external influence. The influence function f_i calculates the weighted sum of incoming influences on i , as described in the paragraph above Definition 1, and subtracts i 's threshold from it. The payoff of each player is defined next.

Definition 2 (Payoff Function [18]). *For an LIG, we define the payoff function $u_i : \{-1, 1\}^n \rightarrow \mathbb{R}$ as $u_i(x_i, \mathbf{x}_{-i}) \equiv x_i f_i(\mathbf{x}_{-i})$, where \mathbf{x}_{-i} denotes the vector of a joint-action of all players except i and f_i is defined in Definition 1.*

The payoff function quantifies the preferences of the players based on the actions of other players. Given the action of all other individuals \mathbf{x}_{-i} and influence function $f_i(\mathbf{x}_{-i})$, an individual will prefer to choose either $+1$ or -1 as follows. When $f_i(\mathbf{x}_{-i})$ is negative, $x_i = -1$ will result in a positive payoff; when $f_i(\mathbf{x}_{-i})$ is positive, $x_i = +1$ will result in a positive payoff. Actions chosen in this fashion in order to result in a positive payoff (i.e., to maximize payoff) is defined as the *best response*. When everyone is playing their best responses simultaneously, we get a pure-strategy Nash Equilibrium (PSNE) as defined below.

Definition 3 (Pure-Strategy Nash Equilibrium [18]). *A pure-strategy Nash equilibrium (PSNE) of an LIG \mathcal{G} is an action assignment $\mathbf{x}^* \in \{-1, 1\}^n$ that satisfies the following condition. Every player i 's action x_i^* is a simultaneous best-response to the actions \mathbf{x}_{-i}^* of the rest.*

We adopt PSNE as the notion of stable outcomes arising from a network of influence. We are interested in questions like how the network changes based on the spheres of legislation and what impact the spheres have on polarization and most influential nodes. For these, we learn the networks using the spheres data.

4 Machine Learning

We use Honorio and Ortiz's machine learning algorithm in order to instantiate an LIG from raw roll-call data [15]. The goal of the algorithm is to capture as much of the ground-truth data as possible as PSNE (the *empirical proportion of equilibria*), without having so many total PSNE (the *true proportion of equilibria*) that the model is meaningless. For example, if all influence weights and threshold levels are 0 ($\mathbf{W} = 0$, $\mathbf{b} = 0$), then all 2^n possible joint actions among n players would be PSNE, trivially covering all observed voting data. However, this is undesirable as it has no predictive power at all. Therefore, we would like to maximize the empirical proportion of equilibria while minimizing the true proportion. To balance the true and empirical proportions of equilibria, the learning algorithm uses a generative mixture model that picks a joint action which is either a PSNE or non-PSNE with probability q and $1 - q$, respectively. Maximizing the empirical proportion of equilibria relative to the true proportion can be framed as a maximum likelihood estimation problem in this generative model. Under mild conditions, the final optimization problem is the following: $\min_{\mathbf{w}, \mathbf{b}} \frac{1}{m} \sum_l \max_i \ell[x_i^{(l)}(\mathbf{w}_{i,-i}^T \mathbf{x}_{-i}^{(l)} - b_i)] + \rho \|\mathbf{w}\|_1$. Here, m is the number of bills, ℓ is the typical logistic loss function, and ρ is an l_1 regularization parameter controlling the number of edges $\|\mathbf{w}\|_1$. This is a gist of Honorio and Ortiz's machine learning algorithm resulting from a very lengthy proof [15].

We solve the above optimization for each sphere of legislation and obtain an influence network. While doing this, we rigorously cross validate to avoid overfitting or underfitting as follows. In the model selection phase, we wish to choose an “appropriate” value of the l_1 regularization parameter ρ . Since ρ penalizes the number of edges, high values of ρ result in sparser graphs at the risk of underfitting, and low values of ρ lead to denser graphs at the risk of overfitting. The number of edges is important, because the problem of computing equilibria is NP-hard [17, 18], and an extremely complex model would have so many edges that equilibrium computation would not finish within days. We use 10-fold cross-validation, track multiple metrics, and choose $\rho = 0.002728, 0.003888, 0.003070, 0.003888$ for the four spheres, respectively. Details are in the Appendix.

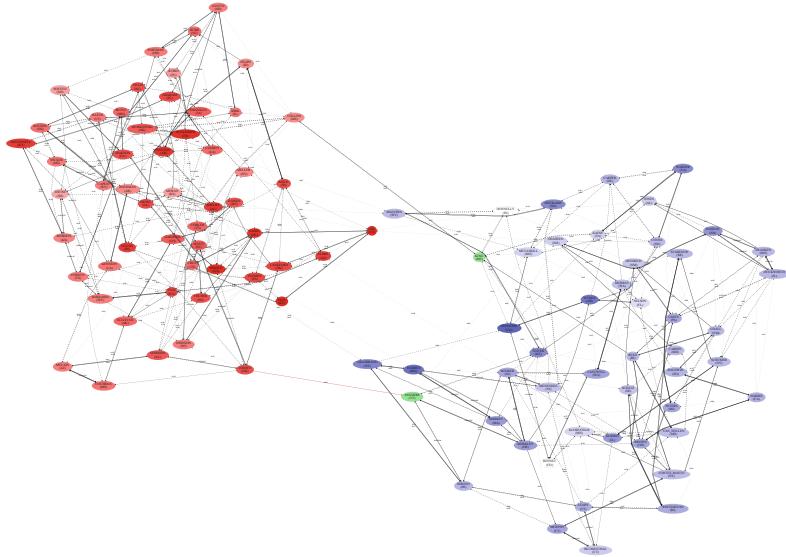


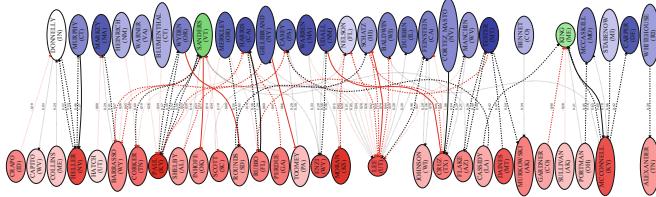
Fig. 1. A bird's eye view of the LIG network for Sphere 2 (Economics & Finance). Red nodes are Republicans, blue Democrats, green Independents. Darker nodes have higher threshold and thicker edges have more influence weights. The strongest 40% incoming and outgoing edges for each node are shown.

5 Polarization in Context

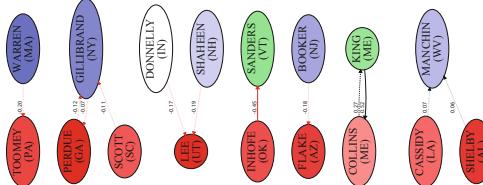
Visualization of the machine learned networks clearly shows that the network structure varies according to the spheres of legislation. In all spheres, however, the force-directed drawing algorithm automatically distinguishes Republicans from Democrats. Figure 1 depicts this LIG visualization for Sphere 2 (Economics & Finance) as a representative example. The visualizations for the remaining spheres can be found in the Appendix.

The boundary between the two parties is interesting for studying polarization. Even though negative edges more often occur at the boundary, the connectivity between the two parties varies a lot according to the spheres of legislation. These are depicted in Fig. 2 for Spheres 1 and 2 (others are in the Appendix).

Figure 2b shows the cross-border edges in Sphere 2 (Economics & Finance), which starkly contrasts those of Sphere 1 (Security & Armed Forces). In Sphere 2, only 12 of the strongest 40% of edges are between members of different parties. Of these, 2/3 are negative, suggesting a very polarized network. Aside from two positive influences between Maine senators King (a left-leaning Independent) and Collins (a center-leaning Republican), the remaining two positive connections are the weakest of all connections shown for this sphere.



(a) Sphere 1 (Security & Armed Forces): 52 boundary edges are positive, 37 negative.



(b) Sphere 2 (Economics & Finance): 4 cross-border edges are positive, 8 negative.

Fig. 2. Graphviz visualization of cross-border edges connecting members of the opposing parties within the strongest 40% of all edges.

Similarly, examining inter-party edges reveals that Sphere 3 (Energy & Infrastructure) is also very polarized. While there are many edges between both parties in this network, about 70% of them are negative. Positive influences come from a few sources, again including the centrist Senator Collins. Incongruously, prominent right-wing senator Tom Cotton (R-AR) also exhibits positive influences with democratic senators. However, most other far-left or far-right leaning senators, including Sanders (I-VT) and Cruz (R-TX) only exhibit negative influences with the opposite party.

Sphere 4 (Public Welfare)'s inter-party edges strike a balance between the polarities exhibited by the previous three spheres. There are slightly more positive edges (9) than negative edges (7), but still a low number of edges overall. Again, there are positive influences between Maine senators King (I-ME) and Collins (R-ME), but also positive influences between Senator McConnell and Democratic senators King (D-ME) and Tester (D-MT).

Overall, each sphere exhibits some level of polarization, but influences within some spheres are far less polarizing than others. Some senators are present in every sphere's inter-party boundary, whether for positive or negative influences. Senators Collins and King often share positive influences with each other, as well as other senators. Senator Lee (R-UT), a conservative libertarian, always exhibits negative edges with members of the other party, although in Sphere 1, he also shares positive influences with senators Harris (D-CA) and Feinstein (D-CA). Meanwhile, left-wing icon Bernie Sanders (I-VT) exhibits the equivalent behavior, with only negative cross-border edges in all spheres *except* Sphere 1. These results suggest that Sphere 1 (Security & Armed Forces) is least polarized, whereas Spheres 2 (Economics & Finance) is highly polarized.

Furthermore, a formal study of polarization rooted in network science produces similar results. Modularity [11, 27, 28] has been widely used as a measure of polarization in networks. We apply the following definition of modularity derived for directed networks with signed weights [14].

$$Q = \frac{1}{2w^+ + 2w^-} \sum_i \sum_j \left[w_{ij} - \left(\frac{w_i^{+,out} w_j^{+,in}}{2w^+} - \frac{w_i^{-,out} w_j^{-,in}}{2w^-} \right) \right] \times \delta(C_i, C_j).$$

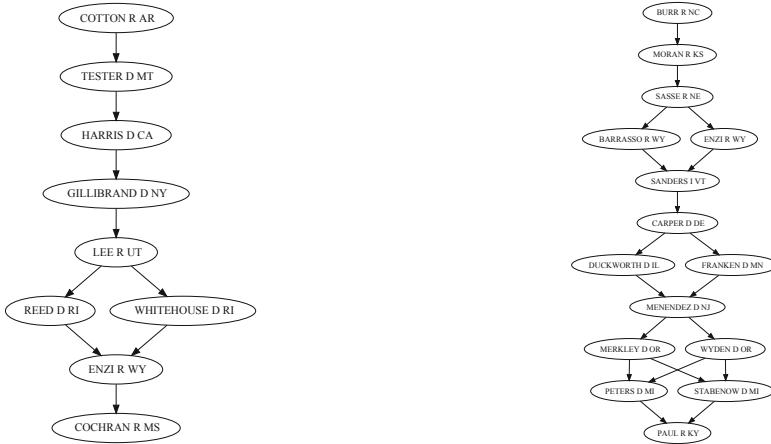
Here, w_{ij} is the weight of edge i to j , $w_{ij}^+ = \max\{0, w_{ij}\}$, $w_{ij}^- = \max\{0, -w_{ij}\}$, and $2w^\pm$ is the total weight of all positive or negative edges, expressed by $\sum_i \sum_j w_{ij}^\pm$. Furthermore, $w_i^{\pm,out}$ is the weighted out-degree $\sum_k w_{ik}^\pm$ and $w_j^{\pm,in}$ is the weighted in-degree $\sum_k w_{kj}^\pm$. The Kronecker delta function $\delta(C_i, C_j)$ is 1 if i and j belong to the same party; it is 0 otherwise.

Applying this definition, We obtain the following modularity scores for the four spheres of legislation respectively: 0.7861, 0.8904, 0.8724, and 0.8857. This shows that Sphere 1 (Security & Armed Forces) is least polarized and Spheres 2 (Economics & Finance), 3 (Energy & Infrastructure), and 4 (Public Welfare) are much more polarized.

6 Most Influential Nodes in Context

There exists a number of centrality measures that are derived from a structural analysis of networks [19]. However, our model is behavioral where nodes adopt their best responses to each other. In a strictly game-theoretic model of behavior, a set of nodes will be called most influential *with respect to achieving a desirable stable outcome* if their choice of actions leads the whole system of influence to that desirable outcome [17, 18]. Here, a crucial aspect is a desirable stable outcome, represented by a PSNE. For example, let us say that our desirable outcome is to pass a bill by a 100-0 vote. A set of senators will be called most influential if their voting together influences every other senator to also vote for the bill, thereby having the desirable outcome as the *unique* PSNE outcome. This concept can be extended to other types of desirable outcomes like passing a bill with at least 60 votes, forcing/avoiding a filibuster, etc.

An approximation algorithm for computing most influential senators was given by Irfan and Ortiz [18], which produces a directed acyclic graph (DAG). The algorithm requires precomputation of all PSNE, which is a provably hard problem [18]. We apply Irfan and Ortiz's PSNE computation algorithm to the LIG for each sphere of legislation. We elaborate this in the Appendix. Having computed all the PSNE, we then compute the DAG representing most influential sets of nodes. Figure 3 shows the results of the most influential nodes algorithm for Spheres 1 and 3, where the desirable outcome is set as passing a bill unanimously. The way to read Fig. 3 is to inspect each DAG and find a top to bottom path. Each of these paths gives a most influential set.



(a) Sphere 1 (Security & Armed Forces): 4 Republicans and 4 Democrats are most influential.

(b) Sphere 3 (Energy & Infrastructure): 5 Republicans and 6 Democrats are most influential.

Fig. 3. Directed acyclic graphs (DAGs) representing sets of most influential nodes for Spheres 1 and 3. Any top to bottom path gives a most influential set.

The sets of most influential senators in each sphere support the inferences gained from analyzing the LIG networks. As illustrated in Fig. 3, in Sphere 1 (Security & Armed Forces), 4 Republicans and 4 Democrats comprise a set of 8 most influential senators. In other words, 8 senators are sufficient to guarantee enough support that a bill will pass unanimously. In Sphere 3 (Energy & Infrastructure), 5 Republicans and 6 Democrats comprise a set of 11. This suggests that Sphere 3 is more polarized than Sphere 1, since it requires a larger body of influencing senators. The DAGs for the other spheres are in the Appendix.

Game-Theoretic vs. Structural Centrality Measures. In the above game-theoretic formulation of most influential nodes, we find that each set of most influential senators across all spheres is comprised of an (almost) equal number of Democrats and Republicans. This signifies the need for bipartisan support to guarantee passing a bill unanimously. As we show next, this also happens to be a distinguishing feature between game-theoretic and structural measures. Table 2 shows various centrality measures and other quantities computed for each sphere. For each sphere, we show the top 10 most central senators with respect to four centrality measures: degree, closeness, betweenness, and eigenvector. Most notably, these centrality measures do not capture the strategic aspects of behavior. Throughout most measures, Republican senators are overrepresented, comprising the majority of the top ten most central nodes. In contrast, the game-theoretic measure gives a balanced coalition between Democrats and Republicans. This is important because when networks are polarized, achieving a desirable outcome requires support from both sides.

Table 2. Network analysis of learned influence networks for different spheres of legislation. Various centrality measures and network-level properties are shown.

	Sphere 1	Sphere 2	Sphere 3	Sphere 4
Number of Edges	1191	1071	1280	1076
Network Diameter	5	4	5	5
Modularity	0.7861	0.8904	0.8724	0.8857
Avg. (Shortest) Path Length	2.2295	2.5132	2.1506	2.5476
Avg. Clustering Coefficient	0.1867	0.2057	0.174	0.2176
Degree Centrality				
Degree (1) 0.5784: LEE R-UT	0.3529: TOOMEY R-PA	0.3725: LANKFORD R-OK	0.3725: COTTON R-AR	
Degree (2) 0.4216: PAUL R-KY	0.3333: PERDUE R-GA	0.3627: SASSE R-NE	0.2941: LEAHY D-VT	
Degree (3) 0.4118: SANDERS I-VT	0.3235: ENZI R-WY	0.3529: WARNEE D-VA	0.2843: CAPITO R-WV	
Degree (4) 0.3824: MORAN R-KS	0.3137: LANKFORD R-OK	0.3333: CAPITO R-WV	0.2843: MURKOWSKI R-AK	
Degree (5) 0.3725: MANCHIN D-WV	0.3137: YOUNG R-IN	0.3235: TOOMEY R-PA	0.2745: AYOTTE R-NH	
Degree (6) 0.3627: RUBIO R-FL	0.3039: COTTON R-AR	0.3235: MURKOWSKI R-AK	0.2745: SHELBY R-AL	
Degree (7) 0.3529: CRUZ R-TX	0.2941: CASEY D-PA	0.3235: COTTON R-AR	0.2647: PERDUE R-GA	
Degree (8) 0.3333: ALEXANDER R-TN	0.2843: CASSIDY R-LA	0.3137: BROWN D-OH	0.2549: ALEXANDER R-TN	
Degree (9) 0.3333: ENZI R-WY	0.2843: WICKER R-MS	0.3137: FEINSTEIN D-CA	0.2549: PETERS D-MI	
Degree (10) 0.3235: LEAHY D-VT	0.2745: CORKEE R-TN	0.3137: PAUL R-KY	0.2549: PAUL R-KY	
Closeness Centrality				
Closeness (1) 0.5862: LEE R-UT	0.5126: PERDUE R-GA	0.5514: WARNER D-VA	0.5204: COTTON R-AR	
Closeness (2) 0.5635: RUBIO R-FL	0.4951: COTTON R-AR	0.5426: LANKFORD R-OK	0.5178: KIRK R-IL	
Closeness (3) 0.5574: PAUL R-KY	0.4766: COLLINS R-ME	0.5368: SASSE R-NE	0.4951: MURKOWSKI R-AK	
Closeness (4) 0.5455: SANDERS I-VT	0.47: ENZI R-WY	0.5368: BENNET D-CO	0.4928: AYOTTE R-NH	
Closeness (5) 0.5455: BALDWIN D-WI	0.4636: SASSE R-NE	0.534: KING I-ME	0.4766: SULLIVAN R-AK	
Closeness (6) 0.5397: ENZI R-WY	0.4636: MANCHIN D-WV	0.5231: MURKOWSKI R-AK	0.4744: MCCONNELL R-KY	
Closeness (7) 0.5368: MORAN R-KS	0.4615: FLAKE R-AZ	0.5178: COTTON R-AR	0.4722: PAUL R-KY	
Closeness (8) 0.5285: CORKER R-TN	0.4595: SHELBY R-AL	0.5152: BROWN D-OH	0.4636: COLLINS R-ME	
Closeness (9) 0.5258: CASEY D-PA	0.4595: YOUNG R-IN	0.5126: CASEY D-PA	0.4554: CAPITO R-WV	
Closeness (10) 0.5231: DURBIN D-IL	0.4595: HEITKAMP D-ND	0.51: CARPER D-DE	0.4554: SANDERS I-VT	
Betweenness Centrality				
Betweenness (1) 0.0696: LEE R-UT	0.0538: PERDUE R-GA	0.0278: LANKFORD R-OK	0.0685: SANDERS I-VT	
Betweenness (2) 0.0362: PERDUE R-GA	0.0468: HEITKAMP D-ND	0.0272: SASSE R-NE	0.0641: WYDEN D-OR	
Betweenness (3) 0.0314: MORAN R-KS	0.0452: GILLIBRAND D-NY	0.0265: WARNEE D-VA	0.0559: COTTON R-AR	
Betweenness (4) 0.0314: KING I-ME	0.045: ENZI R-WY	0.0226: BENNET D-CO	0.0533: MARKEY D-MA	
Betweenness (5) 0.0301: MANCHIN D-WV	0.0434: COLLINS R-ME	0.0206: MURKOWSKI R-AK	0.0532: ALEXANDER R-TN	
Betweenness (6) 0.0288: DURBIN D-IL	0.0383: MERKLEY D-OR	0.0196: BROWN D-OH	0.0421: MURKOWSKI R-AK	
Betweenness (7) 0.0283: RUBIO R-FL	0.0382: COTTON R-AR	0.0194: CORNYN R-TX	0.0381: PAUL R-KY	
Betweenness (8) 0.0283: PAUL R-KY	0.0381: SANDERS I-VT	0.0193: SCHATZ D-HI	0.0376: SASSE R-NE	
Betweenness (9) 0.0253: SANDERS I-VT	0.0344: LEE R-UT	0.0187: SANDERS I-VT	0.0347: HARRIS D-CA	
Betweenness (10) 0.0249: CRUZ R-TX	0.034: TESTER D-MT	0.0185: WICKER R-MS	0.032: AYOTTE R-NH	
Eigenvector Centrality				
Eigenvector (1) 0.271: LEE R-UT	0.2114: COTTON R-AR	0.181: WARNER D-VA	0.2029: KIRK R-IL	
Eigenvector (2) 0.2291: SANDERS I-VT	0.2061: PERDUE R-GA	0.1703: CORNYN R-TX	0.2017: HOEVEN R-ND	
Eigenvector (3) 0.228: PAUL R-KY	0.1866: SULLIVAN R-AK	0.1675: LANKFORD R-OK	0.1727: GARDNER R-CO	
Eigenvector (4) 0.1894: BALDWIN D-WI	0.1865: ENZI R-WY	0.1567: BENNET D-CO	0.1724: PORTMAN R-OH	
Eigenvector (5) 0.1892: RUBIO R-FL	0.183: YOUNG R-IN	0.1551: JOHNSON R-WI	0.1715: CAPITO R-WV	
Eigenvector (6) 0.1696: ENZI R-WY	0.1758: THUNE R-SD	0.1541: SASSE R-NE	0.1706: COTTON R-AR	
Eigenvector (7) 0.1681: BARRASSO R-WY	0.1734: WICKER R-MS	0.1525: MURKOWSKI R-AK	0.1706: ROBERTS R-KS	
Eigenvector (8) 0.161: CASEY D-PA	0.1674: MORAN R-KS	0.1454: KING I-ME	0.1683: FISCHER R-NE	
Eigenvector (9) 0.1607: MORAN R-KS	0.1653: JOHNSON R-WI	0.1426: COTTON R-AR	0.1682: MURKOWSKI R-AK	
Eigenvector (10) 0.1602: MANCHIN D-WV	0.163: GARDNER R-CO	0.1379: BOOZMAN R-AR	0.1646: ISAKSON R-GA	

7 Richer Models: Ideal Point Models with Social Interactions

We also apply a richer model of influence recently proposed by Irfan and Gordon [16] that extends the LIG model by incorporating ideal points of the senators and polarities of bills [29]. Their work showed the value of combining game-theoretic and statistical models for studying strategic interactions in context, but they assume the network to be fixed, regardless of the bill context. We use their model and allow the network to change based on the spheres of legislation. We also perform an analysis of the networks learned.

As a cautionary note, the way Irfan and Gordon's model [16] combines networks with ideal points makes it difficult to disentangle the two. Analyzing the networks alone may be inconclusive, because ideal points also supply the model with predictive power. Moreover, the machine learning algorithm also learns these two components simultaneously. Figure 4 shows the learned network for Sphere 2 (Economics & Finance) under this richer model. It is evident that the

two parties are not as clustered as they are in the LIG model (see Fig. 1 for Sphere 2 under LIG). This is because the ideal points are now being used in addition to social interactions to discriminate the behaviors of opposing senators. More interestingly, an analysis of the cross-party edges show that there are a lot more negative edges between the two parties under this richer model than there are under the LIG model (details are in the Appendix).

Nevertheless, we still study the ideal point distributions and influence networks under this model. We apply ideal point-based polarization metrics [26] as well as network modularity metrics [14] to calculate polarization levels across the four spheres. The ideal point distributions for two of the spheres are depicted in Fig. 5 (others are in the Appendix).

Applying the well-known ideal point-based polarization metric (i.e., distance between the means of the two parties) [26], we obtain values of 0.754, 1.235, 1.126, and 0.889 for Spheres 1 to 4 respectively. Evidently, Sphere 1 is least polarizing with respect to the ideal point distributions alone.

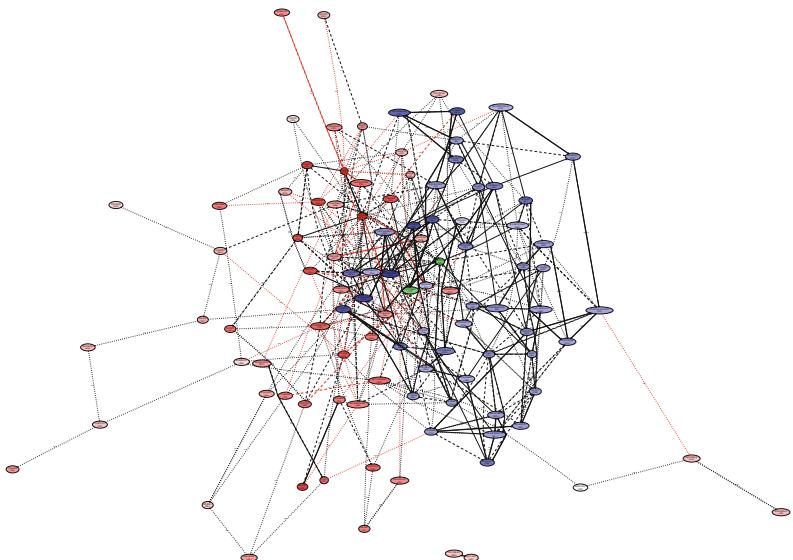
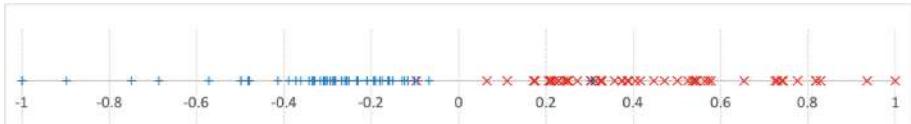


Fig. 4. A bird's eye view of the influence network for Sphere 2 (Economics & Finance) learned under the ideal point model with social interactions. The strongest 33% of the edges are shown. Contrast this with Fig. 1 where the two parties were relatively well separated.

We also analyze the influence networks for the four spheres. The modularity framework discussed in Sect. 5 yields scores of 0.5392, 0.6801, 0.6887, and 0.6229, respectively. Both the ideal point metric and modularity scores indicate that Spheres 2 (Economics & Finance) and 3 (Energy & Infrastructure) are most polarizing, whereas Sphere 1 (Security & Armed Forces) is least polarizing.



(a) Sphere 1 (Security & Armed Forces): The distance between the mean ideal points of the two parties is 0.754. It shows that Democratic and Republican senators are ideologically close to each other when it comes to national security.



(b) Sphere 2 (Economics & Finance): The distance between the mean ideal points of the two parties is 1.235, which shows more polarization compared to Sphere 1.

Fig. 5. The ideal point distributions of Democratic (blue +) and Republican (red x) senators, scaled linearly between -1 and 1 .

Sphere 4 (Public Welfare) sits in between. These results are somewhat similar to our earlier conclusions based on LIG without ideal points.

As mentioned above, investigating the influence networks and ideal points separately does not give us the complete picture since the model combines these two components together to make predictions. Therefore, we should also combine them in a meaningful way to infer polarization. We leave this as future work.

Acknowledgement. This research was partially supported by NSF grant IIS-1910203. We sincerely thank Dr. Stephen Majercik (Bowdoin College) for reading an earlier draft of this paper and giving us many valuable suggestions. We are also thankful to Drs. Honorio and Ortiz for letting us use their codes [15] and to the anonymous reviewers for their suggestions.

Appendix

Link to the Appendix: <http://bit.ly/appendix-spheres-of-legislation>.

References

1. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on facebook. *Science* **348**(6239), 1130–1132 (2015)
2. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981)
3. Chen, W., Collins, A., Cummings, R., Ke, T., Liu, Z., Rincon, D., Sun, X., Wang, Y., Wei, W., Yuan, Y.: Influence maximization in social networks when negative opinion may emerge and propagate. In: Proceedings of the Eleventh SIAM International Conference on Data Mining, pp. 379–390 (2011)

4. Christakis, N.A., Fowler, J.H.: The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* **357**(4), 370–379 (2007). <http://content.nejm.org/cgi/content/abstract/357/4/370>
5. Christakis, N.A., Fowler, J.H.: The collective dynamics of smoking in a large social network. *N. Engl. J. Med.* **358**(21), 2249–2258 (2008). <http://content.nejm.org/cgi/content/abstract/358/21/2249>
6. Clinton, J., Jackman, S., Rivers, D.: The statistical analysis of roll call data. *Am. Polit. Sci. Rev.* **98**(2), 355–370 (2004)
7. Conover, M.D., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., Flammini, A.: Political polarization on Twitter. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
8. Farina, C.R.: Congressional polarization: terminal constitutional dysfunction. *Colum. L. Rev.* **115**, 1689 (2015)
9. Garcia, D., Abisheva, A., Schweighofer, S., Serdült, U., Schweitzer, F.: Ideological and temporal components of network polarization in online political participatory media. *Policy Internet* **7**(1), 46–79 (2015)
10. Gerrish, S.M., Blei, D.M.: How they vote: issue-adjusted models of legislative behavior. *Adv. Neural Inf. Process. Syst.* **25**(1), 2762–2770 (2012)
11. Girvan, M., Newman, M.E.: Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**(12), 7821–7826 (2002)
12. Granovetter, M.: Threshold models of collective behavior source. *Am. J. Sociol.* **83**(6), 1420–1443 (1978). <https://www.jstor.org/stable/27781>
13. Guerra, P.C., Meira Jr., W., Cardie, C., Kleinberg, R.: A measure of polarization on social media networks based on community boundaries. In: 7th International AAAI Conference on Weblogs and Social Media (2013)
14. Gómez, S., Jensen, P., Arenasl, A.: Analysis of community structure in networks of correlated data. *Institut des Systèmes Complexes* **1**(1), 2–3 (2009)
15. Honorio, J., Ortiz, L.: Learning the structure and parameters of large-population graphical games from behavioral data. *J. Mach. Learn. Res.* **16**, 1157–1210 (2015). <http://arxiv.org/abs/1206.3713>
16. Irfan, M.T., Gordon, T.: The power of context in networks: ideal point models with social interactions. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pp. 910–918. International Foundation for Autonomous Agents and Multiagent Systems (2018)
17. Irfan, M.T., Ortiz, L.E.: A game-theoretic approach to influence in networks. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, pp. 688–694 (2011). <http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3746>
18. Irfan, M.T., Ortiz, L.E.: On influence, stable behavior, and the most influential individuals in networks: a game-theoretic approach. *Artif. Intell.* **215**, 79–119 (2014). <http://www.sciencedirect.com/science/article/pii/S0004370214000812>
19. Jackson, M.O.: Social and Economic Networks. Princeton University Press, Princeton (2010)
20. Kearns, M., Littman, M., Singh, S.: Graphical models for game theory. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 253–260 (2001)
21. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003)

22. Kempe, D., Kleinberg, J., Tardos, É.: Influential nodes in a diffusion model for social networks. In: Proceedings of the 32nd International Colloquium on Automata, Languages and Programming (ICALP) (2005)
23. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web (TWEB) **1**(1), 5 (2007)
24. Li, Y., Chen, W., Wang, Y., Zhang, Z.L.: Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 657–666. ACM (2013)
25. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1: Statist, pp. 281–297 (1967). <https://projecteuclid.org/euclid.bsmsp/1200512992>
26. McCarty, N., Poole, K.T., Rosenthal, H.: Polarized America: The Dance of Ideology and Unequal Riches. MIT Press, Cambridge (2016)
27. Newman, M.E.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. **103**(23), 8577–8582 (2006)
28. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)
29. Poole, K.T., Rosenthal, H.: A spatial model for legislative roll call analysis. Am. J. Polit. Sci. **29**(2), 357–384 (1985). Published by: Midwest Political Science Association St. Political Science 29(2), 357–384 (2008)
30. Ross, T.J.: Fuzzy Logic with Engineering Applications, 2nd edn. Wiley, Hoboken (2004)
31. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, New York (1994)
32. Waugh, A.S., Pei, L., Fowler, J.H., Mucha, P.J., Porter, M.A.: Party polarization in congress: a network science approach. arXiv preprint [arXiv:0907.3509](https://arxiv.org/abs/0907.3509) (2009)
33. Zhang, Y., Friend, A.J., Traud, A.L., Porter, M.A., Fowler, J.H., Mucha, P.J.: Community structure in congressional cosponsorship networks. Phys. A Stat. Mech. Appl. **387**(7), 1705–1712 (2008)



Why We Need a Process-Driven Network Analysis

Mareike Bockholt^(✉) and Katharina A. Zweig

Department of Computer Science, Algorithm Accountability Lab, TU Kaiserslautern,
Gottlieb-Daimler-Straße 48, 67663 Kaiserslautern, Germany
{mareike.bockholt,zweig}@cs.uni-krl.de

Abstract. A network representation is a powerful abstraction of a complex system, on which a full range of readily available methods from network analysis can be applied. A network representation is suitable if indirect effects are of interest: if A has an impact on B and B has an impact on C, it is assumed that also A has an impact on C. This implies that some process is flowing through the network. For a meaningful network analysis, the network process, the network representation, and the applied network measure cannot be chosen independently [3, 4, 9, 30]. We propose a process-driven perspective on network analysis, which takes into account the network process additionally to the network representation. In order to show the necessity of this approach, we collected four data sets of real-world processes. As first step, we show that the assumptions of standard network measures about the properties of a network process are not fulfilled by the real-world process data. As second step, we compare the network usage pattern by real-world processes to the usage pattern of the corresponding shortest paths and random walks. Our results support the importance of a process-driven network analysis.

Keywords: Network analysis · Network processes · Dynamics on networks · Paths

1 Introduction

In the last two decades, there has been an increasing interest in the behaviour of complex systems, i.e., systems consisting of entities interacting with each other. Examples of complex systems include social systems of human interactions, biological systems of protein-protein interactions, or transportation systems as the world-wide air transportation system. A popular approach for analyzing such systems is using network analysis [1] where the observations of the system's behaviour are transformed into a graph structure: entities are represented by nodes and their interactions by edges. Having a network representation of a system allows the application of network analytic methods such as measuring the structure of the network, identifying groups of densely connected nodes [12], detecting network motifs [17], or finding the most central node [16].

In the network analysis community, the transformation of a system into a network representation seems natural by now, but it is actually a considerable

simplification of the system which is rarely unique: for the same system, there is always more than one plausible and well-defined network representation [30].

Let us give a small example: Consider a data set containing a sample of passengers' tickets of domestic flights within the US (as introduced in Sect. 3). For each passenger's journey, the data set contains one entry for each non-stop flight connection of the journey, including start and destination airport of the sub-journey, airline, type and size of the airplane, etc. One can easily come up with at least a dozen of different plausible sounding network representations of the same data set: nodes might represent single airports or cities. An edge might be inserted if there is a single flight from one airport to the other, or if there are flights with a minimum volume, or at a regular basis, etc.

It has been shown that seemingly trivial decisions in creating the network representation have an impact on the structure of the network and on the results of the network analytic methods. Butts illustrated that different choices of node aggregations have a substantial impact on the fundamental properties of the resulting network [4]. Similarly, Choudhury experimented with different plausible edge definitions in an email communication network and demonstrated the large effect of different thresholds on the same data set [5]. Hence, the choice of the "right" network representation is crucial for the interpretability and relevance of the results of methods applied on the network representation.

A network is a convenient structure for a system representation if not only pairwise interactions between entities are of relevance. If only pairwise interactions are relevant, a list of separate dyadic relations is a sufficient representation. A network representation (and the applicable methods) is suitable if *indirect effects* are of interest: if there is an effect of entity A on entity B, and an effect of B on C, it is assumed that there is a kind of influence of A on C. The presence of indirect effects, however, implies that something is flowing through the network from node to node by following the edges. In a social system, a person can have an impact on a friend of a friend by forwarding a piece of information [13], by spreading a rumor [8] or by transferring a behavior [6]. In other systems, physical goods are transferred, diseases are spread [20], or humans use the system as infrastructure, such as passengers in transportation systems [7, 14].

We emphasize that only the presence of a *network process* makes a network representation meaningful. The presence of a network process is assumed by the majority of network measures. For example, the concept of graph distance is only meaningful if something is using those paths, hence, all metrics containing path lengths require the presence of a network process. As another example, the classic centrality indices degree, closeness and betweenness centrality were introduced by Freeman with the idea in mind that they measure a node's importance with respect to a specific process [10]. The measures, however, expect the process to be uniform in several aspects: the process is expected to be present at all nodes and edges with the same probability and same intensity, and to start and end in every node with the same probability. Thus, network analytic measures are using a static network representation and a simplified model of possible network flow processes instead of looking at the real network flow process.

We claim that any network analytic approach will benefit from a process-driven perspective where the network representation and the corresponding real process are both taken into account. We call this *process-driven network analysis*. This means that deducing a network representation and applying available network measures needs to be in accordance with the network process of interest.

For demonstrating the relevance of a process-driven network analysis, we collected four data sets containing process data and the corresponding network structure from different scenarios. We investigate whether the usage pattern of the network by the real-world processes are uniform as expected by standard network measures, i.e., whether the simplified model is a good proxy for the real-world process. Furthermore, standard network measures mostly expect the network process to move on shortest paths (e.g., closeness and betweenness centrality) or on random walks (e.g., random walk betweenness centrality [18], Google's PageRank [19], or community detection by WalkTrap [21]). We compare the network usage of the real-world process data to the usage of those extreme cases of process models. We show:

- (i) Neither of the four real-world processes are uniform in node or edge usage. Similarly to real-world networks' degree distribution, there are a few hub nodes which are used many times, and a large number of nodes which are used once or never.
- (ii) For neither of the real-world processes, it holds that all node pairs have the same probability of being start and target of the process.
- (iii) Two of the data sets can be sufficiently approximated by a shortest-path-model, while two others cannot.
- (iv) We simulate the real-world process by a set of random agents. Although, for a fair comparison, starting nodes and outreach potential of the agents are tied to the real-world process, the random agents do not reproduce the usage pattern of the real-world process, for three of the four data sets.

Structure of the Paper. Section 2 gives a short overview of related work. Section 3 describes the used data sets, before Sect. 4 investigates the usage pattern of the networks by the real process. Section 5 compares the network usage by the real process to its usage by shortest paths and random walks. The article concludes with Sect. 6.

Definitions and Notations. A graph $G = (V, E)$ consists of a node set V and an edge set $E \subseteq V \times V$. We consider directed graphs where the edges are ordered tuples. We associate a graph with a weight function $\omega : E \rightarrow \mathbb{R}$ assigning weights to the edges, an unweighted graph is associated with the trivial weight function $\omega(e) = 1 \forall e \in E$. A walk is an alternating finite sequence of nodes and edges $P = (v_1, e_1, v_2, \dots, e_{k-1} v_k)$ with $v_i \in V$ for $i \in \{1, \dots, k\}$ and $e_j = (v_j, v_{j+1}) \in E$ for $j \in \{1, \dots, k-1\}$. If the nodes and edges of P are distinct, we call P a path. If a node v is contained in a walk P , we write $v \in P$. The start and end node of a walk P are denoted by $s(P) = v_1$ and $t(P) = v_k$. The length of a walk P , is defined as $\omega(P) = \sum_{i=1}^{k-1} \omega(e_i)$. The distance from node v to node w , $d(v, w)$, is length of the shortest path from v to w .

2 Related Work

The relevance of network processes for network analysis is not new. An important contribution was made by Borgatti [3] who identified two dimensions by which exemplary network process can be distinguished. He linked the identified process properties to existing centrality indices. Consider closeness centrality as an example: the closeness centrality for a node v is defined as the inverse of the sum of the shortest path lengths from every node to v . Having a high closeness value means that the node can be reached quickly from any other node (on average). Therefore, it is assumed that there is a process flowing through the network using shortest paths or by a broadcasting mechanism. Applying closeness centrality on a network with a process which has neither of those properties, e.g., the spread of a disease, yields non-interpretable results [3]. Dorn et al. [9] bring together Borgatti's insight that network process and network measure are dependent, and Butts' insight that network representation and network measure are dependent [4], and call this interdependence Trilemma of network analysis. This idea is elaborated on by Zweig [30].

Recent works [25, 29] have questioned the common practice of transforming process data (as described in our example above) into a network representation where an edge is inserted from node i to node j if the process data contains a connection from i to j , a so-called first-order network. In this representation, dependencies contained in the process data where the choice of the next node depends on the previously visited nodes, are lost. Xu et al. [29] argue that the network representation itself should reflect those dependencies and propose higher-order networks.

Approaches analyzing data sets of one specific process exist a lot, for example spreading of diseases [20, 23], spreading of rumors [8], propagation of health behaviors [6], or human navigation in information networks [28]. Approaches exploiting the information contained in the process data in order to draw insights about the network itself have been proposed by several authors: Weng et al. show the effect of an information diffusion process on the network evolution [27]. Rosvall et al. derive community structures in the network by incorporating real process data into a Markov chain simulation [24].

3 Data Sets

In order to compare real-world processes with a shortest-path model and model based on random agents, we restrict our analysis to processes with a transfer mechanism, i.e., processes consisting of indivisible entities *moving* from node to node. Suitable data sets satisfies the following requirements and are described in the following paragraphs (see also Table 1): (i) it contains trajectories of process entities, i.e., a set of walks $\mathcal{P} = \{P_1, \dots, P_k\}$ (ii) the process trajectories can be mapped onto the network structure, (iii) the process entities have a target, (iv) which they try to reach as fast as possible.

Table 1. Overview of the used data sets.

Data set	Nodes	Edges	Process
Airline O&D Survey (DB1B) [22]	Airports	Non-stop airline connections	Passengers
London Transport (LT) [26]	Public transport stations	Public transport connections	Passengers
Wikispeedia (Wiki) [28]	Wikipedia articles	Hyperlinks	Players
Rush Hour (RH) [15]	Configurations	Valid game moves	Players

Airline Origin and Destination Survey (DB1B). The Bureau of Transportation Statistics provides an Airline Origin and Destination Survey (DB1B) for every quarter year [22]. It contains a 10% sample of airline tickets from reporting airline carriers within the US. For each itinerary, the data contains its start and destination airport and intermediate stops. The process of interest is passengers traveling by airplane in the network of airports. We use the data of the years 2010 and 2011. Passengers' journeys were split into outbound and return trip. We create a node for every airport, airports for which the journey data contains sub-itineraries done by bus or tram, are merged. A directed edge (v, w) is created if the data contains at least one itinerary with a flight connection from an airport in v to an airport in w . We consider an unweighted version of the network and a version with edges weighted with geographic distance between the airports.

London Transport (LT). Transport of London, a governmental authority responsible for public transport within the region of London, provides the Rolling Origin and Destination Survey [26], containing a 5% sample of all passengers' journeys with an Oyster Card, an electronic ticket, during a week in November 2017. For each journey, the data contains its start and destination station as well as the stations of train changes. We used the Underground timetables to construct a multi-layer network where each layer represents one underground line: a node represents an underground station, an edge in layer i is inserted from v to w if w can be reached from v with line i without changing trains. Note that this does not yield a graph in the form of a chain as a line plan would suggest, but the transitive closure of the chain. An edge (v, w) in layer i is weighted with the minimal travel time from v to w using line i . Note also that we did not take into account the time schedules of the lines. A one-layer network is constructed by merging the layers into one and taking the minimal edge weight.

Wikispeedia (Wiki). West et al. [28] provide logs of persons playing Wikispeedia. In this game, a player is given (or chooses) two Wikipedia articles and the goal is to navigate from the start article to the target article by following the hyperlinks. West collected more than 50 000 logs by offering the game on his web page. The node set is a subset of Wikipedia articles, directed edges are hyperlinks between articles. We consider only walks reaching the target and exclude moves revoked by the player via an Undo button.

Rush Hour (RH). This data set also contains game logs of players, attempting to solve an instance of a Rush Hour game. This single-player sliding-block puzzle consists of a board with 6×6 cells and a designated exit, representing a parking lot. Cars (blocks of width of one cell and of length of two or three cells) are placed on the board horizontally or vertically. Goal is to move the cars (forwards or backwards, not sideways) such that a designated target car can exit the board. The network is the state space of the game instance, containing all board configurations reachable from the start configuration by valid moves, and an undirected edge represents a valid move. The game logs were collected by Jarušek and Pelánek by their web-based tool for education [15]. Three different game instances were used for our analysis: game A is a very easy instance with an optimal solution length 3 while games B and C are of medium difficulty with an optimal solution length 11 and 13, respectively. Only walks ending in a solution configuration are considered.

Table 2. Properties of the used data sets. $|V|$ and $|E|$ denote the cardinality of node and edge set of the underlying graph, $|\mathcal{P}|$ the number of available walks.

Data set	$ V $	$ E $	$ \mathcal{P} $	Path length		Coverage
				Range	Average	
DB1B	462	12499	86 m	[42, 26922]	1909	100%
LT	268	13173	4.8 m	[1, 107]	16.3	100%
Wiki	4592	119804	51306	[1, 82]	5	90%
RH Game A	364	1524	3044	[3, 33]	5	63%
RH Game B	6769	33142	1965	[11, 59]	15	11%
RH Game C	830	4037	1472	[13, 95]	26	53%

4 Uniformity of Network Usage

Our first goal is to investigate whether a process-driven network analysis is a necessary approach at all. Most network measures which assume the presence of a network process, expect the process to be uniform. Consider betweenness centrality as an example: for a node v , the betweenness centrality is defined as $c_B(v) = \sum_{s,t \in V, s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$ where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(v)$ the number of shortest paths from s to t containing v , and with the convention $\frac{0}{0} = 0$. Several assumptions are included in this measure: (i) It is assumed that a process is flowing through the network using shortest paths (see also Borgatti [3]). (ii) It is summed over all node pairs (excluding a few), therefore assumed that there is process flow between each pair of nodes. (iii) Each node pair can contribute a value between 0 and 1 to the betweenness value, hence each node pair is considered as equally important. It needs to be tested if that simple model is a good proxy for the real flow behaviour.

For the four used data sets, however, it turns out—unsurprisingly—that those assumptions are not met. Table 2 shows the coverage of the networks by the corresponding process, i.e., the fraction of nodes which are visited by the process at least once. Except for London Transport and DB1B where the network was constructed from the process data, between 10 and 89% of the nodes were not visited by the process. We compute the usage $n_u(v) = |\{P \in \mathcal{P} | v \in P\}|$ of each node v , i.e., the number of process walks in which the node is contained in.

Figure 1 shows the cumulative distribution of the node usage for each process data set. We find that for all data sets, there are a few nodes which are used by many process entities, while the majority of the nodes is used at most once by the process entities. This disproves—for the used data sets—the assumption that real-world processes are present in all nodes with the same probability and intensity.

We perform the same analysis for node pairs: for each node pair $(s, t) \in V \times V$, we count how many process entities start in s and end in t (referred to as *node pair usage*, n_{pu}). We make a similar observation as before: only a fraction (between 42% and 46% for DB1B and LT, 0.14% for Wiki, and less than 0.01% for the RH games) of all node pairs is used as source and target of the process. While this is not surprising for the RH games since all players start in the same node and end in one of the solution nodes, this is less expected for the two transportation systems, DB1B and LT. The distribution of the node pair usage (of pairs with $n_{pu} > 0$) is shown in Fig. 2a. We find that for the RH games, the frequency of different node pair usage values (with $n_{pu} > 0$) is approximately the same. For Wiki and the transportation systems, the majority of node pairs is used rarely as source and destination, i.e., having a low node pair usage, while there are a few node pairs with a high node pair usage.

For those three data sets, we investigate which node pairs are used more often than expected (assuming each node pair was chosen as start and end with the same probability). For this purpose, we consider the node pairs of the network separately by distance. For a graph distance i , let $np(i) = |\{(s, t) \in V \times V | d(s, t) = i\}|$ be the number of node pairs in the graph with this distance, and let $n_{pu}(i) = |\{P \in \mathcal{P} | d(s(P), t(P)) = i\}|$ the number of process entities with distance i between start and end node. Note that we do not consider the length of the *process walk*, but of the *shortest path*. For the weighted network versions (DB1B with edges weighted by geographic distance, and LT with edges weighted by travel time), we introduce distance intervals of size 500 km and 5 min, respectively, and define $np(i)$ and $n_{pu}(i)$ for a distance interval i accordingly.

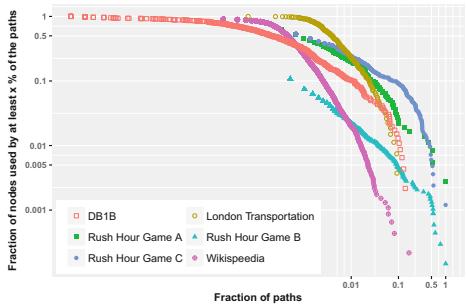
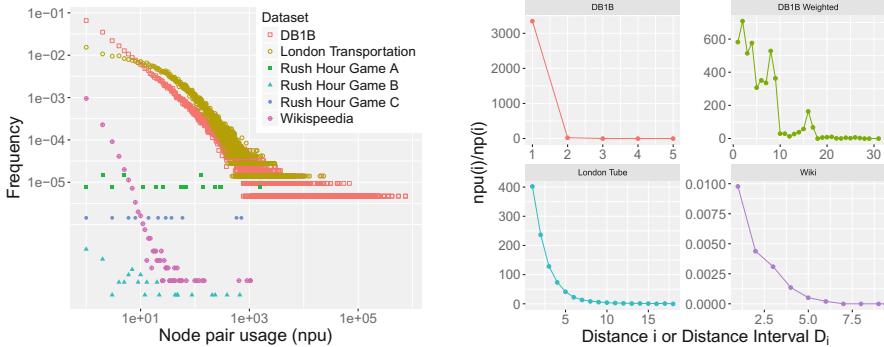


Fig. 1. Relative node usage of the real network processes.



(a) Source-target-frequency (normalized by the total number of node pairs) on a log-log-scale.

(b) Node pair usage normalized by the number of node pairs, by distance (or distance interval) of the nodes.

Fig. 2. Source-target-frequency of the real network processes.

Figure 2b shows the node pair usage $npu(i)$ divided by the number of node pairs $np(i)$ for each distance (or distance interval). We observe that the node pair usage is dependent on the graph distance of the pairs: for each data set, closer nodes are much more used as source and target than nodes which are further apart—far more than expected if node pairs were picked uniformly at random.

This observation might not be surprising for transportation systems, but the same observation has also been done in other networks. Friedkin considered communication networks and found a *horizon of observability* [11]: a distance in communication networks beyond which persons are unlikely to be aware of the existence of another. Also for the data sets used here, there seems to be a horizon of reachability: a distance beyond which there is only a negligible (or no) amount of process flow. All findings, (i) not all nodes are relevant for the network process, (ii) a few nodes are used heavily while most nodes are used only a few times, (iii) especially close nodes are source and target of network processes, could have been expected for the used data sets. They, however, do have consequences for network analysis in general: network analytic measures pretend a homogeneity of the network process—each node, each edge, each node pair is considered to be of the same quality. Our analysis shows that they are not if the real-world network process is taken into account. We argue that the properties of the network process need to be taken into account when deducing the network representation and applying network analytic methods on it.

5 Models of Processes

The last section showed that the network processes at hand do not use the underlying network uniformly. This is in contradiction to the implicit assumptions of most network measures which expect a uniform network process.

The two extreme possibilities of simulating real network processes are shortest paths and random walks. We are going to compare the real trajectories with both extremes.

Processes as Shortest Paths. For comparing a real process walk P to its corresponding shortest path, we compare its length to the length of the shortest path from $s(P)$ to $t(P)$. Figure 3 shows that for the transportation processes DB1B and LT, on average, the length of the real walks is close to the length of the shortest path. This is not true for the game data sets. For the RH games, the lengths of the real walks strongly depend on the game instance. For game B and C, only 14% (game B) and 2% (C) of the real walks have the same length as the optimal path. The real walks of the Wikispeedia game are also longer than their corresponding shortest paths, but on average only by one step.

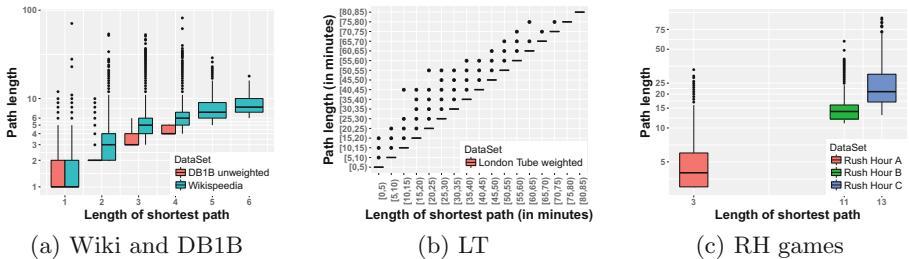


Fig. 3. Path lengths of different network processes compared to the length of the shortest path.

Processes as Random Walks. In contrast to shortest paths, the other extreme of a model for process trajectories are simulations by agent-based random walks: an agent starts at a node, moves to a randomly chosen neighbor node, and continues this procedure until a stopping criterion is reached. In order to compare the real trajectories to the trajectories of the random agents, the random agents are restricted by the real trajectories in the following way: for each real trajectory $P \in \mathcal{P}$, a random agent starts in $s(P)$, and performs a random walk as long it has not exceeded the length of P . By this procedure, it is made sure that the number of agents is equal to the number of real process entities, the node usage of the start nodes is equal for real and random agents, and the random agents have the same “outreach potential” as the real trajectories.

We implement two variants of neighbor choices: the agent picks uniformly at random from all neighbors of the current node (*uniform neighbor choice*), or from the neighbors except of the directly previously visited one (*backwards-restricted neighbor choice*). For data set LT, a second variant for the length restriction is implemented: the agent continues its walk until it has reached the same number of line changes as contained in the real trajectory. For each data set and its walk set \mathcal{P} , sets of random walks are created by repeating the above procedure $N = 500$ times. For the data sets DB1B and LT, we sample a subset of 0.1%

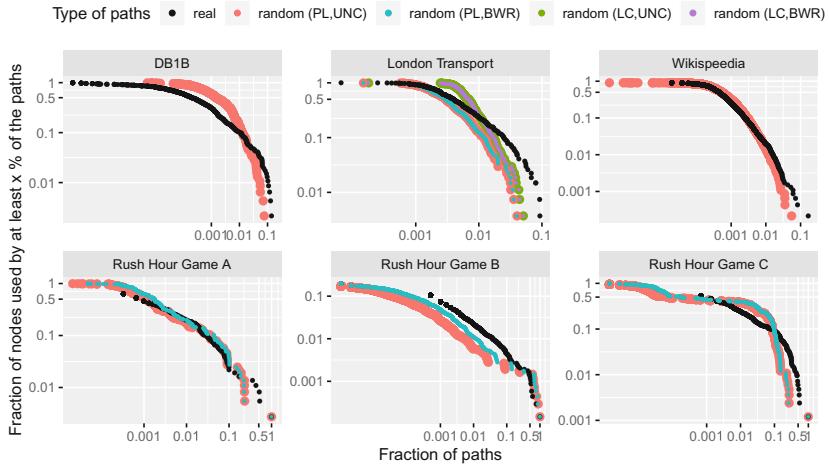


Fig. 4. Cumulative node usage distribution of real and random trajectories. The random walks implement a uniform neighbor choice (UNC) or a backwards-restricted neighbor choice (BWR), their length is restricted by the length (PL) or by the number of line changes of the corresponding real trajectory (LC).

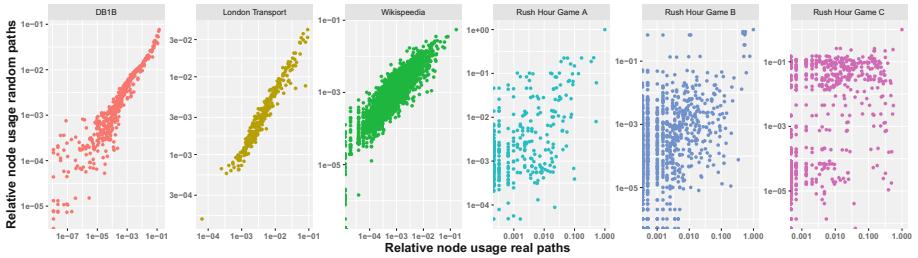


Fig. 5. Node usage of the real paths vs mean node usage of the random walks (UNC) on a log-log scale. Node usage is normalized by the number of (real or random) agents.

(DB1B) and 10 % (LT) of all real trajectories and tie (and compare) the random walks to those subsets.

Figure 4 shows the cumulative node usage by the real-world process and by the random agents (for each node the mean value over the 500 iterations is used). We also compare the node usage by the real process and by the random walks for each single node (see Fig. 5). We observe that for DB1B and LT, the random agents and the real entities yield a different usage pattern. On the same time, there is a high correlation of node usage by the real and random trajectories (Pearson correlation coefficient between 0.81 and 0.87). For the RH games, the findings are opposite: the node usage distribution is similar for random and real trajectories while having a lower correlation (Pearson correlation coefficient 0.77 (A), 0.64 (B) and 0.36 (C)). On the same time, for all data sets, it is the same set of nodes which is used the most often by real or random trajectories.

6 Conclusion and Future Work

A network representation is often a convenient abstraction for a system if indirect effects are of interest. This implies that a process is flowing through the network. A meaningful network analysis needs to take into account the properties of these processes. We call this perspective *process-driven network analysis*. In this work, we collected four data sets containing processes flowing on walks towards a goal. We find that none of those processes shows a uniform usage pattern, i.e., a few nodes are used very often by the process, while most nodes are used at most once. The same holds for node pairs being source and target of the processes. This has consequences for network measures expecting a uniform usage of the network. We furthermore compared the real-world processes to basic simulations, i.e., simulations by shortest paths and by random walks. We find that the transportation paths are, as expected, close to shortest paths in terms of length, while the game paths are not. Random walks where the source and potential outreach are fixed by the real-world process are able to show a similar node usage ranking for three of the four data sets, their node usage distribution is different though. This has consequences for network measures expecting a process moving on shortest paths or randomly through the network.

In order to take these insights into account, the following approaches might be applicable: If data of real-world processes are available, a network of higher order can be constructed and used for analysis [25]. Another option is the incorporation of real-world process data into existing network measures [2]. For considering the horizon of observability of processes [11], it might be a valid approach to separately analyze subgraphs of the network instead of the complete network. In general, we argue for a careful and thoughtful application of network measures on network representations.

References

1. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: structure and dynamics. *Phys. Rep.* **424**(4–5), 175–308 (2006)
2. Bockholt, M., Zweig, K.A.: Process-driven betweenness centrality measures. In: Network Intelligence Meets User Centered Social Media Networks. Lecture Notes in Social Networks, pp. 17–33. Springer (2018)
3. Borgatti, S.P.: Centrality and network flow. *Soc. Netw.* **27**(1), 55–71 (2005)
4. Butts, C.T.: Revisiting the foundations of network analysis. *Science* **325**(5939), 414–416 (2009)
5. Choudhury, M.D., Mason, W.A., Hofman, J.M., Watts, D.J.: Inferring relevant social networks from interpersonal communication. In: Proceedings of the 19th International Conference on World Wide Web - WWW 2010. ACM Press (2010)
6. Christakis, N.A., Fowler, J.H.: The collective dynamics of smoking in a large social network. *N. Engl. J. Med.* **358**(21), 2249–2258 (2008). PMID: 18499567
7. Colizza, V., Barrat, A., Barthélémy, M., Vespignani, A.: The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Nat. Acad. Sci.* **103**(7), 2015–2020 (2006)

8. De Domenico, M., Lima, A., Mougel, P., Musolesi, M.: The anatomy of a scientific rumor. *Sci. Rep.* **3**, 2980 (2013)
9. Dorn, I., Lindenblatt, A., Zweig, K.A.: The trilemma of network analysis. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 9–14 (2012)
10. Freeman, L.C.: A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41 (1977)
11. Friedkin, N.E.: Horizons of observability and limits of informal control in organizations. *Soc. Forces* **62**(1), 54–77 (1983)
12. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *Proc. Nat. Acad. Sci.* **99**(12), 7821–7826 (2002)
13. Guille, A., Hacid, H., Favre, C., Zighed, D.A.: Information diffusion in online social networks: a survey. *SIGMOD Rec.* **42**(2), 17–28 (2013)
14. Guimerá, R., Amaral, L.A.N.: Modeling the world-wide airport network. *Eur. Phys. J. B* **38**(2), 381–385 (2004)
15. Jarušek, P., Pelánek, R.: Analysis of a simple model of problem solving times. In: Intelligent Tutoring Systems, volume 7315 of Lecture Notes in Computer Science, pp. 379–388. Springer (2012)
16. Koschützki, D., Lehmann, K.A., Peeters, L., Richter, S., Tenfelde-Podehl, D., Zlotowski, O.: Centrality indices. In: Brandes, U., Erlebach, T. (eds.) Network Analysis. Lecture Notes in Computer Science, vol. 3418, pp. 16–61. Springer (2005)
17. Milo, R.: Network motifs: simple building blocks of complex networks. *Science* **298**(5594), 824–827 (2002)
18. Newman, M.E.: A measure of betweenness centrality based on random walks. *Soc. Netw.* **27**(1), 39–54 (2005)
19. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)
20. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**(14), 3200–3203 (2001)
21. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: Computer and Information Sciences, pp. 284–293. Springer (2005)
22. RITA TransStat. Origin and Destination Survey database (DB1B) (2016). https://www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=125
23. Rocha, L.E.C., Liljeros, F., Holme, P.: Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput. Biol.* **7**(3), e1001109 (2011)
24. Rosvall, M., Esquivel, A., Lancichinetti, A., et al.: Memory in network flows and its effects on spreading dynamics and community detection. *Nat. Commun.* **5**, 4630 (2014). <https://doi.org/10.1038/ncomms5630>
25. Scholtes, I.: When is a network a network? In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017. ACM Press (2017)
26. Transport for London. Rolling Origin and Destination Survey (RODS) (2017). <http://www.tfl.gov.uk/info-for/open-data-users/our-feeds>
27. Weng, L., Ratkiewicz, J., Perra, N., Gonçalves, B., Castillo, C., Bonchi, F., Schifanella, R., Menczer, F., Flammini, A.: The role of information diffusion in the evolution of social networks. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, pp. 356–364. ACM Press (2013)

28. West, R., Leskovec, J.: Human wayfinding in information networks. In: Proceedings of the 21st International Conference on World Wide Web, pp. 619–628. ACM Press (2012)
29. Xu, J., Wickramarathne, T.L., Chawla, N.V.: Representing higher-order dependencies in networks. *Sci. Adv.* **2**(5), e1600028 (2016)
30. Zweig, K.A.: Network Analysis Literacy. Springer, Vienna (2016)



Gender's Influence on Academic Collaboration in a University-Wide Network

Logan McNichols, Gabriel Medina-Kim, Viet Lien Nguyen, Christian Rapp,
and Theresa Migler^(✉)

California Polytechnic State University, San Luis Obispo, CA 93405, USA
tmigler@calpoly.edu

Abstract. We present a collaboration network of the faculty of a large polytechnic state university in the United States. In our network vertices represent researchers: faculty members at the university and collaborators of these faculty members. Edges between two researchers in this network represent some sort of collaborative experience: conference publication, journal publication, or grant application. In this paper we present a study of this network and various subnetworks with respect to gender.

Keywords: Collaboration network · Interdisciplinary collaboration · Gender

1 Introduction

Co-authorship networks, in which vertices represent authors and two authors are connected if they have co-authored a paper together, are fundamental to our understanding of dynamics among groups of researchers. In particular, there is a strong drive for academics to become *interdisciplinary*, to combine two or more academic disciplines into one activity. For example, the European Union framework program Horizon 2020 has policy actors that state that interdisciplinary research will be the “key to future scientific breakthroughs” [2]. It is predicted that the future of research will become increasingly interdisciplinary [13]. In order to study interdisciplinary collaborations it is necessary to study networks that include as many disciplines as possible. The university being studied has six colleges and 49 distinct departments. We present a large network of 20,822 vertices, 1,855 of these vertices represent the university’s faculty and 18,967 of these vertices represent external collaborators. Two researchers are connected if they have ever shared a collaborative experience: a co-publication of a conference or journal paper or a book or a grant application. We constructed this network with a view to understand the behavior of collaborations across disciplines.

2 Related Work

Collaboration networks are among the most studied networks in network science. Collaboration networks range from the popular “film actor network”, where vertices represent film actors and two film actors are connected if they have ever appeared in a film together [18], to the most often studied co-authorship networks of academics, where vertices represent researchers and two researchers are connected if they have ever co-authored a paper [4, 6, 7, 15, 16]. Most of these collaboration networks are built with respect to a certain field, for example vertices represent mathematicians and two mathematicians are connected if they have published a paper [10]. There have been studies of collaborative networks across several related departments within a university. In 1978 Friedkin studied the collaborative networks of 128 faculty members across six physical science departments [9]. However, to the best of the authors’ knowledge there has been no large-scale university-wide collaboration network across many disciplines constructed or studied.

The analysis of gender with respect to collaborations has been widely studied. Ductor, Goyal, and Prummer showed that women have fewer collaborators, collaborate more often with the same co-authors, and a higher fraction of their co-authors are co-authors of each other [8]. Holman and Morandin found that researchers preferentially co-publish with colleagues of the same gender, and show that this ‘gender homophily’ is slightly stronger today than it was 10 years ago [12]. Using a dataset with more than 270,000 scientists, Araújo, Araújo, Moreira, Herrmann, and Andrade show that while men are more likely to collaborate with other men, women are more egalitarian regardless of how many collaborators each scientist has, with the exception of engineering where the bias disappears with increasing number of collaborators [3]. Abramo, D’Angelo, and Murgia study the scientific work of Italian academics and found that women researchers register a greater capacity to collaborate, with the exception of international collaboration, where there is still a gap in comparison to male colleagues [1]. And West, Jacquet, King, Correll, and Bergstrom showed that on average women publish fewer papers than men [19]. In what follows we will confirm most of these observations for our network.

3 The Network

The university that we are building this collaboration network for has approximately 1,473 current faculty members, both full and part-time. The university is mainly an undergraduate university with 21,037 undergraduate students and 775 graduate students. The university has six colleges: the College of Agriculture, Food and Environmental Sciences, the College of Architecture and Environmental Design, the College of Business, the College of Engineering, the College of Liberal Arts, and the College of Science and Mathematics. Among these six colleges there are 49 distinct departments. We refer to researchers at this university as *university researchers*. Collaborators of university researchers are *external researchers* (although most external researchers work at a different university).

A vertex in the network represents a researcher who is either a university researcher or an external researcher who collaborates with a university researcher. We initially populated the vertex set by scraping each of the 49 departments' webpages for faculty lists. Then as we found collaborative experiences for these university researchers with external researchers, we added these external researchers to our vertex set.

The edges in our network represent *collaborative experiences* between two researchers which represent a scholarly work (a publication or grant application) where a university researcher is included as an author. We populated the collaboration data from five sources:

- Grants applications from the university's grants office
- Publication records from Google Scholar, Microsoft Academic, and *arxiv*
- In the case of the Computer Science Department, personal webpages and curricula vitae for faculty members

Based on many findings that Google Scholar provides a comprehensive coverage that meets or exceeds that of similar bibliographic databases [11], we chose to use collaborations from Google Scholar as our “backbone” set of collaborations. One criticism of Google Scholar is that there may be duplicate entries [11]. We chose not to use services such as ResearchGate because there has been much criticism that such services give equal weight to legitimate journals as well as predatory or hijacked journals [14].

Note that a single collaborative experience with i authors will produce a clique in the graph of size i .

4 Properties of the Networks

We consider three nested networks:

- **The Computer Science Network:** Vertices represent faculty in the Computer Science Department and their collaborators.
 - This network has 54 Computer Science Department university researcher vertices and 1,314 external researcher vertices (1,368 total vertices), and 6,682 edges.
- **The College of Engineering Network:** Vertices represent faculty in the College of Engineering and their collaborators.
 - This network has 207 College of Engineering university researcher vertices and 4,569 external researcher vertices (4,776 total vertices), and 25,438 edges.
- **The University-Wide Network:** Vertices represent faculty at the university and their collaborators.
 - This network has 1,855 university researcher vertices (this includes emeritus faculty) and 18,967 external researcher vertices (20,822 total vertices), and 106,728 edges.

In these three networks edges represent collaborative experiences between two researchers in the network. We exclude collaborative experiences with more than seven researchers as we feel that we cannot draw collaborative information from such large groups. If there are twenty authors we can't expect that every pair of the twenty authors has ever even met the other.

We note that at the university being studied, the Computer Science Department is housed in the College of Engineering, so these networks are truly nested.

The Computer Science Network is the only network that has been human verified. For each computer science university researcher in this network, we verified that all of the collaborations listed on this researcher's curriculum vitae were in the network, and further that there weren't any duplicate entries.

5 Gender and the Networks

In order to analyze our networks with respect to gender we used *Gender API*, a gender inference service. Gender API currently supports 178 countries and provides confidence parameters *samples*, the number of database records matching the request, and *accuracy*, the reliability of the assignment. When comparing with other gender inference tools, Gender API is the best performer in terms of fraction of inaccuracies [17]. Unfortunately, many external researchers in the network first initial to represent their first name. In such cases we mark the gender as *unknown*. Further there are cases where Gender API cannot conclusively identify a name as male or female, we also mark these genders as unknown.

5.1 Basic Gender Statistics

In the University-Wide network there are 5,216 total female vertices, 691 of which are female university vertices, 13,000 total male vertices, and 1,150 of which are male university vertices, 2,606 total unknown vertices, 14 of which are unknown university vertices.

In the College of Engineering Network there are 811 total female vertices, 31 of which are female university vertices, 3,274 total male vertices, 173 of which are male university vertices, and 691 total unknown vertices, 3 of which are unknown university vertices.

In the Computer Science Network there are 282 total female vertices, 11 female university vertices, 966 total male vertices, 43 male university vertices, and 120 total unknown vertices, 0 of which are unknown university vertices. See Fig. 2 for a summary of these counts.

In the following we use the term *internal* to mean: for the University-Wide Network, simply university researchers, for the College of Engineering Network, university vertices who are further College of Engineering faculty members, and for the Computer Science Network university vertices who are further Computer Science faculty members. See Fig. 1 for the Computer Science network with only internal edges and vertices.

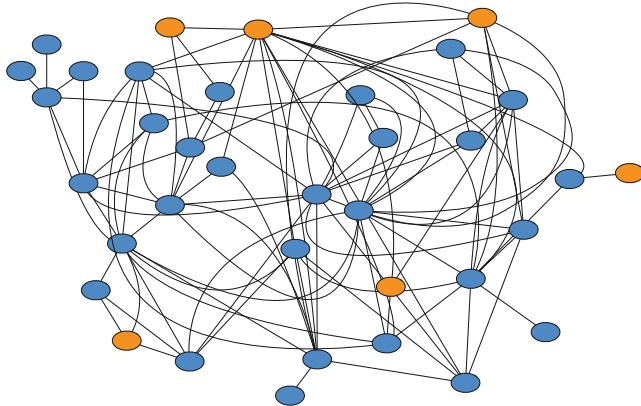


Fig. 1. A subnetwork of the Computer Science network. Vertices represent Computer Science faculty and there is an edge between two faculty members if they are collaborators. All isolated vertices have been removed. Orange vertices represent female faculty members and blue vertices represent male vertices.

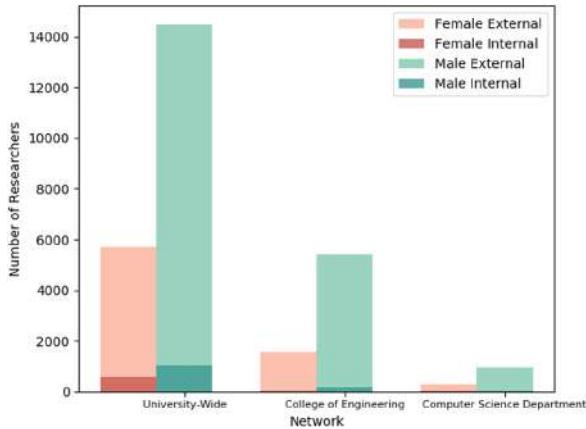


Fig. 2. Counts for female and male internal and external vertices.

5.2 Claim: Women Have Fewer Collaborators Than Men

Ductor, Goyal, and Prummer showed that women have fewer collaborators than men [8].

We tested this hypothesis on each of our three networks by removing multiple edges then comparing the average degree of the female internal vertices with those of the male internal vertices. We found that female researchers have substantially fewer collaborators than the male researchers except for the College of Engineering Network where the difference is not so strong. See Fig. 3.

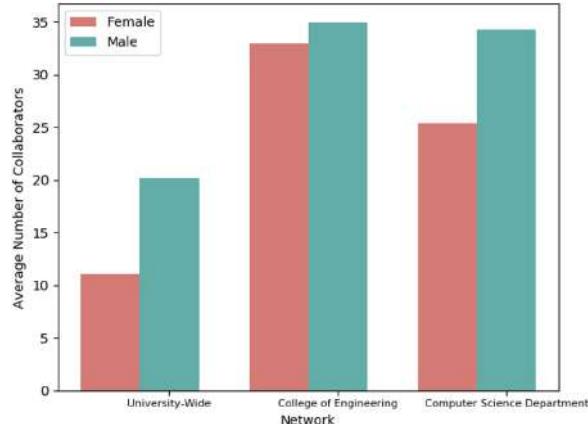


Fig. 3. The average number of female and male collaborators in each of the three networks.

5.3 Claim: Women Collaborate More Often with the Same Co-Authors

Ductor, Goyal, and Prummer showed that women collaborate more often with the same co-authors [8].

To obtain the *repeat collaboration number* for each researcher we summed the number of multiple collaborations that they had with other researchers. For example, if researcher *A* collaborated with researcher *B* once, researcher *C* three times, and researcher *D* five times, then *A*'s repeat collaboration number is six (two repeats with *C* and four repeats with *D*). To test the hypothesis on each of our three networks we averaged the repeat collaboration number for each internal female vertex with the repeat collaboration number for each internal male vertex. We found that the repeat collaboration number is actually higher for males than females in both the University-Wide and Computer Science Department networks. See Fig. 4.

Then we calculate the *repeat collaborator number* for each researcher. This is simply a count of the number of collaborators that the researcher has worked with more than once. For example, suppose researcher *A* has three collaborators, *B*, *C*, and *D*. Suppose that *A* has collaborated with *B* three times, with *C* one time, and with *D* seven times. In this case, *A*'s repeat collaborator number is two (one for *B* and one for *D*). Again, we found that the repeat collaborator number is higher for males than for females in all three networks. See Fig. 5.

5.4 Claim: A Higher Fraction of Women's Co-Authors Are Co-Authors with Each Other

Ductor, Goyal, and Prummer showed that a higher fraction of women's co-authors are co-authors of each other [8].

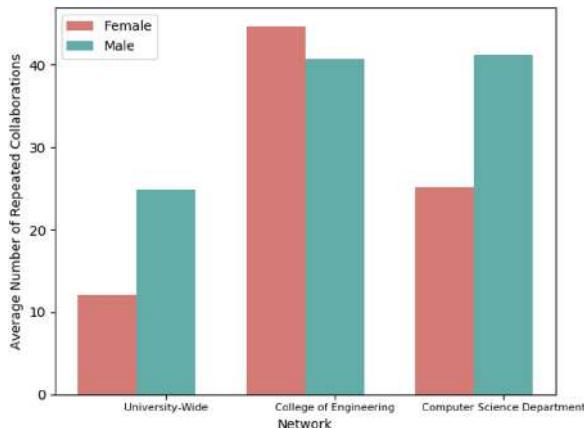


Fig. 4. The average repeat collaboration number for female and male collaborators in each of the three networks.

To test this hypothesis we found the average local clustering coefficient for internal female vertices and compared it with the average local clustering coefficient for internal male vertices. The clustering coefficient for female researchers is higher than that for males, except in the Computer Science Department network where they are very similar. See Fig. 6

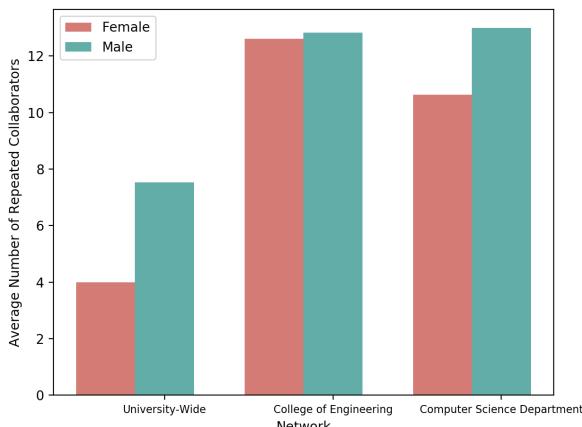


Fig. 5. The average repeat collaborator number for female and male collaborators in each of the three networks.

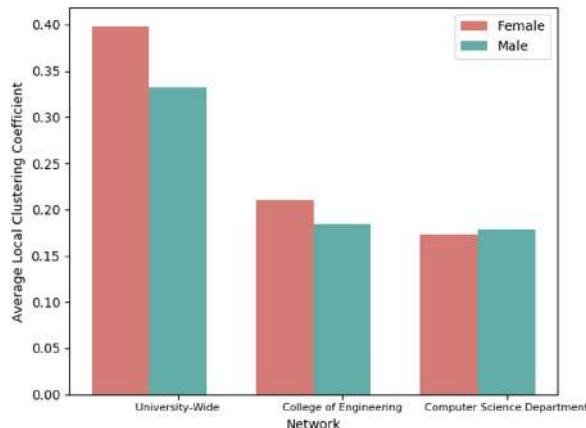


Fig. 6. The average local clustering coefficient for female and male collaborators in each of the three networks.

5.5 Claim: Researchers Preferentially Co-Publish with Authors of the Same Gender

Holman and Morandin found that researchers preferentially co-publish with colleagues of the same gender [12]. However Araújo, Araújo, Moreira, Herrmann, and Andrade show that while men are more likely to collaborate with other men, women are more egalitarian regardless of how many collaborators each scientist has [3].

To test homophily, we took the average over all internal female researchers the ratio of the number of their female collaborators to the number of their

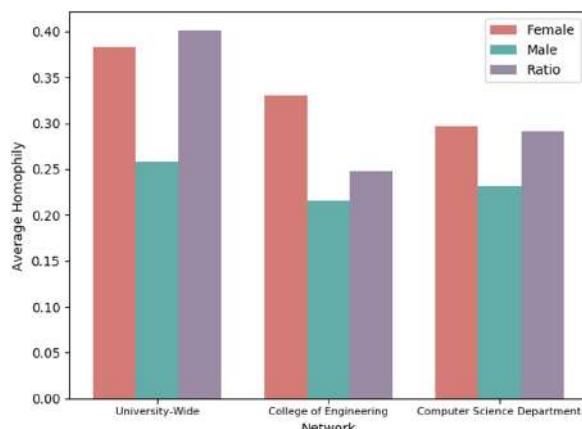


Fig. 7. The average homophily for female and male collaborators in each of the three networks.

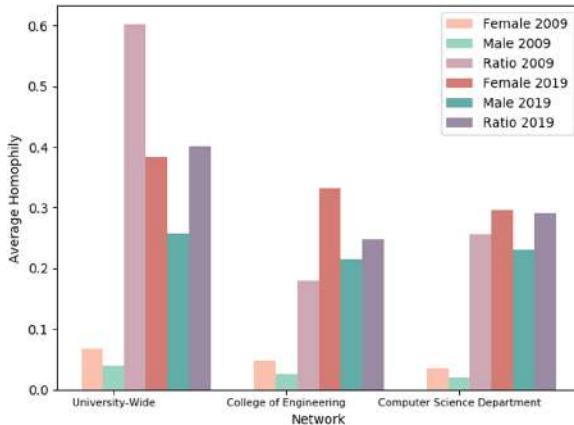


Fig. 8. The average homophily for female and male collaborators with publications until 2019 compared with the average homophily for female and male collaborators with publications until 2009 in each of the three networks

male collaborators. We took the same average for internal male researchers. For reference we include the simple ratio of total female researchers to total male researchers. See Fig. 7.

5.6 Claim: ‘Gender Homophily’ Increases over Time

Holman and Morandin found that ‘Gender homophily’ is slightly stronger today than it was 10 years ago [12].

To test this, we reduced our three networks by including only edges from collaborations that happened before 2009, we removed any isolated vertices then tested homophily in the same way as the previous section. See Fig. 8.

6 Future Work

We intend to manually verify the curriculum vitae of every faculty member at the university in an attempt to form a more complete network. At the time of writing, only the Computer Science faculty members’ CVs have been verified by humans. We further wish to request demographic and job satisfaction information for each university faculty member in the network in the form of an online survey. This survey will include gender and ethnicity. We will also request that each university faculty member states their satisfaction with their position on a Likert scale in an attempt to answer the question: “Does a higher clustering coefficient indicate greater satisfaction?”. In this same survey we also wish for authors to self-identify additional collaborative experiences that may not be documented by our existing collaborative experiences in order to create a more complete network. For example, two faculty members who wrote a paper together that

did not get accepted to any conference or journal. We also intend to study this network as it evolves over the years. NSF found that between 2000 and 2013, the percentage of publications with authors from multiple countries rose from 13.2% to 19.2% [20]. Noting that research is becoming increasingly international, we further wish to analyze the collaborations in our network with respect to location.

Acknowledgements. The authors would like to express their sincerest thanks to Professor Zoë Wood for inspiring this work based on her work with The Center for Research, Excellence and Diversity in Team Science (CREDITS), an integrated research and training program to increase and enhance Team Science and collective intelligence capacity, effectiveness, and excellence in California [5]. We would also like to express our thanks to Professor Foaad Khosmood for teaching our group about data sustainability and data scraping.

References

1. Abramo, G., DâAngelo, C.A., Murgia, G.: Gender differences in research collaboration. *J. Inf.* **7**(4), 811–822 (2013)
2. Allmendinger, J.: Quests for interdisciplinarity: a challenge for the era and horizon 2020. *Science Europe Position Statement* (2012)
3. Araújo, E.B., Araújo, N.A.M., Moreira, A.A., Herrmann, H.J., Andrade, J.S.: Gender differences in scientific collaborations: women are more egalitarian than men. *PLoS ONE* **12**(5), 1–10 (2017)
4. 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), 590–614 (2002)
5. Campbell, L.G., Mehtani, S., Dozier, M.E., Rinehart, J.: Gender-heterogeneous working groups produce higher quality science. *PLoS ONE* **8**(10), 1–6 (2013)
6. De Castro, R., Grossman, J.: Famous trails to Paul Erdős. In: Vitanyi, P.M.B. (ed.) *The Mathematical Intelligencer*, vol. 21, January 1999
7. Ding, Y.: Scientific collaboration and endorsement: network analysis of coauthorship and citation networks. *J. Inf.* **5**(1), 187–203 (2011)
8. Ductor, L., Goyal, S., Prummer, A.: Gender & collaboration. Cambridge working papers in economics, Faculty of Economics, University of Cambridge (2018)
9. Friedkin, N.E.: University social structure and social networks among scientists. *Am. J. Sociol.* **83**(6), 1444–1465 (1978)
10. Grossman, J.W., Ion, P.D.F.: On a portion of the well-known collaboration graph. *Congressus Numerantium* **108**, 129–131 (1995)
11. Harzing, A.-W., Alakangas, S.: Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics* **106**(2), 787–804 (2016)
12. Holman, L., Morandin, C.: Researchers collaborate with same-gendered colleagues more often than expected across the life sciences. *PLoS ONE* **14**(4), 1–19 (2019)
13. Lyall, C., Meagher, L.R.: A masterclass in interdisciplinarity: research into practice in training the next generation of interdisciplinary researchers. *Futures* **44**(6), 608–617 (2012). Special Issue: Politics, Democracy and Degrowth
14. Memon, A.R.: Researchgate is no longer reliable: leniency towards ghost journals may decrease its impact on the scientific community. *J. Pak. Med. Assoc.* **66**(12), 1643–1647 (2016)

15. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. *Proc. Nat. Acad. Sci.* **101**(suppl 1), 5200–5205 (2004)
16. Newman, M.E.J.: Scientific collaboration networks. i. network construction and fundamental results. *Phys. Rev. E, Stat. Nonlinear Soft Matter Phys.* **64**, 0416131 (2001)
17. Santamaría, L., Mihaljevic, H.: Comparison and benchmark of name-to-gender inference services. *Peer J. Comput. Sci.* **4**, e156 (2018)
18. Watts, D.J.: Small Worlds: The Dynamics of Networks Between Order and Randomness. Princeton studies in complexity. Princeton University Press, Princeton (2004)
19. West, J.D., Jacquet, J., King, M.M., Correll, S.J., Bergstrom, C.T.: The role of gender in scholarly authorship. *PLoS ONE* **8**(7), 1–6 (2013)
20. Witze, A.: Research gets increasingly international. *Nature*, January 2016



Centrality in Dynamic Competition Networks

Anthony Bonato¹(✉), Nicole Eikmeier², David F. Gleich³, and Rehan Malik¹

¹ Ryerson University, Toronto, ON, Canada
abonato@ryerson.ca

² Grinnell College, Grinnell, IA, USA

³ Purdue University, West Lafayette, IN, USA

Abstract. Competition networks are formed via adversarial interactions between actors. The Dynamic Competition Hypothesis predicts that influential actors in competition networks should have a large number of common out-neighbors with many other nodes. We empirically study this idea as a centrality score and find the measure predictive of importance in several real-world networks including food webs, conflict networks, and voting data from Survivor.

1 Introduction

While social networks are often studied from the perspective of positive interactions such as friendship or followers, the impact of negative social interaction on their structure and evolution cannot be ignored. Structural balance theory posits positive and negative ties between actors in social networks, and assumes such signed networks will stabilize so that triples of actors are either all mutually friends or possess common adversaries; see [12], and [9] for a modern treatment. The prediction of the signs of edges in a social network was previously studied [15,18,21]. Further, negative interactions as a model for edges was studied in the context of negatively correlated stocks in market graphs [4], and in the spatial location of cities as a model to predict the rise of conflicts and violence [11]. Even in the highly cited Zachary Karate club network [22], the negative interaction between the administrator and instructor was the impetus for the split of the club participants into two communities. We propose that competition or negative interactions are critically important to the study of social networks and more broadly, real-world complex networks, and are often hidden drivers of link formation.

In [6], we investigated properties inherent in social networks of competitors that evolve dynamically over time. Such networks are viewed as directed, where a directed edge from nodes u to v corresponds to some kind of negative social interaction. For example, a directed edge may represent a vote by one player for another in a social game such as the television program Survivor. Directed edges are added over discrete time-steps in what we call dynamic competition networks. Our main contribution in [6] was the presentation of a hypothesis, referred to as the Dynamic Competition Hypothesis, or (DCH), that served as a predictive tool to uncover alliances and leaders within dynamic competition

networks. We provided evidence for the hypothesis using U.S. voting record data from 35 seasons of Survivor.

In the present paper, we focus on a particular implication of the DCH. Namely, the DCH predicts that leaders and central actors in these networks should have a large number of common out-neighbors with other nodes in the network. Consequently, this score should constitute a more accurate and interesting centrality score in competition networks where edges have a negative connotation. We study this score in terms of its ranking of leaders in various kinds of networks ranging from additional international seasons of Survivor, to conflict networks, and to food webs.

We organize the discussion in this paper as follows. In Sect. 2, we formally define dynamic competition networks, and review the DCH as stated in [6], with a focus on the common out-neighbor scores, called CON scores. In Sect. 3, we investigate using CON scores as centrality measures in three distinct sources: (i) voting data from all international (that is, non-U.S.) seasons of Survivor, (ii) from conflict networks arising from the countries of Afghanistan, India, and Pakistan, and (iii) in 14 food webs. We find that the CON scores predict influential actors in the networks with high precision. The final section interprets our results for real-world complex networks, and suggests further directions.

We consider directed graphs with multiple directed edges throughout the paper. Additional background on graph theory and complex networks may be found in the book [5] or [7].

2 The Dynamic Competition Hypothesis

The Dynamic Competition Hypothesis (DCH) provides a quantitative framework for the structure of dynamic competition networks. We recall the statement of the DCH as first stated in [6]. Before we state the DCH, we present some terminology.

A *competition network* G is one where nodes represent actors, and there is a directed edge between nodes u and v in G if actor u is in competition with actor v . A *dynamic competition network* is a competition network where directed edges are added over discrete time-steps. For example, nodes may consist of individuals and edges correspond to conflicts between them; as another example, we may consider species in an ecological community, and directed edges correspond to predation. Observe that dynamic competition networks may have multiple edges if there were multiple conflicts; further, not all edges need be present.

The central piece of the DCH we study here are the *common out-neighbor* scores. Without loss of generality, we assume that the node correspond to integers such that we can use the nodes to address an adjacency matrix as well. Consequently, let \mathbf{A} be the adjacency matrix of given competition network. Entries in the matrix are 0 or positive integers for the number of competition interactions. For nodes u , v , and w , we say that w is a *common out-neighbor* of u and v if (u, w) and (v, w) are directed edges. Alternatively, $A_{uw}A_{vw} \geq 1$. For a pair of distinct nodes u, v , we define $\text{CON}(u, v)$ to be the number of common out-neighbors of u and v . Note that this common out-neighbor score counts multiplicities based

on the minimum number of interactions: $\text{CON}(u, v) = \sum_k \min(A_{uk}, A_{vk})$, which corresponds to a multiset intersection. For a fixed node u , define

$$\text{CON}(u) = \sum_{v \in V(G)} \text{CON}(u, v).$$

We call $\text{CON}(u)$ the *CON score* of u . For a set of nodes S with at least two nodes, we define

$$\text{CON}(S) = \sum_{u, v \in S} \text{CON}(u, v).$$

Observe that $\text{CON}(S)$ is a non-negative integer.

In the DCH, *leaders* are defined as members of a competition network with high standing in the network, and edges emanating from leaders may influence edge creation in other actors. In the context of conflict networks within a country, leaders may be actors who exert the strongest political influence within the country; note that these may not be the largest or most powerful actors. As another example, leaders in a food web would naturally have higher trophic levels (that is, higher position in a food chain). The DCH characterizes leaders as those nodes with high CON scores, low in-degree, high out-degree and high closeness. Recall that for a strongly connected digraph G and a node v , we define the *closeness* of u by

$$C(u) = \left(\sum_{v \in V(G) \setminus \{u\}} d(u, v) \right)^{-1}$$

where $d(u, v)$ corresponds to the distance measured by one-way, directed paths from u to v .

In this paper, we focus on the implication that leaders in competition networks should have high CON scores, which suggests this is a natural centrality measure for these networks. The DCH also involves the notion of alliances, that does not factor into our present study. *Alliances* are defined as groups of agents who pool capital towards mutual goals. In the context of social game shows such as Survivor, alliances are groups of players who work together to vote off players outside the alliance. Members of an alliance are typically less likely to vote for each other, and this is the case in strong alliances. This is characterized in terms of *near independent sets*; see [6] for the formalism.

In summary, the *Dynamic Competition Hypothesis* (or *DCH*) asserts that dynamic competition networks satisfy the following four properties.

1. Alliances are near independent sets.
2. Strong alliances have low edge density.
3. Members of an alliance with high CON scores are more likely leaders.
4. Leaders exhibit high closeness, high CON scores, low in-degree, and high out-degree.

Our focus in this work will be on the validation of the DCH with regards to detecting leaders; in particular, we will focus on items (3) and (4) of the DCH. Note that while we expect leaders to be in alliances (that is, have prominent local influence), leaders are determined via global metrics of the network.

3 Methods and Data

3.1 Survivor

In [6], we studied the voting history of U.S. seasons of Survivor, which is a social game show where players compete by voting each other out. In Survivor, strangers called survivors are placed in a location and forced to provide shelter and food for themselves, with limited support from the outside world. Survivors are split into two or more tribes which cohabit and work together. Tribes compete for immunity and the losing tribe goes to tribal council where one of their members is voted off. At some point during the season, tribes merge and the remaining survivors compete for individual immunity. Survivors voted off may be part of the jury. When there are a small number of remaining survivors who are finalists (typically two or three), the jury votes in favor of one of them to become the Sole Survivor who receives a cash prize of one million dollars. Figure 1 represents a graphical depiction of the voting history of a season of U.S. Survivor.

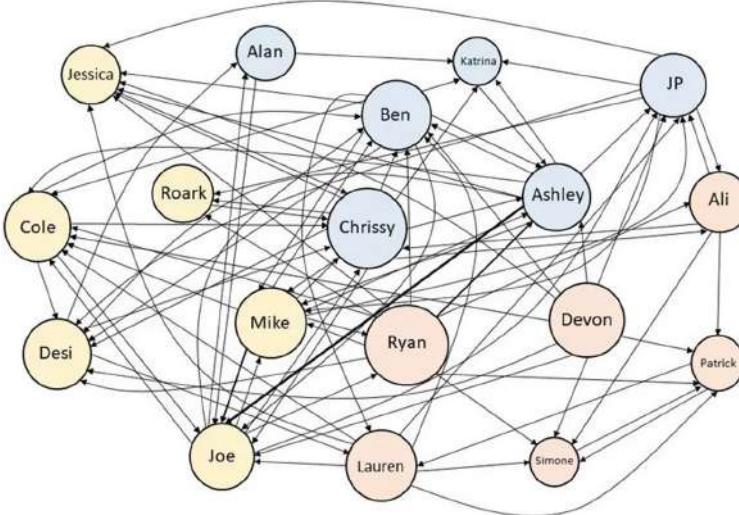


Fig. 1. The Survivor Heroes vs. Healers vs. Hustlers co-voting network. Nodes are scaled by closeness, and color-coded according to their original tribe. Thicker edges represent multiple votes.

Table 1. Three international Survivor seasons. Players are listed by first name, in order from top to bottom with the winner at the top, and the first eliminated player at the bottom. For each player we list the in-degree, out-degree, closeness, and CON score. The horizontal line separates finalists from the rest of the group.

Australian Survivor (2002)					Robinson 2009					Survivor South Africa Malaysia				
Name	ID	OD	C	CON	Name	ID	OD	C	CON	Name	ID	OD	C	CON
Robert	5	10	0.714	44	Ellenor	0	6	0.563	36	Lorette	4	9	0.653	35
Sciona	1	9	0.652	37	Jarmo	7	8	0.557	34	Grant	5	8	0.653	33
Joel	7	8	0.625	35	Anna	2	10	0.645	54	Amanda	4	8	0.622	26
Katie	3	9	0.652	38	Nina	4	7	0.557	38	Mandla	0	8	0.652	32
Sophie	3	8	0.652	38	Erik Bl.	7	9	0.612	47	Angie	6	9	0.688	31
Jane	9	6	0.625	36	Lukas	4	7	0.557	31	Angela	4	6	0.594	28
Lance	8	5	0.577	27	Angela	6	8	0.612	46	Dyke	4	6	0.568	17
Craig	8	8	0.577	18	Ranjit	5	5	0.51	30	Hein	5	4	0.484	22
Naomi	8	7	0.5	25	Christian	3	4	0.51	24	Irshaad	3	5	0.544	25
Caren	10	6	0.5	25	Rafael	5	4	0.49	28	Lisa	11	4	0.484	16
Sylvan	3	5	0.417	30	Erik Bi.	9	5	0.438	26	Rijesh	4	3	0.408	13
Deborah	4	4	0.395	23	Erik R.	5	4	0.422	18	Nichal	6	2	0.363	12
Jeff	5	1	0.395	4	Mika	5	3	0.306	13	Elsie	8	2	0.436	9
David	6	3	0.441	23	Josefine	0	2	0.265	17	Viwe	5	2	0.344	11
Tim	4	2	0.294	10	Erika	7	2	0.306	15	Nicola	5	1	0.304	6
Lucinda	8	1	0.0	7	Beatrice	6	2	0.35	12	Nomfundo	4	1	0.335	8
					Micha	12	1	0.299	7					

We extend the analysis of the 35 U.S. seasons in [6] to 82 international seasons of Survivor. Data used in our analysis was obtained from the Survivor wiki pages https://survivor.fandom.com/wiki/Main_Page. Several seasons (beyond the 82) were excluded for varying reasons. In some cases, a wiki page exists, but there was no voting data. In other cases, much of the voting information was missing, or the rules are significantly different than the traditional version of the game shows. Nevertheless, the number of seasons collected exceeds the number in [6].

In Table 1 we display some of the CON scores for a few example seasons. We distinguish which players are finalists, since the rules change in determining who is the last player eliminated. For example, instead of eliminating the last player via votes *against* players, in survivor many players may return for a final vote *for* who they would like to win.

Table 2. Statistics on international Survivor seasons.

		CON	PageRank	Jaccard similarity	Random set
Survivor	Top 3	57.3	43.9	47.6	11.1–27.3
	Top 5	81.7	78.0	72.0	18.5–45.5

In Table 2, we detail relevant statistics on these networks. For each network, we consider whether the winner of the season had one of the top three or top five CON scores and list the percentage of such networks. For example from Table 2 we see that 81.7% of Survivor winners had one of the largest 5 CON scores. For comparison, we also compute PageRank (on the *reversed-edge* network, where we change the orientation of directed edge) and Jaccard similarity scores, which are both standard ranking scores. Jaccard similarity is a type of normalized CON score; see [10]. We find that the CON scores are a more accurate predictor for determining finalists of Survivor than both PageRank and Jaccard similarity. We observe that these results are consistent with the analysis in [6] for the U.S. seasons. As an added comparison, we list the probability of the winner appearing in a random set of three or five players; note that there is a range of percentages depending on how many players are in a given season. In the interest of space, we refer the reader to <https://eikmeier.sites.grinnell.edu/uncategorized/competition-show-data/> where we house all data on these seasons.

3.2 Political Conflicts

For our second competition network, we extracted data from *The Armed Conflict Location and Event Data Project* (or ACLED), which may be found at <https://www.acleddata.com/data/>. ACLED catalogs information about political conflict and protests across the world. In these *conflict networks*, nodes correspond to actors in a given region, and edges correspond to conflicts between the actors. Many types of metadata are recorded corresponding to each event. Our particular interest is in the actors involved in each event, and where the event took place. More information about this project can be found on the project website.

An important note about the ACLED data is that we do not know which actor *initiated* a given event. Therefore, we do not consider the majority of edges (events) to be directed. The only events which we assume knowledge about directed-ness is when civilians are involved. We restricted our study to a set of events to a particular country; keeping the scale at the country level allows us to keep a larger set of actors. We selected three countries that have a large number of actors and events to analyze: Afghanistan, India, and Pakistan.

We first consider the rankings for Pakistan with commentary.

Pakistan has faced terrorism activities since 2000, with many militant groups attacking civilians and Pakistan armed forces. TTP (Pakistan) is one of the largest radical extremist groups, which is an umbrella organization of many militant groups such as Lashkar-e-Islam, Islamic State (Pakistan), and Jamaat-ul-Ahrar. In Fig. 2, we find that TTP has one of the highest CON scores. TTP has alliances with another terrorist organizations in Pakistan and neighboring countries, which lends to its prominence. In addition, due to the Afghan war, TTP has a strong influence and hold over many Islamic institutions in Pakistan. The Police Forces of Pakistan and Military Forces of Pakistan ensure national security, and they share information for achieving their goals. The Police Forces of Pakistan are an influential actor in the conflict network with another one of the highest CON scores. They perform their duties in all provinces of Pakistan

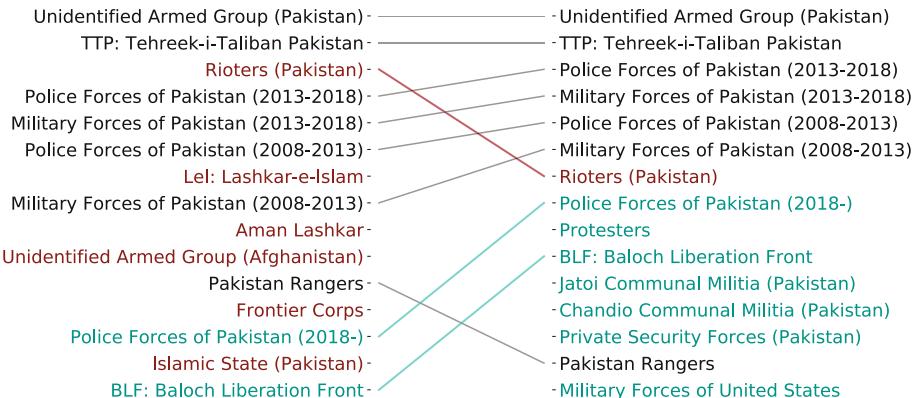


Fig. 2. A Slope Graph to compare the rankings via CON and PageRank. On the left, the top actors in the conflict network Pakistan via CON metrics, while on the right, the top actors in Pakistan via PageRank on the reverse network. Actors are labeled in black if the difference in rankings is less than or equal to three. Actors are labeled in red if the CON ranking is at least four places higher than the PageRank, and in green if the PageRank is at least 4 places higher than the CON ranking. Note that no line appears to connect the left and right side if the actor does not show up in the top 15 of the other ranking.

with the help of their paramilitary forces such as Pakistan Rangers and Frontier Corps, and they maintain law and order, as well as border control. Military Forces of Pakistan (2013–2018) has one of the largest CON scores owing to their increased activities against terrorist groups in recent years.

We also offer commentary on some of the lower ranked actors. The Baloch Liberation Front (or BLF) is an ethnic-separatist political front and militant organization that is currently fighting against the Pakistani government for an independent Balochi state. The BLF is the strongest and most influential militant group of Baluchistan, but there has been no confirmed coordination between the BLF and other Balochi and non-Balochi groups, and they operate independently of one another. This is a large reason that BLF have low CON and closeness scores. The Islamic State is a part of the militant Islamist group: Islamic State of Iraq and Levant (ISIS). The Islamic State was formed by some of the TTP leaders and is more successful in Afghanistan. This organization has had less success in Pakistan largely carrying out isolated, small scale attacks. The Police Forces of Pakistan actively participated with the support of paramilitary forces of Pakistan in 2008–2018 for war against terror. The Police Forces of Pakistan mostly work to maintain the daily law and order in their respective provinces. Likely for these reasons, they have lower CON scores than the years between 2008–2018.

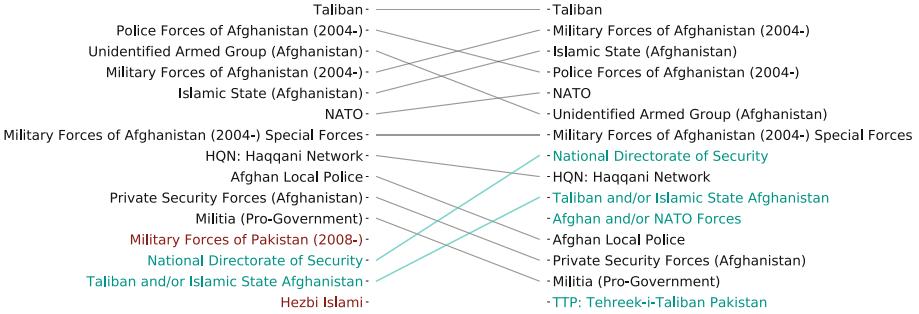


Fig. 3. Top actors in Afghanistan via CON score and PageRank.

We note that the ranking of the top actors using the CON score (on the left in Fig. 2) is not dissimilar to the one using PageRank on the reversed-edge network (on the right in Fig. 2). To quantify the difference in the rankings we used Spearman's rank correlation coefficient. Note that we cannot use Pearson correlation because our data is not at all Gaussian. Recall that Spearman's correlation coefficient is defined as

$$1 - \frac{6 \sum_{i=1}^N d_i^2}{N(N^2 - 1)},$$

where N is the total number of actors, and d_i is the difference in rankings between actor i . A value close to 1 means that the two rankings are very well positively correlated. The Spearman correlation for Pakistan is -0.341 , which suggests that the rankings are not that similar. In fact, the negative value implies that as the CON ranking decreases, the PageRank score increases. There are 741 total actors we consider in the Pakistan data set, and the later rankings clearly vary greatly.

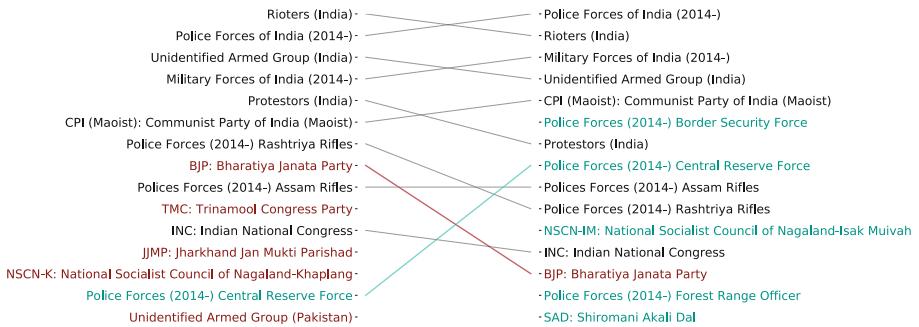


Fig. 4. Top actors in India via CON score and PageRank.

We finish this section with the rankings for Afghanistan and India in Figs. 3 and 4. The Spearman coefficients are 0.604 and -0.267 respectively, indicating

that the rankings provided by CON matches more similarly to PageRank in the Afghanistan dataset. While we do not provide in-depth commentary on these rankings, we find influential actors in both countries with the largest rankings against DCH metrics.

3.3 Food Webs

As a third and final type of data that we analyzed against the DCH, we studied food web datasets from the Pajek website: <http://vlado.fmf.uni-lj.si/pub/networks/data/bio/foodweb/foodweb.htm> [3]. These are 14 food webs in total. In food webs, the nodes are species, and a weighted edge (u, v) exists with weight w if u inherits carbon from v (that is, u preys on v) [2]. We interpret this as a directed negative interaction from node u to node v . A noteworthy difference in these networks (vs. Survivor, say) is that the movement of energy is *balanced*, meaning the in-degree and out-degree for each species is equal.

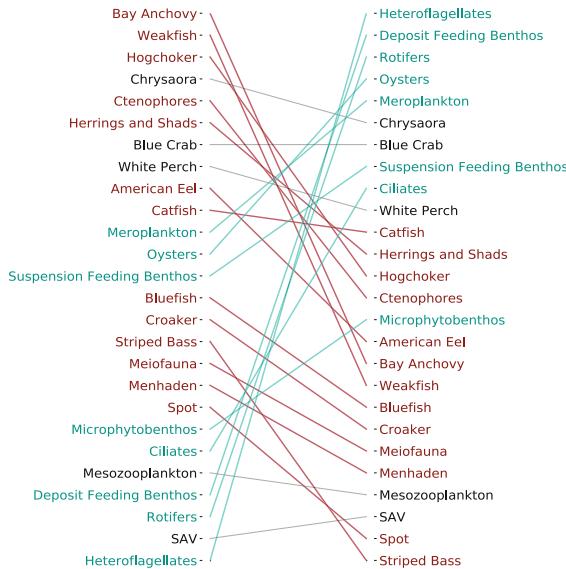


Fig. 5. The Chesapeake Bay Lower food web dataset. On the left, organisms are listed in decreasing order, with the largest CON score at the top. On the right, organisms are listed by decreasing PageRank score.

Rankings of selected food web datasets are in Figs. 5 and 6; the rankings for all the datasets may be found at <https://eikmeier.sites.grinnell.edu/uncategorized/competition-show-data/> along with the computed CON scores, closeness, and PageRank on the reversed-edge network.

In studying the rankings of these 14 food webs, we see a difference between the CON rankings and PageRank. PageRank has been used to study the importance of species in regards to co-extinction [1, 14, 16], which we expect is likely

reflected in the rankings we see here using vanilla PageRank. However, we find a substantially different ranking when using the CON scores; for example, see the placement of Heteroflagellates in Fig. 5. The average Spearman correlation coefficient across these 14 datasets is 0.271, and the range is between 0.004 and 0.554. (Recall that a value close to 1 means very well correlated.) Therefore, we suggest that the CON scores are giving a *different* ranking, which is much closer to *trophic* levels of species. In particular, the CON scores reflect a natural hierarchical structure in ecosystems, and this is consistent with the DCH.

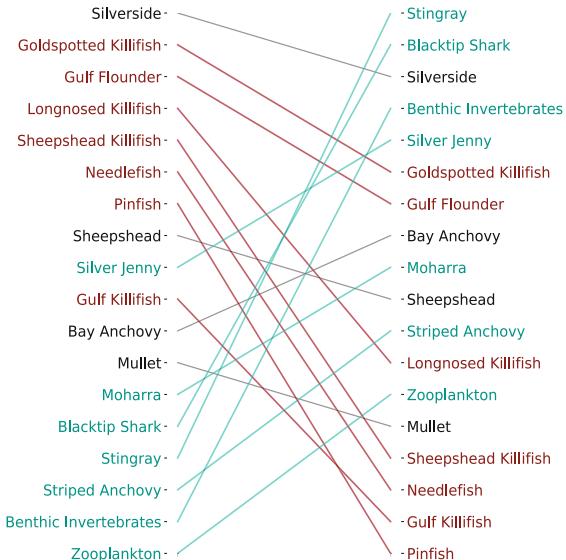


Fig. 6. The CrystalC food web dataset. On the left, organisms are listed in decreasing order, with the largest CON score at the top. On the right, organisms are listed by decreasing PageRank score.

4 Conclusion and Future Directions

We studied an implication of the Dynamic Competition Hypothesis (DCH) for competition networks across several different types of real-world networks. We found that the DCH prediction that high CON scores should correspond to leaders is supported in predicting winners in international seasons of Survivor, in predicting species with high trophic level species in food web, and for determining influential actors in conflict networks in Afghanistan, India, and Pakistan. Metrics such as CON scores outperformed PageRank as an indicator of influential actors in the competition networks we studied.

While our results provide support for the DCH, more work needs to be done. We did not address items (1) and (2) of the DCH regarding alliances in our data

sets, and that would be an important next step. Another direction is to consider an aggregate score, based on the CON score, closeness, and in- and out-degree, as a measure of detecting leaders in competition networks. An interesting direction would be to study more closely the dynamic aspects of competition networks, analyzing them over time to predict leaders. For example, we could analyze the co-voting network of Survivor of each episode of a season, and determine if temporal trends in network statistics predict finalists.

An open question is whether CON score centrality is applicable to large-scale networks exhibiting adversarial interactions, such as in Epinions and Slashdot (which give rise to signed data sets with tens of thousands of nodes, and available from [13]). Epinions was an on-line consumer review site, where users could trust or distrust each other. Slashdot is a social network that contains friend and foe links. A challenge with these data sets from the view of validating the DCH is that there is no inherently defined ranking, as there is in Survivor (via the order contestants were voted off), food webs (trophic level), and in conflict graphs (via political and strategic influence).

Acknowledgments. The research for this paper was supported by grants from NSERC and Ryerson University. Gleich and Eikmeier acknowledge the support of NSF Awards IIS-1546488, CCF-1909528, the NSF Center for Science of Information STC, CCF-0939370, and the Sloan Foundation.

References

1. Allesina, S., Pascual, M.: Googling food webs: can an eigenvector measure species' importance for coextinctions? *PLoS Comput. Biol.* **59**, e1000494 (2009)
2. Baird, D., Ulanowicz, R.E.: The seasonal dynamics of the Chesapeake Bay ecosystem. *Ecol. Monogr.* **594**, 329–364 (1989)
3. Batagelj, V., Mrvar, A.: Pajek food web datasets. <http://vlado.fmf.uni-lj.si/pub/networks/data/>
4. Boginski, V., Butenko, S., Pardalos, P.M.: On structural properties of the market graph. In: *Innovation in Financial and Economic Networks*, Edward Elgar Publishers, pp. 29–45 (2003)
5. Bonato, A.: *A Course on the Web Graph*. American Mathematical Society Graduate Studies Series in Mathematics, Rhode Island (2008)
6. Bonato, A., Eikmeier, N., Gleich, D.F., Malik, R.: Dynamic competition networks: detecting alliances and leaders. In: *Proceedings of Algorithms and Models for the Web Graph (WAW 2018)* (2018)
7. Bonato, A., Tian, A.: Complex networks and social networks. In: Kranakis, E. (ed.) *Social Networks. Mathematics in Industry Series*. Springer, Berlin (2011)
8. Brandes, U., Erlebach, T. (eds.): *Network Analysis: Methodological Foundations*. LNCS 3418. Springer, Berlin (2005)
9. Easley, D., Kleinberg, J.: *Networks, Crowds, and Markets Reasoning about a Highly Connected World*. Cambridge University Press, Cambridge (2010)
10. Gower, J.C., Warrens, M.J.: Similarity, dissimilarity, and distance, measures of, Wiley StatsRef: Statistics Reference Online (2006)
11. Guo, W., Lu, X., Donate, G.M., Johnson, S.: The spatial ecology of war and peace, Preprint (2019)

12. Heider, F.: The Psychology of Interpersonal Relations. John Wiley & Sons, Hoboken (1958)
13. Leskovec, J.: The Stanford large network dataset collection. <http://snap.stanford.edu/data/index.html>
14. McDonald-Madden, E., Sabbadin, R., Game, E.T., Baxter, P.W.J., Chadès, I., Possingham, H.P.: Using food-web theory to conserve ecosystems. *Nat. Commun.* **7**, 10245 (2016)
15. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web (WWW 2010) (2010)
16. Stouffer, D.B., Bascompte, J.: Compartmentalization increases food-web persistence. *Proc. Nat. Acad. Sci.* **1089**, 3648–3652 (2011)
17. Survivor Wiki. http://survivor.wikia.com/wiki/Main_Page
18. Tang, J., Chang, S., Aggarwal, C., Liu, H.: Negative link prediction in social media. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM 2015) (2015)
19. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998)
20. West, D.B.: Introduction to Graph Theory, 2nd edn. Prentice Hall, New Jersey (2001)
21. Yang, S-H., Smola, A.J., Long, B., Zha, H., Chang, Y.: Friend or frenemy?: predicting signed ties in social networks. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2012) (2012)
22. Zachary, W.W.: An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977)



Investigating Saturation in Collaboration and Cohesiveness of Wikipedia Using Motifs Analysis

Anita Chandra^(✉) and Abyayananda Maiti

Department of Computer Science and Engineering, Indian Institute of Technology Patna,
Patna 801103, Bihar, India
{anita.pcs15,abyaym}@iitp.ac.in

Abstract. Wikipedia is a multilingual encyclopedia that works on the idea of virtual collaboration. Initially, its contents such as articles, editors and edits grow exponentially. Further growth analysis of Wikipedia shows slowdown or saturation in its contents. In this paper, we investigate whether two essential characteristics of Wikipedia, collaboration and cohesiveness also encounter the phenomenon of slowdown or saturation with time. Collaboration in Wikipedia is the process where two or more editors edit together to complete a common article. Cohesiveness is the extent to which a group of editors stays together for mutual interest. We employ the concept of network motifs to investigate saturation in these two considered characteristics of Wikipedia. We consider star motifs of articles with the average number of edits to study the growth of collaboration and 2×2 complete bicliques or “butterfly” motifs to interpret the change in the cohesiveness of Wikipedia. We present the change in the count of the mentioned network motifs for the top 22 languages of Wikipedia upto May 2019. We observe saturation in collaboration while the linear or sudden rise in cohesiveness in most of the languages of Wikipedia. We therefore notice, although the contents of Wikipedia encounter natural limits of growth, the activities of editors are still improving with time.

Keywords: Natural limits of growth · Bipartite networks · Network motifs · Wikipedia

1 Introduction

The advancement in Web 2.0 allows web users to interact, organize, generate and collaborate on the Internet. Its one of the most valuable and successful consequences is Wikipedia and that is now the fifth most visited website¹. Wikipedia is an online multilingual encyclopedia that permits any web user to create, read, share, delete and edit articles. In 2001, it was introduced only in English edition while it is available in 294 active languages² at present. Several registered editors, anonymous editors and automated bots are consistently improving the contents of Wikipedia by editing the existing articles. All the activities including recent changes performed by editors are recorded as history. Numerous studies carried out in the past [1, 16] show tremendous initial growth

¹ https://en.wikipedia.org/wiki/List_of_most_popular_websites.

² https://en.wikipedia.org/wiki/List_of_Wikipedias.

in Wikipedia with an exponential increase in the number of articles, editors, edits and views. Further, researchers in [6, 7, 9, 15] found that growth in its contents is slowing down or getting almost saturated. Some of the possible reasons reported for the saturation are unfriendly behaviors towards newly joined editors where their edits are likely to be reverted by experienced editors [7], increased overhead costs of coordination and production, and Wikipedia has probably reached the natural limits of growth [6]. This phenomenon of slowdown or saturation in growth is also discussed in Wikipedia³ itself. In this Wiki page, the previous exponential growth model is revised and two distinct novel growth models *i.e.*, Gompertz function and modified Gompertz function are introduced to characterise the saturation in Wikipedia. The Gompertz function model predicts that the contents grow and ultimately asymptotically approach zero. Its modified model states that content increases continuously but diminishes significantly in the later phase of development of Wikipedia networks.

In [10], authors have discussed eighteen important characteristics for effective regulation of teamwork. Out of these eighteen characteristics, they have mentioned the significance of collaboration and cohesiveness in teamwork in detail. Similarly in [2], the authors have outlined essential characteristics for effectual teamwork of multidisciplinary scientific collaborators. They discussed features for proper regulation and production of this teamwork which includes intra-cohesion, collaboration, communication and distribution of mutual roles and responsibilities. In [18], the authors accessed cooperation/collaboration in Wikipedia by correlating article quality and the number of edits. Thus, we can comprehend the significance of collaboration and cohesiveness in Wikipedia as a consequence of efficient teamwork.

In this paper, we investigate whether collaboration and cohesiveness of Wikipedia show slowdown or saturation with time. Though Wikipedia has several characteristics, we believe collaboration and cohesiveness to be important ones. We use the concept of network motifs to examine saturation in these two characteristics of Wikipedia. Network motifs are recurrent small subgraphs with specific interconnection patterns present in the network. We consider star motifs of articles with the average number of edits to illustrate the growth of collaboration in Wikipedia. Whereas to study the change in cohesiveness, we calculate 2×2 complete bicliques or “butterfly” (χ) motifs. This “butterfly” (χ) motif is the simplest cohesive unit of the bipartite network. The aforementioned motifs can also be used to explore other characteristics of Wikipedia. In particular, we investigate two research questions in this study:

- **Q1.** Does collaboration in Wikipedia show slowdown or saturation with time?
- **Q2.** Does cohesiveness in Wikipedia show slowdown or saturation with time?

The rest of the paper is organized as follows: we discuss related works in Sect. 2. In Sect. 3, we briefly present the algorithms which we use to calculate the network motifs. Section 4 provides results and discussions on the decline or saturation phenomenon. At last, we conclude our paper and discuss some future works in Sect. 5.

2 Related Work

In this paper, we aim to investigate whether collaboration and cohesiveness show slowdown or saturation in several wikipedias. We present related works to various aspects of

³ https://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia%27s_growth.

our present work: (i) Wikipedia growth analysis; (ii) Wikipedia growth saturation; (iii) Network motifs calculation in bipartite networks; (iv) Network motifs applications.

Wikipedia Growth Analysis: The examination of statistical properties and the growth of directed Wikigraph is presented in [3], where nodes are represented as articles and edges as hyperlinks. The authors observed scale-invariant property for both in and out hyperlink distributions which were characterized by local rules such as preferential attachment mechanism. Although editors are generally responsible for the growth of Wikipedia. A growth model for Wikigraph is proposed in [20] which incorporates the concept of preferential attachment and information exchange via reciprocal arcs. The authors achieve a good fit between in-link distributions from the model and the real wikipedias by extracting a few parameter values of the model. Another growth model was proposed for the Wikipedia network by Gandica *et al.* [5] where it was described as an editor-editor network. According to it, the probability of an editor editing an article is proportional to the number of edits she already has (*i.e.*, Preferential attachment), her fitness and her age.

Wikipedia Growth Saturation: Gibbons *et al.* [6] discovered slowdown or saturation in the growth of different languages of Wikipedia. The authors proposed the Low-Hanging Fruit hypothesis to explain the reasons behind the slowdown. This hypothesis states that larger the site becomes and more the knowledge it contains, the more difficult it becomes for editors to make a novel, lasting contributions. They mentioned that the editors might have already delivered a lot on common topics and leaving complicated topics that require more time and effort. Similarly in [15], the authors reported a decline in the number of articles and editors. They measured growth, population shifts and patterns of the editor as well as administrator activities to show the slowdown or saturation. Their study also mentioned possible reasons for this sudden decline in Wikipedia: (a) frequent edit reverts of occasional editors (b) more cost overhead involved in coordination and bureaucracy (formulating and discussing policies) (c) maintaining quality of the tools used by editors and administrators. Kittur *et al.* [9] show the growth of conflicts and coordination in Wikipedia at the global, article and user levels. This paper reported a decline in production of administrators as given statistics showed that administrators performed only around 10% of edits in 2006 whereas their contributed edits were 50% in 2003.

Network Motifs Calculation in Bipartite Networks: Network motif is one of the significant properties of the networks. Simmons *et al.* [14] proposed an algorithm “Bmotif” for ecological bipartite networks to determine the count of 44 motifs consisting of six species and 148 positions of these 44 motifs. Wang *et al.* [17] calculate rectangles or 4-cycles (\boxtimes) in a bipartite graph. The authors have proposed three different algorithms which are applied to different size of datasets and available computational resources. The efficiency of their algorithms is verified over large real-world and synthetic bipartite networks. In [12], the authors proposed an efficient ExactBFS algorithm for the calculation of “butterfly” (\boxtimes) motifs by enhancing the previously proposed algorithm in [17]. The algorithm starts the calculation of \boxtimes from the partition whose sum of degree squares is more and this reduced its computational time dramatically. They presented a suite of randomized algorithms that can quickly approximate the number of butterflies in very large bipartite networks.

Network Motifs Applications: In [8], the authors represent star motifs of an article with different types of editors (registered, anonymous, bots and administrators) with distinct revisions types (add, delete, edit and revert). Applying above-defined motifs, they classified pages as combative or cooperative and also understand dynamics of editors behaviors to study the growth of Wikipedia. They also showed the emergence of slowdown or saturation in contents of Wikipedia using the generative model where these defined temporal motifs are used as features. In [13], the authors presented evidence of the financial crisis that happened in 2007 and the economic recession of 2008–2009 through temporal variations in motifs structures of bipartite World Trade Web (WTW). In [11], the authors revealed that structural features like motifs are sufficient to detect strong or weak ties in Twitter with high precision. In [19], the authors used motif structures of geo-tagged photos of Flickr to design travel recommender, which assists tourists in planning their trips more efficiently.

3 Network Motifs Calculation

3.1 Datasets Description and Filtration:

We have downloaded datasets of the top 25 languages of Wikipedia based on the count of articles from Wiki dumps⁴. It contains various information such as a full history of the articles, the log event of all the editors and articles, *etc.*. All these information are available in different file formats such as XML, HTML, SQL and JSON. In our study, we have not considered Cebuano, Waray and Serbo-Croatian wikipedias because their size (*i.e.*, count of articles) is too small. We parse the downloaded XML files to extract editors ID, articles ID and timestamps at which the editors have done revisions. We have taken all wikipedias from their date of publishing to May 2019 (31/05/2019). We discard all the multiple edits of editors and edits performed by automated bots on articles while we retain the edits of anonymous editors. Multiple edits are more than one edits performed on the same article by the same editor. The ExactBFS algorithm proposed in [12], which we apply to compute “butterfly” (X) motifs, drops the multiple edits. Anonymous or unregistered editors are Wiki contributors whose IP addresses of their machines are recorded. In our previous paper [4], we provided % of anonymous editors and their contributed edits for the top 20 wikipedias. These statistics showed that more than 75% of edits are performed by anonymous editors in all wikipedias. Since the quantity of their contributions is too large, discarding them will lead to a huge loss of data. Bots are automated tools that perform repetitive tasks to maintain the articles of Wiki. Bots perform numerous edits at a time which distort the original network of edits. Thus, we discard contributions of all the registered bots available in Wiki page⁵.

3.2 Natural Limits in Growth of Wikipedia:

In several research papers [6, 7, 9, 15], researchers noticed slowdown or saturation in the contents of Wikipedia after its enormous initial growth. Saturation in the contents

⁴ <https://dumps.wikimedia.org/>.

⁵ https://en.wikipedia.org/wiki/Category>All_Wikipedia_bots.

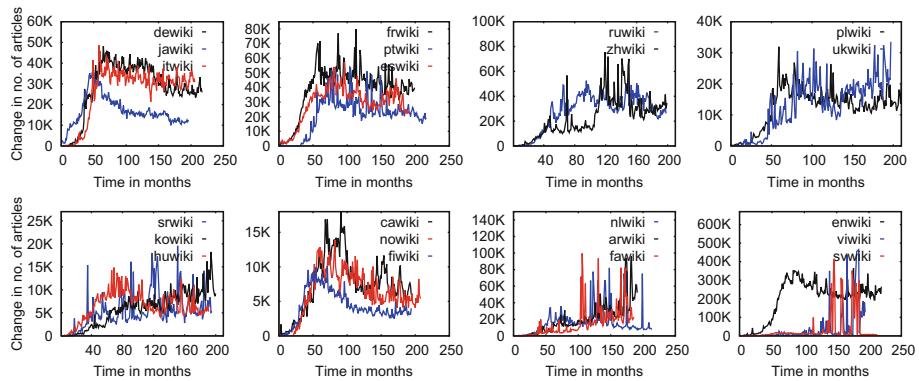


Fig. 1. Changes in the number of articles added for the top 22 languages of Wiki upto May 2019. Here, *x*-axis shows time in months and *y*-axis gives change in no. of articles.

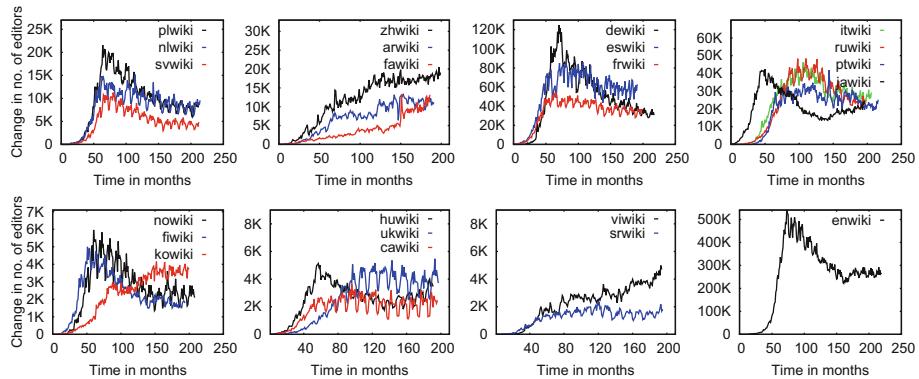


Fig. 2. Changes in the number of editors arrived for the top 22 languages of Wiki upto May 2019. Here, *x*-axis shows time in months and *y*-axis gives change in no. of editors.

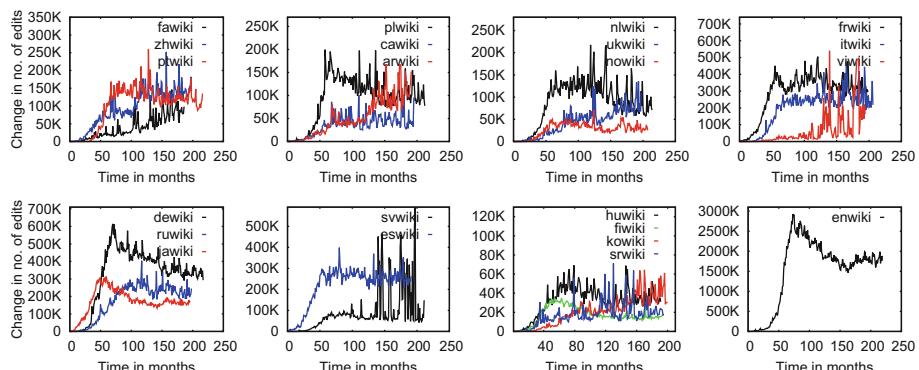


Fig. 3. Changes in the number of edits performed for the top 22 languages of Wiki upto May 2019. Here, *x*-axis shows time in months and *y*-axis gives change in no. of edits.

of Wiki means the rate of creation of articles, arrival of editors and their edits asymptotically approach zero. However, none of the above mentioned research showed natural limits of growth in all the contents such as articles, editors and edits of several languages of Wiki. Moreover, the growth patterns were only reported upto 2009 which depicted decline followed by saturation in English Wiki from 2007. Hence, in Figs. 1, 2 and 3, we present the change in the count of all contents of the top 22 languages of Wiki upto May 2019 (current year). We drop the outliers occurring due to sudden jump (eg., auto or bulk posting) or fall (eg., server performance issues) in the contents of Wikis. In Fig. 1, we present the change per month in the count of articles for top 22 languages of Wiki. We observe from Fig. 1 that, in Arabic, Persian, Korean, Serbian, Swedish, Ukrainian and Vietnamese Wikis, the change in the count of articles shows no saturation, which means it is still growing. Whereas, in the case of the remaining 15 Wikis such as English, German, Japanese, Italian, French, Portuguese, Spanish, Chinese, Russian, Polish, Hungarian, Catalan, Norwegian, Finnish and Dutch Wikis, the change patterns decrease and then get almost plateaued. Although Swedish Wiki is 2nd and Vietnamese Wiki is 10th largest developed Wikis, we have not yet noticed decline or saturation in their contents. Instead, we notice several big spikes that happen due to auto or bulk posting behaviors of editors. In Fig. 2, we can see that, in all the considered Wikis except Persian Wiki and Vietnamese Wiki, the change in the count of editors either gets saturated or declines followed by saturation. Similarly in Fig. 3, besides Arabic, Persian, Korean, Serbian and Vietnamese Wikis, all other 17 Wikis show slowdown followed by saturation or saturation in the change of the count of edits. Interestingly, we notice in the case of Arabic, Korean and Serbian Wikis that the change in editors' count saturates while the count of articles and their corresponding edits keeps on growing with time. Thus, from Figs. 1, 2 and 3, we encounter slowdown or saturation in all the contents in most of the considered languages of Wiki.

3.3 Network Motifs Calculation

We intend to investigate whether collaboration (union) and cohesiveness (strength) in Wikipedia encounter slowdown or saturation. In Fig. 4(a), we provide a star motif of an article with edits that represent a collaborative unit where two or more editors edit the same article. We assume that all revisions performed on an article are edits. However these revisions can be of different types such as add, delete and revert other than edit. We consider the cohesive unit as 2×2 complete biclique or “butterfly” (\boxtimes) motif shown in Fig. 4(b). In this motif, a set of two editors edit the same pair of articles. We calculate this “butterfly” (\boxtimes) to study saturation in the growth of cohesiveness in the networks of Wikipedia. To examine the cohesiveness of editors, we calculate the butterfly per editor ($\overline{\chi}_e$) as shown in Fig. 4(c), i.e., the number of butterflies in which an editor e is present. We can see from Fig. 4(c) that editors e_1 , e_2 and e_3 are present in 2, 2 and 1 butterflies respectively. We calculate the count of mentioned motifs to investigate saturation in the growth of characteristics of several wikipedias.

The collaboration of an article in Wikipedia is stated as collaborative edits performed by a random set of editors for a while⁶. Thus, to study the growth of

⁶ <https://en.wikipedia.org/wiki/Wikipedia:Collaborations>.

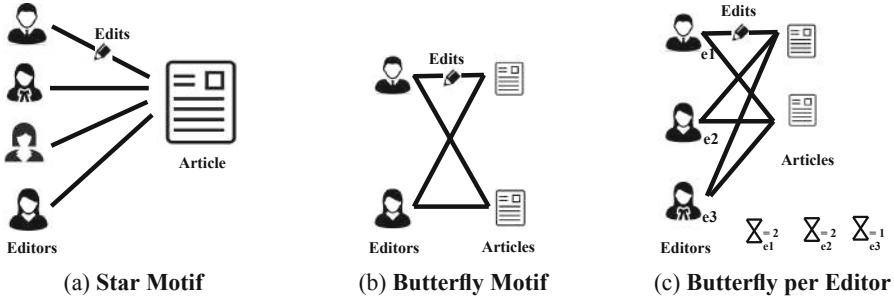


Fig.4. (a) Star Motif ($>$) : collaborative units of network (b) Butterfly Motif (\times) : cohesive units of network (c) Butterfly per Editor ($\overline{\times}_e$) : cohesive units of editors

collaboration in Wiki, we consider the star motif of articles with an average number of edits, as shown in Fig. 4(a). We consider these star motifs of articles with an average number of edits as collaborative units because they represent the whole network. We calculate butterfly motifs (\bar{X}) as shown in Fig. 4(b) to examine the growth of cohesiveness in Wiki networks. To calculate these cohesive units (\bar{X}), we apply an efficient ExactBFS algorithm proposed in [12]. This ExactBFS algorithm is almost same as the WFC algorithm proposed in [17]. The only difference is the selection of partition of a bipartite network to which the algorithm starts. This constraint of ExactBFS algorithm reduces its computational time dramatically. The authors reported in [12] that it is 700 times faster than WFC for a Journal network. Further, we also examine the cohesiveness of editors by calculating butterfly per editor (\bar{X}_e), as shown in Fig. 4(c). We refer vBFS algorithm proposed in [12] to calculate \bar{X}_e . According to vBFS algorithm, it selects any editor e from a set of editors and then applies ExactBFS algorithm on it. If we calculate \bar{X}_e for all the editors of any Wiki network, then its computational cost increases much. Thus, we again refer the VSamp algorithm introduced in the same paper [12], which gives the approximated count of \bar{X} in a network. In this algorithm, first of all, it uses random sampling to select a set of random editors and then apply vBFS algorithm on it. In [12], the authors presented the theoretical proof using Chebyshev's inequality to show that the variance of the VSamp estimator is not much as compared to ExactBFS. The source code for all the referred algorithms is publicly available on Github⁷. Further, we see how the change in the count of $<$, \bar{X} and \bar{X}_e helps us to understand the growth of collaboration and cohesiveness in Wikipedia.

4 Results and Analysis

In this section, we investigate saturation in the growth of collaboration and cohesiveness in Wikipedia using the concept of network motifs.

⁷ <https://github.com/beginner1010/butterfly-counting>.

Table 1. Names and codes, average edit values, sample size, number of editors and edits in the considered sample of top 22 languages of Wikipedia upto May 2019.

Wikipedia languages		Collaboration		Cohesiveness		
Names	Codes	Average Number of Edits (AE)	AE Values	Sample size (%)	Editors Count	Edits
English	enwiki	2× AE	32	01	12,94,619	33,57,425
Swedish	svwiki	3× AE	12	50	6,32,986	98,84,645
German	dewiki	1× AE	25	10	16,53,891	70,98,438
French	frwiki	1× AE	16	10	10,83,748	57,38,813
Dutch	nlwiki	2× AE	20	50	10,26,956	98,23,203
Russian	ruwiki	2× AE	28	50	23,15,038	1,56,43,104
Italian	itwiki	2× AE	28	30	21,01,851	1,23,10,465
Spanish	eswiki	1× AE	16	10	15,55,896	41,98,135
Polish	plwiki	2× AE	28	50	12,14,858	1,01,21,072
Vietnamese	viwiki	3× AE	6	70	3,24,170	1,39,64,610
Japanese	jawiki	1× AE	19	30	18,54,540	1,01,85,714
Portuguese	ptwiki	2× AE	16	50	11,65,331	1,09,01,488
Chinese	zhwiki	2× AE	16	50	14,10,497	98,13,450
Ukrainian	ukwiki	3× AE	24	70	3,91,701	60,35,474
Persian	fawiki	3× AE	12	70	5,45,331	55,90,990
Catalan	cawiki	3× AE	30	70	2,78,293	50,87,221
Arabic	arwiki	3× AE	15	70	9,87,574	71,55,738
Serbian	srwiki	1× AE	6	70	1,85,836	25,80,431
Norwegian	nowiki	2× AE	18	70	4,03,159	42,76,638
Finnish	fiwiki	2× AE	22	70	5,69,509	48,09,653
Hungarian	huwiki	1× AE	12	70	3,71,512	42,81,402
Korean	kowiki	2× AE	18	70	5,47,605	58,93,577

4.1 Investigating Saturation in the Collaboration of Wikipedia

In Fig. 5, we provide the change per month in the count of star motifs upto May 2019 for the top 22 languages of Wikipedia. We provide codes for all the Wikis in Table 1, which we use further in result discussions. In Table 1, we report integral values of the average number of edits of articles for all the considered Wiki networks. However, we consider star motifs of articles with twice/thrice the number of average edits where the edit distributions of articles are highly right skewed. Skewed edit distribution means large number of articles have smaller number of edits and less number of articles are heavily edited. We consider the star motif of articles with only the average number of edits even for the skewed networks when its size is too small *e.g.*, srwikipedia and huwiki. We include multiple edits of articles to study the growth of collaboration. We discard the outliers to get discernible change patterns of star motifs. Outliers are the changes having the counts either too large or too small. The change in collaborative units can also be negative as its absolute count values can decrease with time.

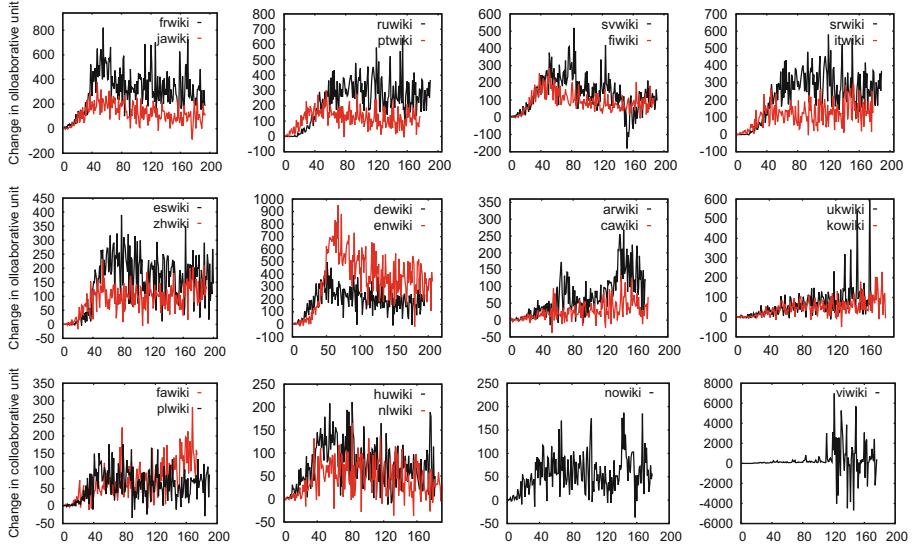


Fig. 5. Changes in the count of star motifs for top 22 languages of Wiki. Here, x -axis has time in months and y -axis gives change in collaborative units.

In Fig. 5, we notice three different trends such as saturated, not saturated and decline followed by saturation in the change in the count of star motifs of articles. We notice from Fig. 5 that the change in star motifs is not saturated and still growing in the case of arwiki, cawiki, fawiki, kowiki, ukwiki and viwiki. Interestingly, we notice a peculiar patterns in viwiki where the change encounters multiple big spikes (sudden surge in collaborative units) after a decade. These patterns show inconsistencies in collaboration and rather they present burst editing behaviors of editors. While in the rest of 16 Wikis, the change after exponential growth gets either saturated or declined followed by saturation. For instance, we notice in the case of itwiki, nowiki, ruwiki, srwiki, nlwiki and zhwiki that change in star motifs is getting saturated with time. Whereas in dewiki, enwiki, eswiki, fiwiki, frwiki, huwiki, jawiki, plwiki, ptwiki and svwiki, the change manifests slowdown and then gets saturated. Interestingly in Fig. 5, we can clearly notice more negative values in declined and saturated parts of the growth patterns of collaboration. Thus, we observe that collaboration also shows slowdown or saturation phenomenon in most of the considered languages of Wikipedia.

4.2 Investigating Saturation in the Cohesiveness of Wikipedia Networks

In this section, we investigate saturation in the cohesiveness of networks and editors of wikipedias with the help of butterfly (χ) motifs. First, we examine cohesiveness in Wiki networks and then present it for the editors.

Saturation in Cohesiveness of Wikipedia Networks: In Fig. 6, we present the change in the count of butterfly motifs (χ) per month upto May 2019 for the top 22 languages

of Wikipedia. Change in the count of Σ has either zero or positive values as the absolute count of this motif either remains constant or increases with time. In case of some Wikis, we retain the outliers as they are parts of the change patterns of cohesiveness.

In Fig. 6, we notice three different trends in the change of the count of butterfly motifs: linear increase, sudden increase after a certain duration and saturation. We observe from the plots of Fig. 6 that, in eswiki, dewiki, frwiki, svwiki, zhwiki, jawiki and enwiki, the change in count of (Σ) motifs is increasing linearly followed by big spikes in recent few years. For Example, in the case of dewiki and enwiki, count of Σ motifs is growing linearly till the first 190 months (≈ 15 years), increasing more in recent 2 years (big spikes). These patterns infer that cohesiveness in these Wikis is constantly improving with time and it rises even more in the last few years. Further, we also notice that the slope of linear growth is less in the case of svwiki and zhwiki as compared to eswiki, dewiki, frwiki, jawiki and enwiki Wikis. This infers that cohesiveness grows more slowly but constantly in such Wikis. We notice a huge surge in the change of Σ count after around 10 years in itwiki, arwiki, cawiki, fawiki, ruwiki, kowiki, srwiki and viwiki. As we can notice numerous big spikes in the change of cohesive units after a decade as compared to their initial growth of networks. This shows that cohesiveness in these Wikis enhanced suddenly after a specific time duration. Whereas we notice saturation in the change of Σ only in case of fiwiki, nlwiki and ptwiki Wikis. There are no specific patterns in srwiki and ukwiki Wikis to report. Thus, we observe the linear or sudden rise in cohesiveness in most of the considered languages of Wiki networks except few which shows saturation.

Saturation in the Cohesiveness of Editors in Wikipedia: In Fig. 7, we present the average change in the count of butterfly per editor ($\bar{\Sigma}_e$) for the top 500 editors and all editors for a few wikipedias. Top editors are those who are present in the maximum number of butterflies. In Table 1, we provide a sample size of Wiki networks, number

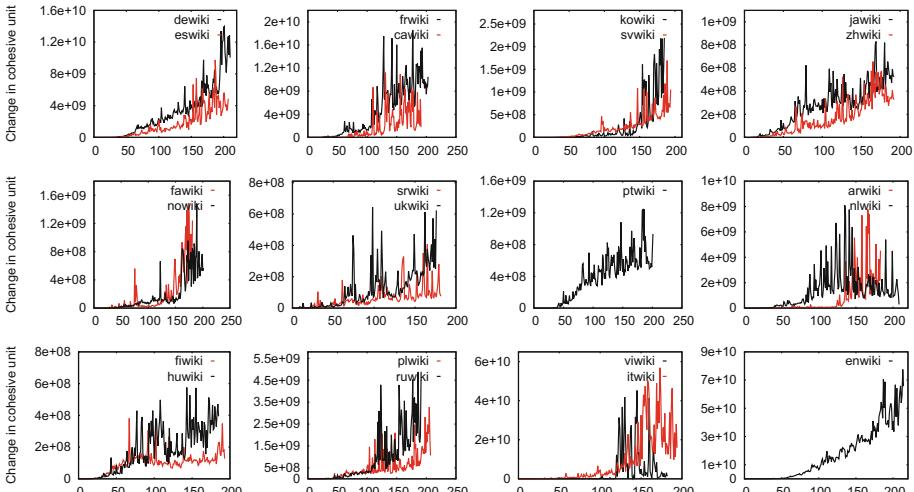


Fig. 6. Changes in the count of cohesive units for top 22 languages of Wiki. Here, x -axis has time in months and y -axis gives change in cohesive units.

of editors and edits in their samples to examine cohesiveness of editors. We calculate the average of \bar{X}_e over all editors for each month till May 2019. Interestingly, we notice that the growth patterns in cohesiveness of top 500 editors or all editors are similar to cohesiveness in Wiki networks as given in Fig. 6. Thus, we have not provided all the plots. Although the change patterns \bar{X} of networks and \bar{X}_e of editors are the same, interpretations are entirely different. For example, in Fig. 7(a), we present growth in the cohesiveness of top 500 German editors, we see that editors are present in an average 50 cohesive units in 100 months and which increases to 300 cohesive units in 200 months. This means that the number of cohesive units in which editors belong to in 200 months is 6 times more than the count in 100 months. This infers that these German editors are getting more involved in the cohesive environment with time. We present the cohesiveness of all the editors of the considered sample for enwiki in Fig. 7(b), dewiki and itwiki in Fig. 7(c) and ptwiki and frwiki in Fig. 7(d). We notice similar linear or sudden rise in the cohesiveness of editors in most of the considered languages of Wikipedia. In enwiki (English), the most popular Wiki, we notice that the contents and collaboration show decline or saturation while cohesiveness of both network and editors increase linearly with time.

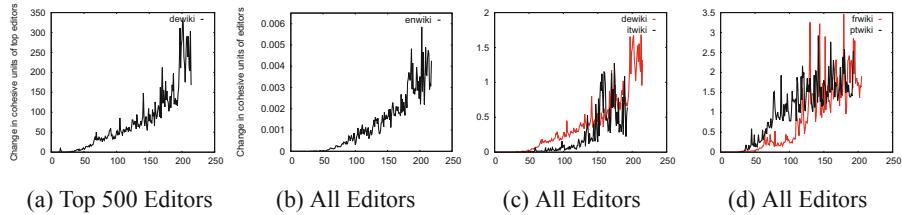


Fig. 7. Average changes in the count of cohesive units for (a): top 500 editors in dewiki; (b): all editors of enwiki; (c): dewiki, itwiki (d) ptwiki, frwiki. Here, x -axis shows time in months and y -axis gives average change in cohesive units of editors.

5 Conclusion and Future Work

We know that the contents of Wikipedia such as articles, editors and edits show slowdown or saturation after initial exponential growth. In this paper, we investigate whether collaboration and cohesiveness of Wikipedia also encounter slowdown or saturation. We apply the concept of network motifs to investigate this saturation phenomenon. We consider star motifs of articles with the average number of edits as collaborative units and complete 2×2 biclique motifs or “butterfly” as cohesive units. We present the change in the count of defined motifs upto May 2019 for the top 22 languages of Wikipedia. We observe slowdown or saturation in collaboration, whereas the linear or sudden rise in cohesiveness in most of the considered languages of Wikipedia. Although most of the studies reported that the contents of Wikipedia has reached natural limits of growth, we notice a rise in the count of butterfly motifs (*i.e.*, cohesiveness) in several wikipedias which indicates a steady growth in the smallest units of star motif (*i.e.*, wedge: collaboration of articles with 2 edits). This infers that the activities of editors in Wikipedia are

still progressing with time. In this research, we provide the empirical growth analysis of collaboration and cohesiveness in Wikipedia. In the future, we want to propose the analytical growth model to examine the collaboration patterns of Wikipedia.

References

1. Almeida, R.B., Mozafari, B., Cho, J.: On the evolution of wikipedia. In: ICWSM. Citeseer, Princeton (2007)
2. Bennett, L.M., Gadlin, H.: Collaboration and team science: from theory to practice. *J. Invest. Med.* **60**(5), 768–775 (2012)
3. Capocci, A., Servedio, V.D., Colaiori, F., Buriol, L.S., Donato, D., Leonardi, S., Caldarelli, G.: Preferential attachment in the growth of social networks: the internet encyclopedia wikipedia. *Phys. Rev. E* **74**(3), 036116 (2006)
4. Chandra, A., Maiti, A.: Modeling new and old editors' behaviors in different languages of wikipedia. In: International Conference on Web Information Systems Engineering, pp. 438–453. Springer (2018)
5. Gandica, Y., Carvalho, J., dos Aidos, F.S.: Wikipedia editing dynamics. *Phys. Rev. E* **91**(1), 012824 (2015)
6. Gibbons, A., Vetrano, D., Biancani, S.: Wikipedia: Nowhere to grow (2012)
7. Halfaker, A., Kittur, A., Riedl, J.: Don't bite the newbies: how revertors affect the quantity and quality of wikipedia work. In: Proceedings of the 7th International Symposium on Wikis and Open Collaboration, pp. 163–172. ACM (2011)
8. Jurgens, D., Lu, T.C.: Temporal motifs reveal the dynamics of editor interactions in wikipedia. In: Sixth International AAAI Conference on Weblogs and Social Media (2012)
9. Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, she says: conflict and coordination in wikipedia. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 453–462. ACM (2007)
10. Mickan, S., Rodger, S.: Characteristics of effective teams: a literature review. *Aust. Health Rev.* **23**(3), 201–208 (2000)
11. Rotabi, R., Kamath, K., Kleinberg, J., Sharma, A.: Detecting strong ties using network motifs. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 983–992. International World Wide Web Conferences Steering Committee (2017)
12. Sanei-Mehri, S.V., Sariyuce, A.E., Tirthapura, S.: Butterfly counting in bipartite networks. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2150–2159. ACM (2018)
13. Saracco, F., di Clemente, R., Gabrielli, A., Squartini, T.: Detecting early signs of the 2007–2008 crisis in the world trade. *Sci. Rep.* **6**, 30286 (2016)
14. Simmons, B.I., Sweering, M.J., Schillinger, M., Dicks, L.V., Sutherland, W.J., Di Clemente, R.: bmotif: a package for motif analyses of bipartite networks. *BioRxiv*, p. 302356 (2018)
15. Suh, B., Convertino, G., Chi, E.H., Pirolli, P.: The singularity is not near: slowing growth of wikipedia. In: Proceedings of the 5th International Symposium on Wikis and Open Collaboration, p. 8. ACM (2009)
16. Voss, J.: Measuring wikipedia (2005)
17. Wang, J., Fu, A.W.C., Cheng, J.: Rectangle counting in large bipartite graphs. In: 2014 IEEE International Congress on Big Data, pp. 17–24. IEEE (2014)
18. Wilkinson, D.M., Huberman, B.A.: Assessing the value of coooperation in wikipedia. arXiv preprint cs/0702140 (2007)
19. Yang, L., Wu, L., Liu, Y., Kang, C.: Quantifying tourist behavior patterns by travel motifs and geo-tagged photos from flickr. *ISPRS Int. J. Geo-Inf.* **6**(11), 345 (2017)
20. Zlatić, V., Štefančić, H.: Model of wikipedia growth based on information exchange via reciprocal arcs. *EPL (Europhysics Letters)* **93**(5), 58005 (2011)



ESA-T2N: A Novel Approach to Network-Text Analysis

Yassin Taskin, Tobias Hecking^(✉), and H. Ulrich Hoppe

University of Duisburg-Essen, Duisburg, Germany
{taskin,hecking,hoppe}@collide.info

Abstract. Network-Text Analysis (NTA) is a technique for extracting networks of concepts appearing in natural language texts that are linked by a certain measure of proximity. In prior works it has been argued that those networks are a representation of the mental model of the author. Extracting those networks often requires a high amount of domain knowledge of the analyst to specify relevant concepts in advance. Grammatical approaches that discover concepts automatically. However, the resulting networks can contain noisy concept nodes and meaningless edges, and thus, are less interpretable. In this paper, we present a new method that bridges between both approaches for extracting networks from text using Wikipedia as a knowledge base to map phrases occurring in the text to meaningful concepts. The utility of the method is demonstrated along a case study where pivotal moments in the evolution of Brexit debates in the British House of Commons in 2019 are discovered in speech transcripts.

Keywords: Network-text-analysis · Information extraction · Discourse analysis

1 Introduction

Network-Text Analysis (NTA) is a family of methods for extraction and analysis of networks from text corpora. In such networks terms (or concepts) are represented as nodes and an edge between the nodes represents a relationship between the corresponding concepts. These relationships can be based on textual proximity, co-occurrence, or grammatical relationships. Such networks represent the structure of the text and make the relations between the concepts, which may not initially be explicit, more visible. Networks of concepts interlinked based on certain notions of proximity originate from psychometrics, as for example pathfinder networks [14]. In their seminal work, Carley and Palmquist [1] argue that concept networks extracted from texts with concepts interlinked based on textual proximity can be considered as a representation of the text author's mental model. The explanation is that if two words are close in a person's mind, the person will also tend to put them close to each other when writing a text. This network-based representation allows for applying network analysis techniques

in order to identify roles of concepts or other structural characteristics of the textual network [2] which leads to a better understanding of the relationships between entities mentioned in a text. Application domains are, for example, scientometrics [15], media and communication [6], discourse analysis [9], or the analysis of narrative structures [10].

While there are many different methods to derive networks from text (c.f. [16]), the two main challenging steps are always identification of important terms (concepts) as nodes and the extraction of relations between concepts. Initially, NTA involved a high amount of manual coding of the network [1, 14], but meanwhile tool like Automap [5] for automated network extraction from text are available. In Automap, the user can specify different types of concepts that are located in the input texts and linked if they are not separated by more than n words. This automated processing enables the application of network text analysis on larger datasets. Automap and its successor ConText [4] allow for specifying controlled vocabularies to deal with synonyms, different spellings, and to assign further meaning to concepts. On the one hand, these vocabularies have to be specified manually which requires prior analysis of the texts or domain knowledge, but on the other hand all nodes in the resulting network represent meaningful concepts. The tool VOSviewer [15] differs from this approach in that it requires a similarity matrix that can typically be derived from co-occurrences of words. An example for a completely automated tool is InfraNodus [11] which follows a more grammatical approach to identify words in the text that should appear as nodes in the final network. However, on the downside it may happen that the final network contains many words that are not relevant in the application domain.

Irrespective of the application and the concrete methods applied for concept discovery and relation extraction, there is always a tradeoff between accuracy and the amount of manual work needed to extract meaningful network structures from text. The approach described in the following uses Wikipedia as a source of knowledge instead of using a controlled vocabulary. The idea is that if a concept is detected in the text for which there exist an article on Wikipedia, it is likely to be meaningful. Furthermore, the strength of the relation between two concepts can also be derived by examining the text similarity of the two corresponding articles. An implementation of our approach as a web service can be found on Github¹.

The utility of our approach will be demonstrated through a case study on analysing the relation between concepts that can be found in transcripts of evolving discussions between different speakers in Brexit debates of the British House of Commons. This introduces additional complexity since dynamic texts lead to dynamic concept networks. We generate networks after each speaker contribution and retain the information about added concepts and established relations, which allows for characterisation of the network over the course of the debate. The progress of the discussions are then analysed based on the dynamic networks extracted from discussion transcripts to identify different phases of the debate and to identify pivotal contributions changing the discourse.

¹ https://github.com/collide-uni-due/esa_ttn_web_service.

2 Approach

As described before, the primary aim is to increase the interpretability of a concept network by reducing the amount of nodes not representing meaningful concepts. Furthermore, edges between nodes should only be established if there is enough evidence that the corresponding concepts are related. In order to avoid too much manual effort for creating vocabularies or for filtering relationships, Wikipedia is utilised as a publicly available and processable knowledge source. The general idea is to map terms occurring in a text to Wikipedia articles (concepts) and to use the semantic relatedness of articles to get evidence about the strength of edges in the final concept network. One core element of the process of creating networks from text presented in this paper is based on Explicit Semantic Analysis (ESA) [7] which is a method of representing the meaning of text as vector of concepts extracted from a large thematic text corpus such as Wikipedia. Therefore we will first briefly explain ESA and introduce how the text representation utilising ESA is integrated into a pipeline to create network representation of an input text.

2.1 Explicit Semantic Analysis (ESA)

The ESA approach aims to convey the semantic meaning of a term by representing it as a high dimensional vector of concepts. These concepts are usually the titles of an article corpus, usually Wikipedia. The association between a term and a Wikipedia article is determined by the *tf-idf* score [13] for the term given the article corpus. This score consists of two components. The first is the term frequency which counts the occurrences of a term in an article. The more often it appears the more relevant a term is. The second part is the inverse document frequency. It is determined by calculating the fraction of documents of a corpus that a term appears in. It is used as a measure of specificity of a term. Terms with high *idf* are likely domain specific vocabulary and therefore more relevant to the meaning of an article.

As a basis a Wikipedia corpus is processed the *tf-idf* scores for all term-article pairs are calculated. Consequently one can assign a given term t and ESA vector $v_{article}(t) \in \mathbb{R}^n$, where n is the number of Wikipedia articles and each article corresponds to one dimension. The elements of $v_{esa}(t)$ denote the scores the term has for each of the n articles. This vector is typically sparse, i.e. most scores are 0. In the same way one can also create vectors $v_{term}(a) \in \mathbb{R}^m$ for an article a , where the elements give the association of a given article a to each of the m terms. We make use of this possibility in our method of creating text networks.

2.2 Architecture

The architecture for the processing pipeline is depicted in Fig. 1 and each step is explained in the following:

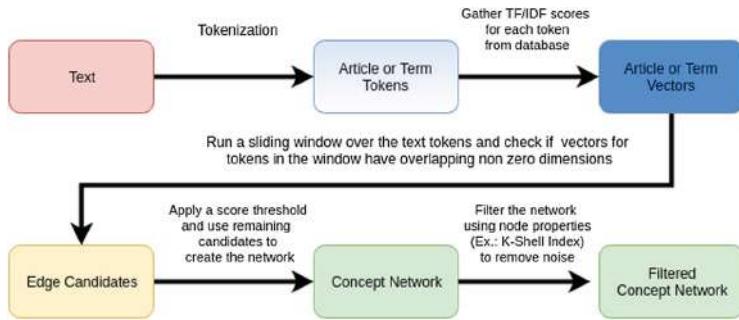


Fig. 1. Pipeline for text network generation

... the functional products of **natural selection** or **sexual selection** in human evolution ...

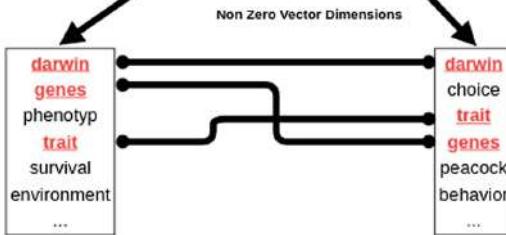


Fig. 2. Example of identifying edge candidates in a sliding window, with tokens being article names.

Tokenisation and Concept Identification. In the first step the input text is tokenised into single terms. Next, these terms are then matched against a collection of titles of Wikipedia articles. First single terms are matched if possible. Phrases consisting of two, three, ..., terms that occur successively in the text are also mapped to Wikipedia articles in order to account for multi-word expressions.

Concept Representation. The next step is the generation of ESA vectors $v_{term}(a_i)$ for the matched articles a_i . Note that it would also be possible to use the article associations of the terms occurring in the text but this is not applied in this work. Term/article pairs can easily be looked up from the *tf-idf* database generated from the Wikipedia corpus in the previous step.

Identifying Edge Candidates. To generate candidates for edges in our text network a sliding window approach is used as in other network extraction tools (c.f. [4,5,11]). A sliding window comprising of k words is moved over the text word by word. Whenever two discovered concepts occur in one window (are

not separated by more than $k - 2$ words) an edge candidate is established. In contrast to older approaches, this technique is enriched with the data available from the ESA database. An illustrated example can be seen in Fig. 2 where two terms in a window are matched to different Wikipedia articles that have some overlap in their associated ESA term vectors. The ESA vectors of each pair of tokens in a window are examined and the matching between their non zero dimensions is calculated. For matching tokens the corresponding scores are multiplied resulting in a combined score for the edge candidate in the following format: (*concept A, concept B, connecting terms, score*). Afterwards the scores are accumulate for all candidates for the entire text.

Concept Network Generation. To generate a concept network, first the list of edge candidates is filtered according to a specified threshold of their scores. The rationale behind this is that edge candidates having low scores are likely to be from common use words which have a low *idf*, and thus, likely an overall low score. An properly set threshold filters out noise and keeps mostly connections between domain specific concepts. Higher thresholds put more focus on commonly known relations in a domain that are reflected in the Wikipedia knowledge base, and thus, reduces noise, while lower thresholds allow for the discovery of potentially unknown relationships. After this, either an unipartite network can be generated by adding the connecting terms and scores as edge properties, or alternatively, a bipartite network consisting of concepts linked to connecting terms.

Filtered Concept Network. Although the prior filtering of edge candidates that do not represent domain specific relationships connections might be left that have some meaning according to the ESA vectors of the two endpoints but might still not be representative for the topic discussed in the text. Therefore, the network is further reduced by removing nodes with a very low k-core index [3] (typically $k = 2$). This would eliminate nodes that do not have a sufficient amount of connections to other nodes that are also well connected forming a k-core.

3 Case Study

In order to demonstrate the utility of the WikiT2N approach, we analyse the transcripts of the Brexit debates in the British House of Commons. The goal is to distinguish different episodes of the discussion by analysing dynamic concept networks created from the statements made by the members of the parliament. This kind of analysis is inspired by the work of Min and Park [10] who used network-text-analysis to characterise the structure of narratives, which can be also adapted to texts representing discourses. The full transcripts of all utterances of members of the House of Commons during the debates are fully accessible at the House of Commons Hansard Website². Even though the debates are

² <https://hansard.parliament.uk/commons>.

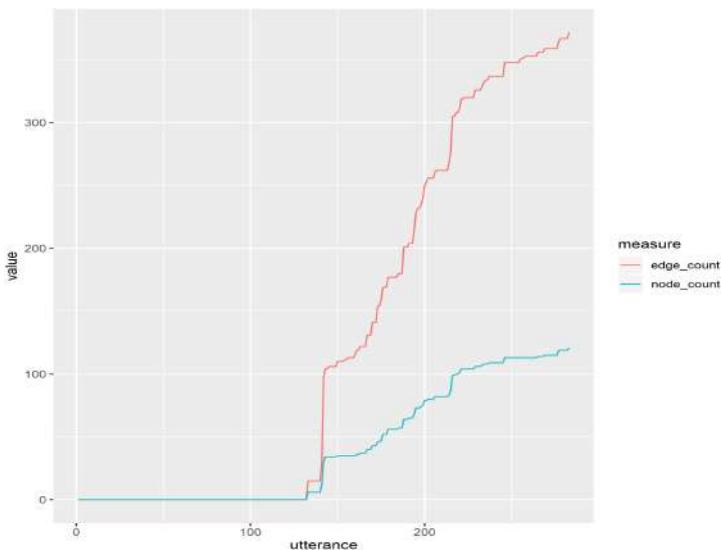


Fig. 3. Aggregated number of nodes and number of edges over the utterances of the debate on 15 January 2019

public and fully documented online, in the following only initials of parliamentarians are used instead of full names in order to depersonalise the interpretation and to emphasise scientific neutrality. The data has been processed into a list of contributions by different speaker which forms the input to the algorithm described above.

3.1 Network Construction

For network generation the Brexit discussion transcripts were decomposed into series of speaker contributions. From this a corpus of texts growing by one contribution each time is generated. From these texts a dynamic network is generated that successively grows with each speaker contribution. The generated network is unipartite. Text tokens were treated as article titles giving us networks of connected Wikipedia concepts. The sliding window over the text had a size of 20 and we filtered the networks, removing all nodes with a k-core index of less than 2. The window size of 20 was chosen because on average it encompasses the words of two successive sentences, which allows us to examine connections of concepts across sentence borders. This is reasonable since statements made by the speakers are likely to be split into multiple sentences.

3.2 Discourse Episodes

Because of space limitation only the results of the analyses of two sittings of the House of Commons are presented that are representative of different discourse patterns that could be identified for all transcripts.

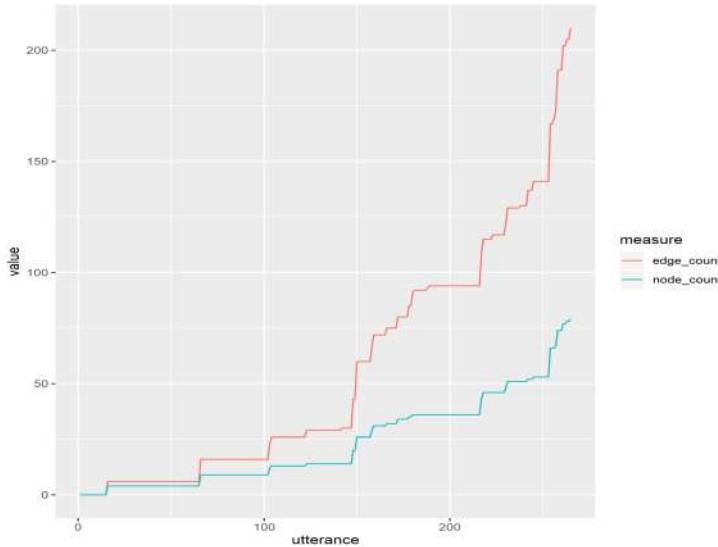


Fig. 4. Aggregated number of nodes and number of edges over the utterances of the debate on 27 February 2019

Debate on 15 January 2019. The first debate in 2019 on the United Kingdom leaving the European Union took place on 15 January. The number of nodes and number of edges over the course of the discussion is depicted in Fig. 3. Surprisingly due to the filtering of nodes that have a k-core index less than 2 the resulting network is empty for the first 132 utterances until I. B. starts to participate in the discussion and exchanges arguments with several others. From contribution 142 on the number of edges raises drastically while the increase of the number of nodes is not as steep. This can be seen as the pivotal moment in the discussion on 15 January since from this contribution on the concept network is very dynamic. The contribution 142 was made by B. C. who is one of the strongest critics of the EU. In the statement the parliamentarian mostly establishes links between different European countries and regions by giving a longer speech expressing a critical view on past and current developments in various countries. Further new edges are, for example, “central europe” – “youth unemployment” or “parliament” – “constitutional crisis”. The fact that the number of nodes does only slightly increase afterwards compared to the number of new connections introduced in every following utterance indicates that the parliamentarians do not introduce many new topics but they interrelate different terms that were introduced by previous speakers.

Debate on 27 February 2019. The debate of the 27 February is an example for a different pattern from the one shown above. In Fig. 4 it can be seen that the evolution of the concept network over time can be partitioned into multiple episodes.

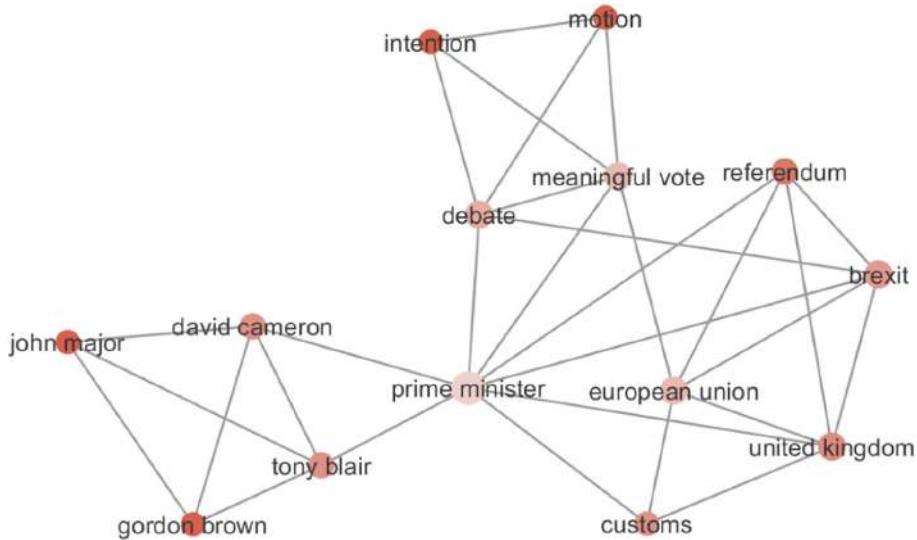


Fig. 5. Concept network build from utterance 1 to 145 in the Brexit debate on 27 Feb. 2019

At the beginning there is a dialogue between D. L. and other members of the parliament. In this episode the number of nodes and edges stay relatively stable. There are some moments when the discussion slightly shifts and concept relations are introduced before the resulting network remains again almost stable for a certain period.

The pivotal moment appeared after utterance 147, which is the announcements of the results of the various votes on different aspects of exiting the EU (deferred divisions). After that follows a short speech by C. S. advocating voting for a deal with the European Commissions that follows into a converse discussion involving several members of the parliament.

When visualising the concept network derived from utterance 1 to 145 (Fig. 5) and utterance 146 to 200 (Fig. 6) one can see very clearly the shift of the debate after the voting results were announced. While in the beginning the most central concept is “prime minister” and the debate is much focused on people, afterwards the terms “brexit” and “democracy” are very central, as expected, while “prime minister” is only one of the peripheral nodes. Furthermore, the concept “time” becomes important since the timing of leaving the EU is mentioned by several people.

The colour of the nodes in Figs. 5 and 6 corresponds to the local clustering coefficient of the nodes. Darker colours indicate a higher tendency of the nodes neighbours to be neighbours themselves. It can be seen that the resulting networks tend to form small clusters representing different facets of the discussion. These are connected by nodes representing the overarching themes of the discussion, and thus, can be considered as boundary concepts with a lower clustering coefficient.

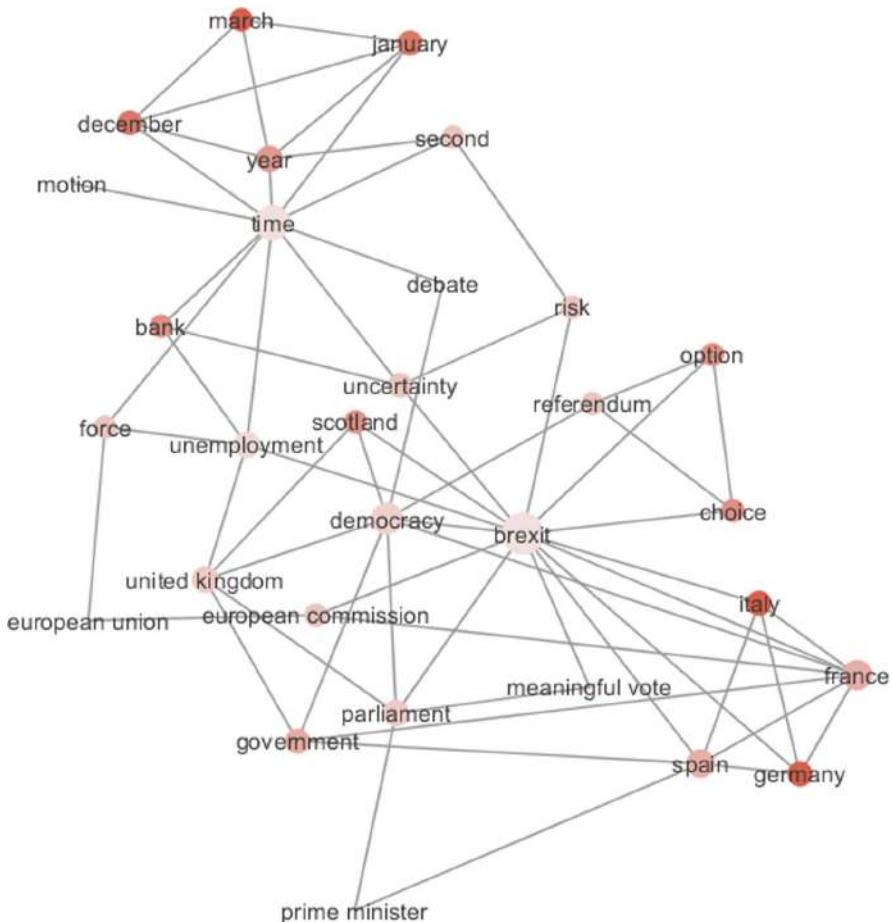


Fig. 6. Concept network build from utterance 146 to 200 in the Brexit debate on 27 Feb. 2019

4 Conclusion and Future Work

In this paper we presented a novel method to transform natural language texts into a network representation of connected concepts. This process relies on Wikipedia as a knowledge source to increase the interpretability of the result by filtering irrelevant concepts and relations. This approach is in contrast to many existing text-to-network techniques where terms are simply connected to each other if they appear close in the text or other notions of co-occurrence. To get meaningful results this is often supplemented by manually generating a concept and deletion list to filter for relevant words. Furthermore a generalisation thesaurus can be provided which maps different text strings to the same concept. This, however, requires much manual work and often profound domain

knowledge. Through the use of Wikipedia as a knowledge source we can attain similar effects automatically without manual intervention from the user. Only having Wikipedia articles as candidates filters out non relevant text phrases which are then only connected if they have semantic similarity which further reduces noise. Generalisation similar to a thesaurus can be achieved by making connecting concepts explicit. Furthermore, this automatic approach does not suffer from the possibility of excluding relevant concepts due to human error and is thus suitable for exploratory analyses.

The utility of the implemented technique for exploratory data analysis was demonstrated by applying it to a real world dataset of parliamentary discussions about the Brexit in 2019. It could be shown that the extracted networks are meaningful and a means to characterise different phases in the debates, particularly identifying pivotal contributions that change the structure of the discourse afterwards.

In further research the suitability of this approach for different tasks could be evaluated. A network approach has the advantage of incorporating concept relations instead of simple counting of word scores for query terms. For example clustering of concept networks extracted from text can be helpful to supplement topic models for increasing the interpretability of topic [8]. Furthermore, calculation of document similarity and document clustering could be based on network similarity [12].

References

1. Carley, K., Palmquist, M.: Extracting, representing, and analyzing mental models. *Soc. Forces* **70**(3), 601–636 (1992)
2. Carley, K.M.: Network text analysis: the network position of concepts. *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts* **4**, 79–100 (1997)
3. Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y., Shir, E.: A model of internet topology using k-shell decomposition. *Proc. Nat. Acad. Sci.* **104**(27), 11150–11154 (2007)
4. Diesner, J.: Context: software for the integrated analysis of text data and network data. paper presented at the social and semantic networks in communication research preconference at international communication association (ica) (2014)
5. Diesner, J., Carley, K.M.: AutoMap 1.2 : extract, analyze, represent, and compare mental models from texts (2004). https://kilthub.cmu.edu/articles/AutoMap_1.2-extract_analyze_represent_and_compare_mental_models_from_texts/6621194
6. Diesner, J., Rezapour, R., Jiang, M.: Assessing public awareness of social justice documentary films based on news coverage versus social media. IConference 2016 Proceedings (2016)
7. Gabrilovich, E., Markovitch, S., et al.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJcAI* **7**, 1606–1611 (2007)
8. Hecking, T., Leydesdorff, L.: Can topic models be used in research evaluations? reproducibility, validity, and reliability when compared with semantic maps. *Res. Eval.* **28**(3), 263–272 (2019)

9. Introne, J.E., Drescher, M.: Analyzing the flow of knowledge in computer mediated teams. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work, pp. 341–356. CSCW 2013, ACM, New York, NY, USA (2013). <http://doi.acm.org/10.1145/2441776.2441816>
10. Min, S., Park, J.: Mapping out narrative structures and dynamics using networks and textual information (2016). arXiv preprint [arXiv:1604.03029](https://arxiv.org/abs/1604.03029)
11. Paranyushkin, D.: Infranodus: generating insight using text network analysis. In: The World Wide Web Conference, pp. 3584–3589. WWW 2019, ACM, New York, NY, USA (2019). <http://doi.acm.org/10.1145/3308558.3314123>
12. Paul, C., Rettlinger, A., Mogadala, A., Knoblock, C.A., Szekely, P.: Efficient graph-based document similarity. In: European Semantic Web Conference, pp. 334–349. Springer (2016)
13. Robertson, S.E., Jones, K.S.: Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* **27**(3), 129–146 (1976)
14. Schvaneveldt, R.W.: Pathfinder Associative Networks: Studies in Knowledge Organization. Ablex Publishing, New Jersey (1990)
15. Van Eck, N., Waltman, L.: Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics* **84**(2), 523–538 (2009)
16. Vega, D., Magnani, M.: Foundations of temporal text networks. *Appl. Netw. Sci.* **3**(1), 25 (2018)



Understanding Dynamics of Truck Co-Driving Networks

Gerrit Jan de Bruin^{1,2,3(✉)}, Cor J. Veenman^{1,4}, H. Jaap van den Herik²,
and Frank W. Takes¹

¹ Leiden Institute of Advanced Computer Science, Leiden University,
Leiden, The Netherlands

g.j.debruin@liacs.leidenuniv.nl

² Leiden Centre of Data Science, Leiden University, Leiden, The Netherlands

³ Human Environment and Transport Inspectorate,
Netherlands Ministry of Infrastructure and Water Management,
The Hague, The Netherlands

⁴ Data Science Department, TNO, The Hague, The Netherlands

Abstract. The goal of this paper is to learn the dynamics of truck co-driving behaviour. Understanding this behaviour is important because co-driving has a potential positive impact on the environment. In the so-called co-driving network, trucks are nodes while links indicate that two trucks frequently drive together. To understand the network’s dynamics, we use a *link prediction* approach employing a machine learning classifier. The features of the classifier can be categorized into spatio-temporal features, neighbourhood features, path features, and node features. The very different types of features allow us to understand the social processes underlying the co-driving behaviour. Our work is based on a spatio-temporal data not studied before. Data is collected from 18 million truck movements in the Netherlands. We find that co-driving behaviour is best described by using neighbourhood features, and to lesser extent by path and spatio-temporal features. Node features are deemed unimportant. Findings suggest that the dynamics of a truck co-driving network has clear social network effects.

Keywords: Transport networks · Mobility · Co-driving behaviour · Spatio-temporal networks · Link prediction

1 Introduction

Nowadays an increasing volume of published studies concerning social network analysis is combined with spatio-temporal data. Much of the research performed so far used either GPS [5, 11], WiFi [13] or calls from mobile phones [16]. In this study, we analyze 18 million truck movements in the Netherlands.

The goal is to study social phenomena amongst truck drivers, so that we may understand why truck drivers initiate so-called *co-driving behaviour* with other drivers. In simple terms, co-driving is the activity where two trucks drive

together, i.e., are frequently at the same place at the same time. We assume a direct and natural relation between a truck and its driver, meaning that a truck driver only drives one truck. A second assumption is that every truck is always driven by the same driver. To make sure that only *intentional* co-driving activity is investigated, a number of strict selection criteria are used. These criteria are explained in Sect. 4.2.

Co-driving behaviour is known to have a potential positive impact on the environment through optimizing logistics and consequently reducing fuel use [15]. Hence, an improved understanding of co-driving behaviour may stimulate additional co-driving behaviour. Moreover, innovative forms of transportation, such as autonomous driving, may have major implications for this behaviour.

Without traditional spatio-temporal data mining techniques, we construct a so-called *co-driving network* from the data. The nodes of this network are trucks, while a link is drawn between two nodes when these two trucks frequently show co-driving behaviour. The obtained co-driving network shows properties similar to other networks often analyzed in the field of complex network analysis [1]. We mention three of them. First, we see that the network has a giant component containing the majority of nodes and edges. Second, the relatively low average shortest path length suggests a small-world network structure [17]. Third, the degree distribution appears to follow a power law, indicating that the network may be scale-free [1].

Previous work on similar data focused on communities and static properties of the co-driving network [3]. In contrast, the goal of this work is to learn the *dynamics* of the co-driving network. To this end, we use a link prediction approach [8]. More concretely, we develop a machine learning classifier which predicts for all possible pairs of trucks that are not connected, whether a link is formed in the future. We then investigate the importance of each type of feature that occurs in the link prediction classifier. These allow us to understand what is assessed important by the classifier and hence contributes to co-driving behaviour. The features used can be categorised into four different types of features.

1. Spatio-temporal features, aiming to summarize registrations at different locations over a given period.
2. Neighbourhood features, related to the local embedding of considered trucks in the co-driving network.
3. Path features, describing distance-related properties of truck pairs based on the global structure of the network.
4. Node features, related to static meta-information of trucks.

The overall structure of this paper takes the form of the different steps taken to attain our goal and is as follows. In Sect. 2, relevant work is provided on analyzing dynamics in social networks including spatio-temporal data. Section 3 describes the spatio-temporal truck data. Section 4 reports how a co-driving network is constructed from the data. In this section we also discuss the characteristics of the obtained network. Section 5 provides a formal description of

the link prediction approach. It also explains how the different features are constructed from both the data and the obtained network. Section 6 outlines the experimental setup, demonstrates the performance of the link prediction approach and assesses the feature importance. Finally, in Sect. 7 we conclude our paper and provide suggestions for future work.

2 Related Work

From the substantial body of related work on spatio-temporal data we select three different approaches which have been frequently used to study dynamics in networks at the level of individual nodes. These three different approaches have in common that they all try to understand the underlying social network by studying node attributes available in the data.

First, Sekara *et al.* use Bluetooth sensors to measure proximity of students [14]. The authors show that when high-resolution data is available (both in time and location), groups of interacting nodes can be observed directly. Hence, making sense of individual node attributes using network measures can be performed directly. As an example, the authors show that the students explore new locations in groups during the weekend, while the groups tend to be at the same location.

Secondly, Kossinets and Watts analyze e-mail data gathered from students and employees at a university [7]. Unlike our truck data, e-mail data does not contain spatial information. However, different attributes are collected and analyzed in this work such as professional status, gender and age.

Finally, Wang *et al.* analyze the mobility patterns by tracking both the mobility and interactions of millions of mobile phone users [16]. A social network is constructed from phone calls, where users are connected when they communicate. They provide three findings. First, the authors find that spatial trajectories of two users strongly correlate when they are close in the social network Second, finding is that mobility features have high predictive power on which nodes will connect, comparable to that of network proximity features. Third, the link prediction performance by using both network proximity and mobility features. We will use a similar link prediction approach in our work. In addition we build on the other works [7, 14] by distinguishing between weekends and weekdays, and using a combination of both network and static attributes.

3 Data

Data collection took place at 18 different locations throughout the Netherlands between 2016 and 2018. Using an automatic number-plate recognition (ANPR) system, every truck passing these locations is registered. The data is obtained by the same systems used in earlier work [3]. At some locations the registration systems faced an unexpected downtime. To ensure a sufficiently valid range of data, only registrations from 6 out of 18 systems have been considered. These systems were placed near the port of Rotterdam. Furthermore, registrations

with low quality data are removed, such as invalid characters in license plates and non-existing countries.

The aforementioned quality selections reduce the number of registrations from 18,678,420 to 9,202,764. The number of registrations over time is shown in Fig. 1. We observe that the number of registrations after applying the quality selections is more stable over time. In Fig. 2 the distribution of the number of registrations per truck is shown (note that both axes have logarithmic scales). The distribution of the number of registrations per truck remains similar.

4 The Co-Driving Network

We start with a description on how the co-driving network is constructed in Sect. 4.2. Section 4.3 continues then with general statistics of the obtained network.

4.1 Definition of an Intentional Co-Driving Event

We will now provide a more formal definition of a co-driving event. Our dataset of all registrations (as discussed in Sect. 3) is denoted by \mathcal{D} . We use \mathcal{D}_u to refer to all registrations x_i in dataset \mathcal{D} from truck u with license plate $lp_i = u$. More formally, $\mathcal{D}_u = \{x_i \in \mathcal{D} : lp_i = u\}$. We speak of a co-driving event (u, v, t) when two registrations $x_i \in \mathcal{D}_u$ and $x_j \in \mathcal{D}_v$ from trucks u and v exists at the same location $loc_i = loc_j$ at time t_i with at most $\Delta t = |t_i - t_j|$ seconds between them. There are two ways in which these co-driving events can occur: (1) randomly because two trucks just happen to be at the same place around the same time or (2) intentionally because two trucks were involved in *intentional co-driving*. Our goal is to study intentional co-driving behaviour while keeping the random co-driving events to a minimum.

4.2 Network Construction

The following two steps are taken to ensure that only intentional co-driving is studied. First, we separate the intentional co-driving events from the random

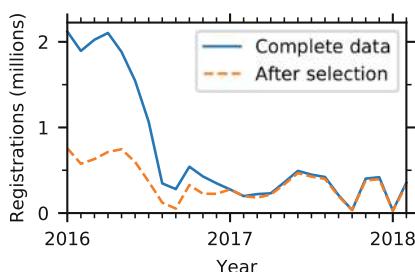


Fig. 1. Number of registrations over time.

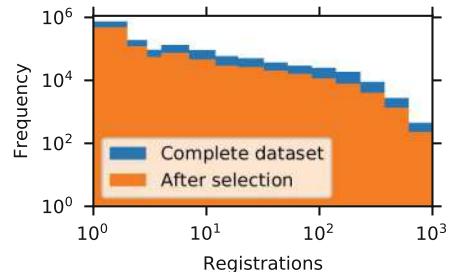


Fig. 2. Histogram of registrations per truck. Note logarithmic axes.

ones by selecting only co-driving events in which trucks u, v at least twice drive at most $\Delta t \leq \Delta t_{\max}$ seconds apart. We will discuss the Δt_{\max} parameter shortly. Second, at least two of those co-driving events should have a gap of more than two hours, i.e., there exists two co-driving events (u, v, t_i) and (u, v, t_j) for which $|t_i - t_j| \geq 2$ h. With the latter requirement, we ensure that the two co-driving events originate from different truck journeys. By applying these criteria, we minimize the probability that a random co-driving event is marked as an intentional co-drive.

The temporal network $G = (V, E)$ is constructed. In this network, the nodes are the trucks $u, v \in V$ that frequently show (intentional) co-driving behaviour. The links of this network consist of the obtained co-driving events $(u, v, t) \in E$ between those trucks. Note that multiple links (u, v, t) exist between two nodes u and v with different t as a result of the first step taken to select only intentional co-driving. We refer to the number of links between u and v as $w_{u,v}$, with $w_{u,v} \geq 2$ as a result of the selection criteria discussed above. When no links exist between u and v , the weight $w_{u,v}$ equals 0.

Then, we need to find the appropriate value for the previously discussed parameter Δt_{\max} . There is a trade-off when setting the value of Δt_{\max} . High values will result in selecting a large share of random co-driving events, while low values will result in the omission of intentional co-driving behaviour. We present three thoughts when determining the value of Δt_{\max} .

First, Fig. 3 shows the distribution of the time gap between two co-drivings events. We observe that distinct behaviour is present for random and intentional co-driving events. Two trucks involved in intentional co-driving drive closer together than randomly co-driving trucks. Note that the time gap in intentional co-driving trucks peaks at around 2 s, close to the 1.3 s which is considered a safe driving gap between two trucks [9]. After $\Delta t \approx 8$ s the relative frequency of intentional co-driving trucks becomes similar to that of randomly co-driving trucks. This may indicate that from this value onward only random co-driving events are selected as intentional co-driving.

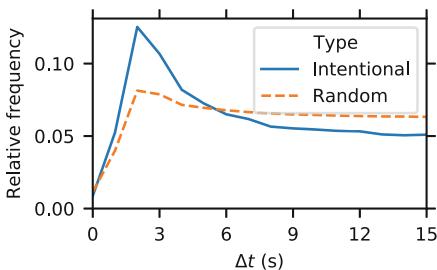


Fig. 3. Frequency distribution of time gap measured between the two trucks in a co-driving event.

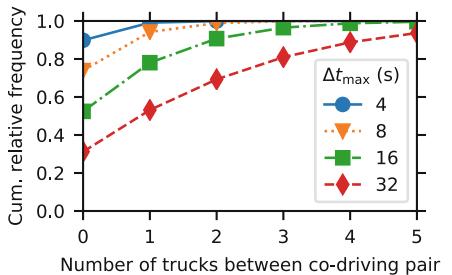


Fig. 4. Frequency distribution of trucks driving between the two trucks in a co-driving event.

Second, Fig. 4 shows the distribution of the number of trucks driving between two trucks involved in intentional co-driving for various values of Δt_{\max} . For values between 4 and 8 s, we observe that virtually all trucks are driving with at most one truck in between them. Higher values result in a non-negligible probability that more than two trucks are driving between the two co-driving trucks. We consider it unlikely that trucks are intentionally co-driving when more than two trucks drive between these trucks. This is the case for values of $\Delta t_{\max} \geq 16$ s.

Third, we rationalize that following a truck intentionally is only possible when there is at most a few hundred of meters between the two trucks. Provided that trucks in our data drive typically at a speed around 20 m/s, this means that reasonable values for Δt_{\max} should be at most 20–30 s.

In summary, we are conjectured that the robustness checks above enable us to properly select intentional co-driving behaviour for further analysis in the remained of this paper.

4.3 Network Statistics

We continue by providing general statistics of the obtained network in Table 1. For definitions of these statistics, see [2]. Note that multiple links are present between nodes. This is caused by the first measure taken to select intentional co-driving events in Sect. 4.2. In Fig. 5a the distribution of the number of neighbours of each node is shown. We show in Fig. 5b the distribution of node strengths. For a node, this value is equal to the sum of the weights of the nodes connected to it. Note that both distributions appear to have power-law behaviour. Together with the presence of a giant component and a relatively low average shortest path length the power-law behaviour suggests that the network is remarkably similar to other social networks and scale-free networks commonly observed in real-world settings [2, 17].

Table 1. General network statistics

Property	Value
Number of nodes	25,553
Number of links	73,059
Number of connected node pairs	27,986
Fraction nodes in giant component	62%
Fraction links in giant component	79%
Density	2.2×10^{-4}
Power law exponent γ	3.3
Average shortest path length	7.8
Diameter	24

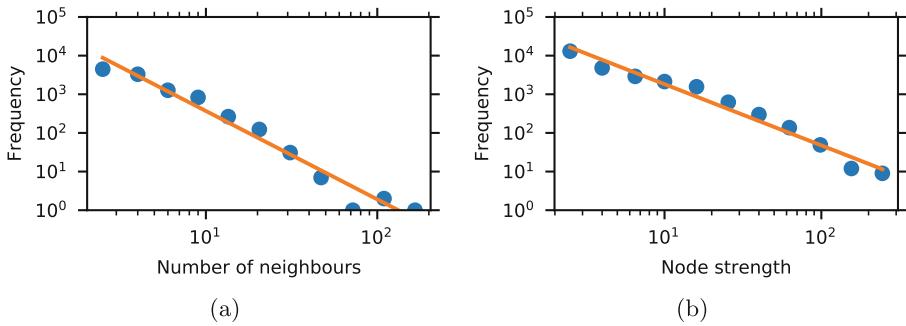


Fig. 5. The number of neighbours (a) and node strength (b) distribution of the network. Note the logarithmic axes.

5 Approach

This section presents our approach to the analysis of the dynamics of the co-driving network. We start with a description of the proposed link prediction approach in Sect. 5.1. The feature construction steps are outlined in Sect. 5.2. Finally, we provide the measures taken to reduce the observed class imbalance in Sect. 5.3.

5.1 Link Prediction

We start with a formal description of the link prediction problem. Given a snapshot of the network at time τ , the link prediction classifier needs to predict newly formed links in the evolved network after time τ . In doing so, the classifier is able to use present information to predict future links. The input of this classifier is a feature matrix X , which is based on a snapshot of the network at time τ : $G_\tau = (V_\tau, E_\tau)$ with $E_\tau = \{(u, v, t) \in E : t \leq \tau\}$ and $V_\tau = V$. The feature vector is calculated for each candidate node pair which is not linked (yet) in G_τ : $\{(u, v) \in V_\tau \times V_\tau : u \neq v, (u, v, t) \notin E_\tau\}$. The target of the classifier, y , denotes for a node pair whether a link is present in the evolved network:

$$y_{u,v} = \begin{cases} 0 & \text{if } (u, v, t) \notin E \\ 1 & \text{if } (u, v, t) \in E \end{cases} \quad \text{for some } t > \tau$$

Note that only link formation is predicted; we do not predict the weight of the link. Accordingly, this classifier can be seen as a supervised binary classifier.

A random forest classifier is used. We choose this classifier because random forests are known to generalize well on unseen data and allows one to determine the importance of each feature [4, 6]. The random forest classifier contains 128 decision trees, since larger values usually bring no significant performance gain [10]. Each individual decision tree is trained on a randomly drawn selection of variables. The number of randomly drawn features is equal to the square root

of the total number of variables, which is a common value used in classification [12]. By sampling randomly with replacement from the data, randomness is increased for the individual decision trees. The splitting criteria of the nodes are determined by considering the reduction in Gini impurity [6]. The random forest classifier allows to obtain the importance of each feature by determining the reduction in the Gini impurity for splitting nodes with a certain variable [6]. Recall that this is important, as it enables us to understand the dynamics of truck co-driving behaviour.

Because only a sample of the data is used for each tree, we can calculate an estimate of the performance for the decision forest on the remaining part of the data [6, 12]. We use this estimate to assess the optimal value for the depth of the decision trees in the random forest. The performance of the classifier is calculated on the test set, which is a 10% random sample of the data. This data is not used in the training of the classifier nor in finding the optimal value for the depth of the decision trees.

5.2 Feature Construction

We continue with the explanation of the feature vector which is used for each candidate truck pair (a, b) . Table 3 provides all 52 features used by the link prediction classifier. These features can be categorised into four different types, each described in more detail below.

1. Node features, constructed from information available about the trucks.
2. Spatio-temporal features.
3. Neighbourhood features, which consider relevant operations related to micro-level properties of the nodes of the candidate pair. The neighbourhood of a node is defined by $\Gamma(a) = \{v \in V : (a, v, t) \in E \text{ for some } t\}$. The strength of a node is the total number of links connected to a node, $s_a = \sum_{\{u \in V\}} w_{a,u}$.
4. Path features, that consider macro-level properties of the network. We consider only the shortest path length in this work.

Truck Properties. The ANPR-system determines the license plate and country ($country_u$) of each truck u passing by. We use \mathcal{D}_u to denote all registrations x_i available of truck u , as explained in Sect. 4.2. The registration systems are also equipped with sensors to measure the length ($length_i$), mass ($mass_i$) and number of vehicle axes ($axes_i$) of each truck. These measurements may slightly differ between registrations. Therefore, we calculate the averages shown in Table 2 for each truck in the network. The *driving_hours* and *weekend_driver* features are calculated because they are known to vary between trucks operating in different industrial sectors.

Spatio-Temporal Information. The goal of this type of feature is to capture both spatial and temporal behaviour for the truck pair under consideration.

Table 2. Information available about truck u , collected from its registrations \mathcal{D}_u .

Property		Description	Type
$truck_country_u$		Country of registration	String
$truck_axes_u$	Median $axes_i$ $\{x_i \in \mathcal{D}_u\}$	Number of axes	Number
$truck_length_u$	Median $length_i$ $\{x_i \in \mathcal{D}_u\}$	Length	Number
$truck_mass_u$	Median $mass_i$ $\{x_i \in \mathcal{D}_u\}$	Mass	Number
$driving_hours_u$	Mean $ t_i(h) - 12h $ $\{x_i \in \mathcal{D}_u\}$	Usual driving hours	Number (0–12)
$weekend_driver_u$	Mean $\begin{cases} 0 & \text{if } t_i = \text{weekday} \\ 1 & \text{if } t_i = \text{weekend} \end{cases}$ $\{x_i \in \mathcal{D}_u\}$	Fraction driving in weekend	Number (0–1)

We do so by counting the number of registrations in different time periods. As an example, for feature $last_day_l(a+b)$ registrations are counted for trucks a and b at location l in the last day before the considered time. We consider time periods of one week, one month and one year. These periods are chosen in such a way that they cover a broad window of possible relevant time periods.

5.3 Class Imbalance

It is well-known that in real-world networks the link prediction classifier comes with a large class imbalance [16]. The performance of the employed random forest classifier may drop if there is a large class imbalance. To overcome this limitation, we use the following three measures. First, both classes are given a weight, such that in total both classes have equal weight. Second, we consider only truck pairs where both trucks are involved in co-driving events in the last two months before time τ . This will both reduce the number of considered truck pairs and reduce class imbalance. Third, we consider only node pairs with both trucks in the giant component of the network. An additional advantage of this measure is that the shortest path feature is well-defined.

6 Experiments and Results

To determine important factors that govern the dynamics in the truck co-driving network, the approach as set out in Sect. 5 is applied to the network discussed in Sect. 4.

The value of τ is chosen such that 50% of the links are formed. In this way, the number of considered truck pairs and class imbalance are reduced, while ensuring that at least 1,000 truck pairs are present that will make a link. For this value of τ we find a class imbalance of 1:61,000. By taking the measures mentioned in Sect. 5.3, the class imbalance is reduced to 1:15,000. The random

Table 3. Features for truck pair (a, b) used in the link prediction model. The rightmost column lists the feature importance calculated using Gini importance as provided by the random forest classifier.

Index	Feature	Type	Importance
X_1	$\text{truck_country}(a) = \text{truck_country}(b)$	Node	0.005
X_2	$\text{truck_ax}(a) + \text{truck_ax}(b)$	Node	0.006
X_3	$ \text{truck_ax}(a) - \text{truck_ax}(b) $	Node	0.008
X_4	$\text{truck_len}(a) + \text{truck_len}(b)$	node	0.017
X_5	$ \text{truck_len}(a) - \text{truck_len}(b) $	Node	0.040
X_6	$\text{truck_mass}(a) + \text{truck_mass}(b)$	Node	0.016
X_7	$ \text{truck_mass}(a) - \text{truck_mass}(b) $	Node	0.030
X_8	$\text{driving_hours}(a) + \text{driving_hours}(b)$	Node	0.016
X_9	$ \text{driving_hours}(a) - \text{driving_hours}(b) $	Node	0.030
X_{10}	$\text{weekend_driver}(a) + \text{weekend_driver}(b)$	Node	0.014
X_{11}	$ \text{weekend_driver}(a) - \text{weekend_driver}(b) $	Node	0.019
$X_{12}-X_{19}$	$\text{last_week}_l(a+b)$ for $l = 1, \dots, 8$	Spatio-temporal	0–0.027
$X_{20}-X_{27}$	$\text{last_month}_l(a+b)$ for $l = 1, \dots, 8$	Spatio-temporal	0–0.057
$X_{28}-X_{45}$	$\text{last_year}_l(a+b)$ for $l = 1, \dots, 8$	Spatio-temporal	0.010–0.060
X_{46}	$ \Gamma(a) + \Gamma(b) $	Neighbourhood	0.117
X_{47}	$ \Gamma(a) - \Gamma(b) $	Neighbourhood	0.013
X_{48}	$ \Gamma(a) \cup \Gamma(b) $	Neighbourhood	0.093
X_{49}	$ \Gamma(a) \cap \Gamma(b) $	Neighbourhood	0.021
X_{50}	$s_a + s_b$	Neighbourhood	0.056
X_{51}	$ s_a - s_b $	Neighbourhood	0.017
X_{52}	Shortest path length in G	Path	0.111

forest is used as implemented in Python sci-kit learn 0.21.2. We find an optimal value of three for the maximum depth of the decision trees in the random forest.

We report the trade-off between true positives and false positives to assess the accuracy of the classifier. The relation between these two values can be shown using the well-known Receiver Operator Characteristic (ROC)-curve [10], shown in Fig. 6. The area under the ROC curve (AUROC) is 0.84. This value indicates that the classifier is able to accurately select the links that will appear. This allows us to analyze the feature importance.

The feature importance of each individual feature is provided in Table 3. In Fig. 7 the features are shown for each of the aforementioned four types. We observe that the neighbourhood features score highest, closely followed by the single path feature. The two neighbourhood features with the highest scores are X_{46} and X_{48} . These features provide the sum of the degrees of the node pairs and of the union of their neighbourhoods, respectively. Both the spatiotemporal and node features score lower.

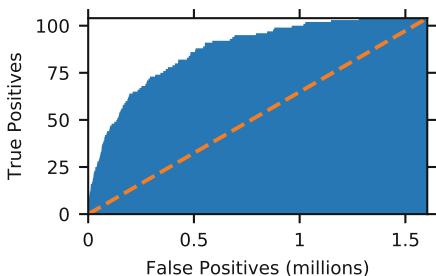


Fig. 6. Receiver Operator Characteristic curve of the random forest link prediction classifier. The AUROC is 0.84.

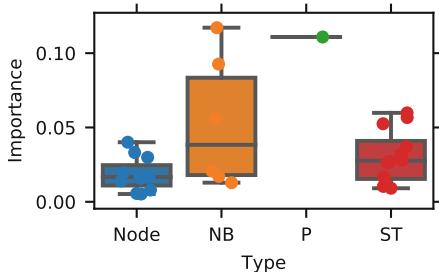


Fig. 7. Gini feature importance. NB, P and ST are the neighbourhood, path and spatiotemporal features resp.

Since the features based on network metrics have a higher feature importance, from these experiments we conclude that the network perspective on this data is fruitful. A next step is to interpret these findings in the infrastructure domain; we leave this step for future work.

7 Conclusions and Outlook

In this work we assessed to what extent the dynamics of co-driving behaviour are stimulated by the characteristics of the nodes involved. The research has shown that features based on network measures are able to explain the dynamics of the studied co-driving network. This means that the network perspective on this spatio-temporal dataset of truck driving in the Netherlands is meaningful and worth exploring further. Our findings may also suggest that, in general, the link prediction approach is suitable to analyze spatiotemporal datasets containing social behaviour.

In future work, we will extend the devised link prediction approach to other datasets in the infrastructure domain. This will allow us to investigate how different type of features perform in different contexts. A second interesting angle is to use a similar approach to predict which nodes will turn inactive, i.e., will not form any new links. This will result in a substantially smaller set of candidate nodes for the link prediction algorithm. Finally, future work will focus on interpreting and applying the knowledge observed from the micro-level dynamics in the infrastructure domain.

References

1. Barabási, A.L.: Network Science. Cambridge University Press, Cambridge (2016)
2. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999)

3. de Bruin, G.J., Veenman, C.J., van den Herik, H.J., Takes, F.W.: Understanding behavioral patterns in truck co-driving networks. In: International Conference on Complex Networks and their Applications, pp. 223–235. Springer (2018)
4. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends Comput. Graph. Vis. **7**(2–3), 81–227 (2012)
5. Cuttone, A., Lehmann, S., González, M.C.: Understanding predictability and exploration in human mobility. EPJ Data Sci. **7**(1), 2 (2018)
6. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York (2009)
7. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. Science **311**(5757), 88–90 (2006)
8. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
9. Mazurek, U., van Hattem, J.: Rewards for safe driving behavior: influence on following distance and speed. Transp. Res. Rec. **1980**(1), 31–38 (2006)
10. Oshiro, T.M., Perez, P.S., Baranauskas, J.A.: How many trees in a random forest? In: International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 154–168. Springer (2012)
11. Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D., Giannotti, F.: Understanding the patterns of car travel. Eur. Phys. J. Spec. Top. **215**(1), 61–73 (2013)
12. Probst, P., Wright, M.N., Boulesteix, A.L.: Hyperparameters and tuning strategies for random forest. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **9**(3), e1301 (2019)
13. Sapiezynski, P., Stopczynski, A., Gatej, R., Lehmann, S.: Tracking human mobility using wifi signals. PloS one **10**(7), e0130824 (2015)
14. Sekara, V., Stopczynski, A., Lehmann, S.: Fundamental structures of dynamic social networks. Proc. Natl. Acad. Sci. **113**(36), 9977–9982 (2016)
15. Tsugawa, S., Kato, S.: Energy ITS: another application of vehicular communications. IEEE Commun. Mag. **48**(11), 120–126 (2010)
16. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabási, A.L.: Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1100–1108. ACM (2011)
17. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. Nature **393**(6684), 440 (1998)



Characterizing Large Scale Land Acquisitions Through Network Analysis

Roberto Interdonato^{1(✉)}, Jeremy Bourgoin¹, Quentin Grislain¹,
Matteo Zignani², Sabrina Gaito², and Markus Giger³

¹ Cirad, TETIS,
TETIS, Univ. of Montpellier, APT, Cirad, CNRS, Irstea, Montpellier, France
roberto.interdonato@cirad.fr

² Università degli Studi di Milano, Milan, Italy

³ Centre for Development and Environment (CDE),
University of Bern, Bern, Switzerland

Abstract. Large Scale Land Acquisitions (LSLAs) by private companies or states have seen a sudden increase in recent years, mainly due to combined and increasing demands for biofuel (i.e., caused by the increase in oil prices) and food (i.e., caused by the increase in world population and changes in dietary habits). These highly controversial phenomena raise many questions about production models, people's rights, resource governance, and are often at the root of conflicts with local populations. A valuable source of open access information about LSLAs, which fosters the study of such phenomena, is the database collected by the Land Matrix initiative. The database lists land deals of at least 200 ha and details for example, their nature (e.g. agriculture, infrastructure, mining), their current status (e.g. ongoing, abandoned, pending), and the investing companies. The information about land deals collected in the Land Matrix database comes from heterogeneous sources such as press articles, government data, individual contributions and scientific publications. In this work, we focus on a land trade network built upon the Land Matrix data about top companies and target countries related to each deal in the database. Modeling the information about LSLAs in a land trade network allows us to leverage on network analysis techniques, which will help to characterize land acquisition deals from an original point of view. In order to take a first step in this direction, we provide: (i) a centrality based analysis of the land trade network, including an analysis based on the correlation of centrality measures with different country development indicators, and (ii) an analysis based on network motifs (i.e., recurring, statistically significant subgraphs), which provides an insight into higher order correlations between countries, thus providing fresh knowledge about recurring patterns in transnational land trade deals.

Keywords: Large scale land acquisition · Network motifs · Network analysis · Land trade network

1 Introduction

Land acquisitions are not new as a phenomenon, but they have received important attention in recent years, and more specifically since the combined financial and food crisis which took place in 2007/2008. This period triggered an important increase in large-scale national and transnational commercial land transactions, or Large-Scale Land Acquisitions (LSLA) [3]. For instance, the increase in world food prices which created a global crisis (i.e., political and economic instability) also resulted in an increase in food demand, worsening an already dramatic situation caused by increasing population and changes in dietary habits. LSLAs are a highly controversial phenomenon, which raises many questions about production models, people's rights, resource governance, and are often at the root of conflicts with local populations. In fact, while proponents highlight the economic opportunities that these investments could lever, a number of authors and NGOs warn against the risks of corruption, and other threats to the rural poor's livelihoods, including loss of land and a progressive marginalization [4, 8, 10]. Large-scale land acquisitions have seldom been fairly negotiated with the farmers, and are often connected to unfair trade arrangements, thus recalling the colonial power asymmetries between the global North and the global South [1]. Whether the phenomenon is not new and rural communities have lived for centuries with insecure land rights, authors stress that the rate of large-scale acquisitions is increasingly jeopardizing people's access to land (see [2]). Researchers and NGOs rapidly underlined the issue of transparency regarding land deals and the urgent need of relevant and accountable data on their forms and dynamics [7]. In many countries, established procedures for decision-making on land deals do not exist, and negotiations and decisions do not take place in the public realm. Unofficial (e.g. NGO assessments) and official data sources at the country level often show discrepancies, and none may actually reflect reality on the ground. This acknowledgement led the International Land Coalition and other partners to launch the Land Matrix Initiative in 2009 to provide multi-source and open access information about LSLAs¹. The information collected in the Land Matrix database comes from heterogeneous sources such as press articles, government data, individual contributions and scientific publications. Land deals captured by the database are related to land over 200 ha that is being leased or sold [2].

While the interest in the initiative, and the number of deals contained in the database, are constantly growing, the data have rarely been used to extract new knowledge through data mining and network analysis techniques, which would allow better understanding the dynamics of this complex phenomenon, and to characterize relations between countries. To the best of our knowledge, the only examples are the work by Seaquist et al. [19] and Mechiche-Alami et al. [13]. The work of Seaquist et al. (2014) mostly consists in the analyses of basic structural characteristics (e.g., betweenness, clustering coefficient, assortativity) of a

¹ <https://landmatrix.org>.

land trade network obtained by merging data from Land Matrix and GRAIN² (another LSLA database built upon from scientific publications and press articles). However, the current version of the Land Matrix is way richer than the one from 2012 used in [19] (i.e., 2,809 transnational deals vs 1,006), so that the current land trade network is more complex and structurally different, making most analyses carried out in [19] obsolete. The work engaged by [13] is rather focused on the identification of different phases of trade activity and network dynamics, but not specifically on the relations between countries. Moreover, we do not consider only countries and deal sizes, but we also acknowledge the heterogeneity of these deals, namely their implementation status and purpose.

The aim of this work is to move the analysis of the data contained in the Land Matrix one step forward, by modeling up-to-date information about LSLAs in a land trade network, and providing a quantitative analysis consisting of two main steps:

- a centrality based analysis of the land trade network, including a study based on the correlation of centrality measures with different country development indicators;
- an analysis based on network motifs (i.e., recurring, statistically significant subgraphs), which provides an insight into higher order correlations between countries, thus providing fresh knowledge about recurring patterns in transnational land trade deals.

Motifs are induced subgraphs that occur significantly more often in an observed network than would in a randomized network with same network properties. They help revealing the existence of underlying non-random structural or evolutionary design principles that might have been involved in growing the network. Network motifs have been firstly leveraged in biological networks, where it has been shown that a small set of network motifs appears to serve as basic building blocks in [14], but recently have attracted attention as a tool for studying many different networks. For sake of example, network motifs, even temporal annotated, have been used to characterize communication patterns, homophily and information cascades in social networks, to perform fraud detection in financial networks or to identify special connections among firms in economic and financial networks. The motivation for carrying out an analysis of the land trade network based on network motifs is that this will allow to discover higher order correlations between the entities (i.e., investing and target countries), thus providing new insights about the dynamics of the LSLA phenomena. For instance, specific motifs may be highly correlated to notable statuses of the implementations (e.g., abandoned ones), or to trades between specific countries.

The rest of the paper is structured as follows: in Sect. 2 we introduce the Land Matrix land trade network, in Sect. 3 we present the centrality based analysis of the network, while in Sect. 4 we present the analysis based on network motifs. Section 5 concludes the work and discusses future directions.

² www.grain.org.

Table 1. Structural characteristics of the Land Matrix land trade network.

#nodes	#edges	reciprocity	avg_path_length	avg_clustering_coeff	transitivity	assortativity
161	956	0.01	2.45	0.20	0.06	-0.10

2 The Land Matrix Land Trade Network

The Land Matrix (LM) is a global and independent initiative for monitoring land deals. It is facilitated by a partnership of organizations concerned by decision-making over large-scale land deals, their implications for communities and the environment, and the fact that many directly affected stakeholders are currently excluded from such decision-making. The Land Matrix provides a tool for widening citizen involvement in making data available and understandable, thus promoting transparency and accountability [4]. The Beta version of the Global Observatory was launched by the Land Matrix in April 2012, with the aim of creating a reliable source of data to feed debate and provoke informed action on large-scale land deals [15]. It is important to keep in perspective that, by nature, the Land Matrix does not provide a complete and comprehensive database of land deals through time. While it encourages participation and contribution, data is only available for cases that have been reported.

In order to provide a network analysis based characterization of the LSLA phenomenon, we introduce here the directed Land Matrix land trade network. In this network nodes represent countries, and an edge (u, v) means that a company from country u has at least a land trade deal involving country v as target country. Edge weights model the total size (in hectares) of deals between the two countries (i.e., sum of the deal sizes between the two countries). In the context of this work, we will take into account all the transnational deals included in the Land Matrix database (i.e., all deals except for the national ones). Structural characteristics of the land trade network are reported in Table 1. It can be noted how the network shows a relatively dense connectivity 0.04, with an average path length computed on the undirected graph of 2.4, which accounts for the compactness of the network, as also supported by the average clustering coefficient (computed on the undirected network) of 0.2. The reciprocity of the network is equal to 0.01 indicating that most links are not likely to be bidirectional. This low value of the reciprocity gives a quantitative measure of the colonial power asymmetry phenomenon which clearly emerges from the land acquisition database. This asymmetry impacts also on the transitivity of the directed network, which is rather low (0.06), and on the number of strongly connected node pairs (21%, for a directed average path length of 2.95), both indicating that paths are not likely to be bidirectional. To deepen this latter aspect, in Sect. 4 we perform a census analysis of triads and more, providing the actual deal transfers among subsets of countries. Finally, it can be noted how the network is slightly disassortative (i.e., with an assortativity of -0.1), which is expected due to the divide between North and South (and more in general between *poor* and *rich* countries).

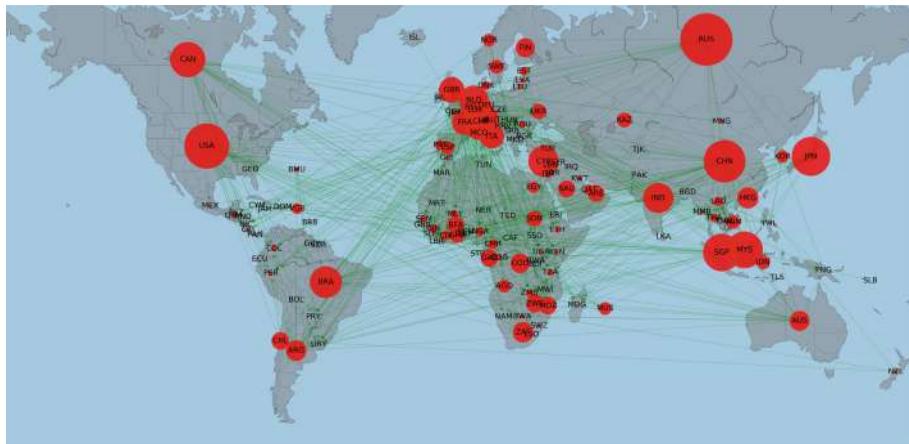


Fig. 1. Land trade network with node size proportional to the total surface (ha) of outgoing deals, taking into account only deals in operation.

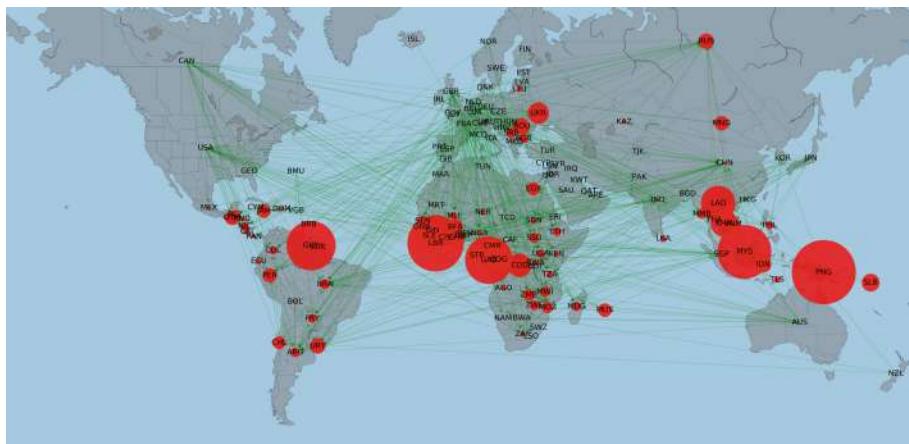


Fig. 2. Land trade network with node size proportional to the ratio sum of deals in production divided by the total of agricultural lands for target countries.

3 Centrality Based Analysis

Figures 1 and 2 show two different views of the network³. In Fig. 1 node size is proportional to the weighted outdegree of the nodes (i.e., total size of outgoing deals), i.e., bigger nodes indicate countries acquiring a larger quantity of land through transnational deals (countries *buying* land). Conversely, Fig. 2 reports a node size proportional to the ratio of the weighted indegree of each country

³ The maps were obtained using the *Basemap* and *networkx* python libraries.

(i.e., total size of ingoing deals) divided by the total of agricultural land of each country (countries *selling* more land), i.e., bigger nodes indicate countries with a large proportion of land acquisitions. The choice to show this ratio instead of a simple *indegree* is driven by the observation that by doing so we can emphasize the entity of the deals by also taking into account the available land in each country. In order to obtain fair and significant results, in both figures we took into account only deals that are marked as *in operation* in the Land Matrix, i.e., disregarding the abandoned ones and the deals already in the startup phase (not yet in production).

As pictured by Fig. 1, a worldwide network of companies are investing in foreign lands. This first acknowledgement allows us to put in perspective the common divide between North and South. Nevertheless, the fact that deals are originating from all over the world should not overcast the important correlation between the size of deals in operation and the location of investors. Indeed, biggest investors originate from G20 countries and/or are strongly linked to Paris Club, a group of 22 creditor countries coordinating loans to countries in financial difficulties. The importance of BRICS can also be observed, associating five major emerging national economies, namely Brazil, Russia, India, China and South Africa. Also, China (through companies and State) has been investing massively since 2010, mainly outsourcing agricultural production to Russia, Guyana, Congo Democratic Republic and Mozambique. Other African countries, Mali and Madagascar have been targeted for biofuel investments. China is also investing in South-East Asia, mainly in the forestry sector and in Vietnam, Myanmar, Lao PDR, Cambodia and Indonesia. This sector is also reaching South America with investments in Brazil and Bolivia for example.

Figure 2 highlights the fact that not only land deals are targeting countries in the South, but in some countries, they account for a large share of total agricultural land. On the one hand, we can hypothesize that foreign investments in these poor countries will foster drastic increase in food production and achieve food security, as the major part of donors and scholars stress the key role of the private sector and foreign investments in the process to closing the yield gap and revitalizing agricultural production through agribusiness-led development [6, 9, 12, 18]. Other authors also underline the benefits of foreign investments for poor people to improve their livelihoods and grow out of poverty at the local scale, generating jobs and creating opportunities for smallholders, while respecting the right of local communities and protecting environment [5, 6]. This situation raises other issues regarding sovereignty and a number of scholars have specifically underlined the potential risks caused by the global land acquisition phenomenon [17, 20]. In fact foreign agricultural investments could result in “enclaves of advanced agriculture” offering little benefit to the host nations and “resulting in purely extractive neo-colonialism” [1, 11]. Looking at Liberia, the country experienced a drastic increase in land deals in 2007–2008, with investments from Europe (Italy, UK, Luxembourg), Asia (Singapore, Malaysia, China) and other West African countries (Nigeria, Ivory Coast) in agriculture and forestry. We also note Canadian interest in the mining sector. Since the end of the civil war in 2003, Liberia

has been struggling to enact a land law, which, coupled with weak land and natural resources management, contributed to tenure insecurity and this increased rate of large-scale land concessions to private investors. In an attempt to address this issue, the Liberian government started a land reform process in 2009 with the establishment of a Land Commission. Since then the investment activity registered by the Land Matrix has drastically slowed down.

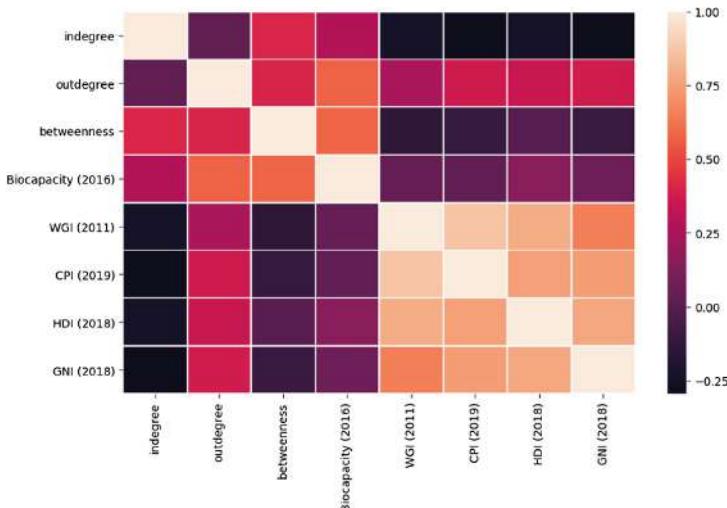


Fig. 3. Heatmap showing the correlation (pearson correlation coefficient) of the indegree, outdegree and betweenness centrality of the land trade network with different country development indicators.

3.1 Correlation with Country Development Indicators

To further characterize the role of different countries in the network, we perform a correlation analysis which aims at putting the network activity in perspective using different indicators describing specific characteristics of each country. We used different indicators to encompass social, environmental, economic and governance dimensions. The different indicators are described as follows:

- Biocapacity is an indicator of the Global Footprint Network, which calculates the area of productive land available to produce resources or absorb carbon dioxide waste, given current management practices. It is measured in standard units called global hectares.
- The Worldwide Governance Indicator (WGI) reports aggregate and individual governance indicators for over 200 countries and territories over the period 1996–, for six dimensions of governance; voice and accountability, political stability and absence of violence, government effectiveness, regulatory quality, rule of law and control of corruption.

- Corruption Perceptions Index (CPI) is an indicator of Transparency International. The index, which ranks 180 countries and territories by their perceived levels of public sector corruption according to experts and business people, uses a scale of 0 to 100, where 0 is highly corrupt and 100 is very clean. More than two-thirds of countries score below 50 on this year's CPI, with an average score of just 43.
- Human Development Index (HDI) was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not economic growth alone. It is a statistic composite index of life expectancy, education, and per capita income indicators.
- Growth National Income (GNI) calculates the total income earned by a nation's people and businesses, including investment income, regardless of where it was earned. It also covers money received from abroad such as foreign investment and economic development aid.

Figure 3 shows a heatmap reporting the correlation (Pearson's correlation coefficient) of the *indegree*, *outdegree* and *betweenness* centrality of the land trade network with respect to such indicators.⁴ From what can be observed in Fig. 3, land deals are directed towards countries experiencing low development rates (economic, social, governance). This is displayed by the negative correlation between *indegree* and the last four indicators (WGI, CPI, HDI, GNI), showing correlations in the range $-0.29 \leq \rho \leq -0.24$. Investing countries are characterized by higher CPI, HDI and GNI, with positive correlations between the *outdegree* and such indicators in the range $0.34 \leq \rho \leq 0.37$. The *biocapacity* indicator suggests that land deals originate from countries with higher levels of *biocapacity* than countries where deals are materialized and in operation. Outsourcing primary production to foreign countries seems to have a direct negative externality, measured by this environmental indicator. In this regard, it is interesting to note how *betweenness* shows a high positive correlation with this indicator (0.58), indicating that countries with a high biocapacity may serve as a sort of *flow hubs* in the network. This leads to two observations: (i) their land is involved in transnational deals, but at the same time they have enough economic power to invest themselves in transnational deals, (ii) these (ingoing and outgoing) investments happen in separated markets, i.e., between countries not directly connected between them.

4 Analysis Based on Network Motifs

In this section we present an analysis of the land trade network based on network motifs. The idea of network motifs was introduced by Milo *et al.* [14] in biological networks to identify recurrent over-represented patterns of interaction, which may correspond to the functional or organizational building blocks of the network. In the last years the analysis of network motifs has been applied to different kinds of networks, but with the same purpose, i.e. to highlight the

⁴ Heatmap was produced using the *seaborn* python library.

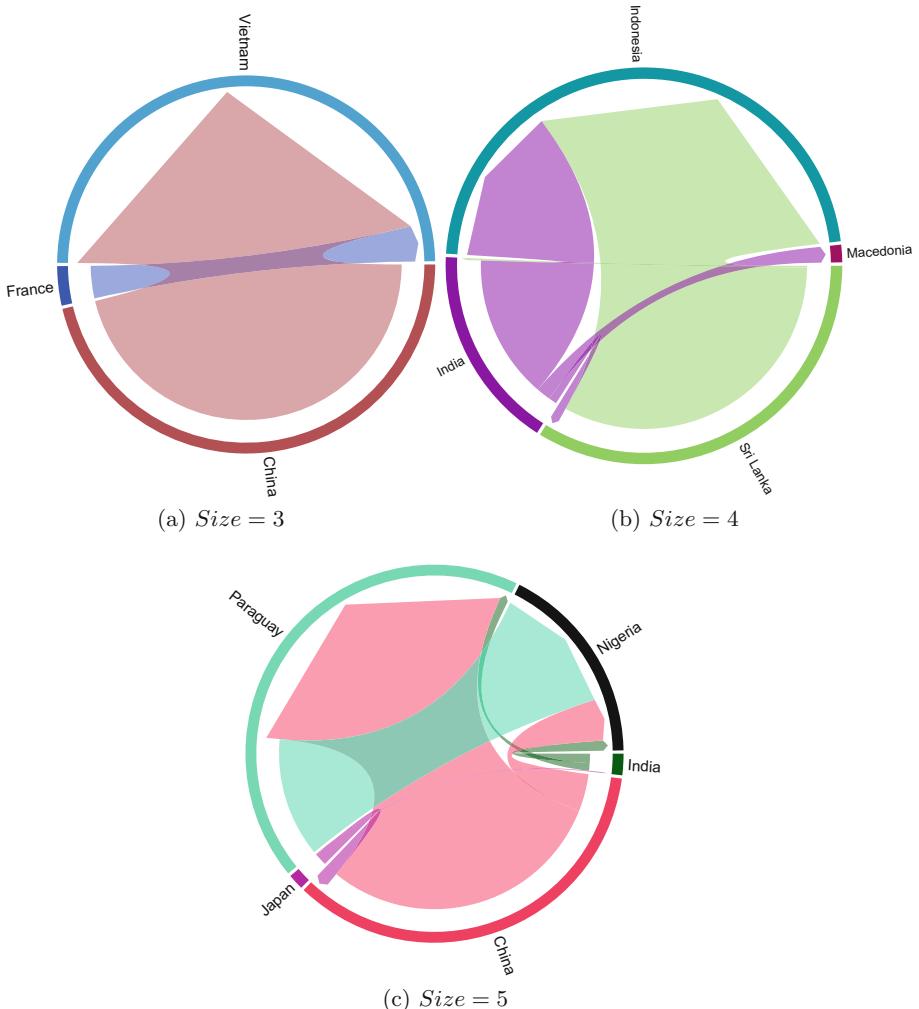


Fig. 4. Chord diagrams visualizing examples motifs of size 3, 4 and 5, with edges sizes proportional to edge weights (i.e., total size of the deals between two countries).

basic building blocks which characterize the network. Here we analyze directed 3-, 4- and 5–network motifs on the directed land trade network, neglecting the edge weights. Since the size of the land trade network is relatively small, we apply an exact network motif algorithm to calculate the census of all subgraphs of size 3, 4 and 5. Specifically, we adopt the state-of-art solution proposed in [16] based on *g-tries* to do the census of all the k -subgraphs in the original network and in the random networks used to assess the significance of each subgraph. In our case, we generate 100 random networks preserving the in and out degree

and we measure the statistic significance of a subgraph D_k by its *z-score*:

$$z\text{-score}(D_k) = \frac{f(D_k)^G - \hat{f}_{rand}(D_k)}{\sigma(f_{rand}(D_k))} \quad (1)$$

where $f(D_k)$ represents how many times the subgraph D_k occurs in the network G , while \hat{f}_{rand} and $\sigma(f_{rand})$ correspond to the mean and standard deviation of frequency of D_k in the random networks. Since we are interesting in over- and under-represented network motifs, we focus our analysis on network motifs with $|z\text{-score}(D_k)| \geq 1.5$ and $f(D_k)^G \geq 4$.

Even though we performed a complete analysis of statistically significant network motifs of size 3, 4 and 5 on the land trade network. Here we discuss an example for each size. Figure 4⁵ reports chord diagrams visualizing example motifs of size 3, 4 and 5, with edge sizes proportional to edge weights (i.e., total size of the deals between two countries). The aim of Fig. 4 is to represent the interconnection between countries by insisting on the globalization of deals.

The example motif of size 3 (Fig. 4a) shows a situation including two investing countries and one target country. This motif is rather common in our network, with a frequency of 5747 and a *z-score* of -1.76 . In the example we show an instance involving France and China as investing countries and Vietnam as target country. By investigating on the nature of the deals showed in the example, we found that more than 95% of investments are coming from China. France is investing in Biofuels when China is diversifying investments in agriculture (2 deals), mining (1 deal) and industry (8 deals). It is important to note that the mining deal accounts for 96% of total land invested.

The example motif of size 4 (Fig. 4b) represents a pattern involving two target countries (Indonesia and Macedonia in the example) and two countries that are both investing and target countries (India and Sri Lanka in the example). This motif has a frequency of 100 and a *z-score* of -1.79 . The example network shows massive investments in food crops from Sri Lanka to Indonesia (11 deals which account for more than 128 000 ha). The former is receiving investments from India (investments in tourism and agriculture). India is investing in all countries of the network (e.g. Macedonia in livestock and tourism and Sri Lanka in agriculture and tourism).

As concerns the example motif of size 5 (Fig. 4c), it represents five countries: one target country (Nigeria in the example), one investing country (Japan in the example) and three countries that are both investing and target countries (China, India and Paraguay in the example). This motif shows a frequency of 50 and a *z-score* of 5.66. In the situation shown in the example, Japan is investing in two countries, in industries in India and agriculture and plantations (exploitation and carbon sequestration) in China. The latter is massively investing in mining in Paraguay (185 000 ha), which in return invests extensively in Nigeria for food crops (77 000 ha). India also invests in Nigeria for Biofuel and crops (up to 6000 ha).

⁵ Chord diagrams were produced by using the *circlize* R library.

5 Conclusion

In this work, we focused on the analysis and characterization of large scale land acquisitions by leveraging on network analysis techniques. We introduced a land trade network built upon the information about relations between companies and target countries extracted from the public Land Matrix database, and we carried out two analysis stages in order to deepen our knowledge of such phenomenon: a first one based on centrality measures, and a second one based on network motifs. The results show that (i) land investment dynamics are global (all continents are affected) and (ii) both North-South dynamic (e.g. France invests in Vietnam) and South-South dynamic (e.g. Paraguay invests in Nigeria) hold in this context. Nevertheless, the analysis also showed how there is a strong correlation with country development indicators, with the biggest investing countries characterized by higher CPI, HDI and GNI (i.e., G20 countries). In addition, investments are characterized by the diversity of the purposes (biofuel, agriculture, tourism, industry, etc.), the area of land purchased (from 200 ha up to 185,000 ha) and the number of deals.

Since this work represents a first step towards analysis and characterization of LSLAs through network analysis, several future directions are open. First of all, the qualitative investigation of network motifs should be deepened. For instance, specific motifs may be highly correlated to notable statuses of the implementations (e.g., abandoned ones), or to trades between specific countries. Moreover, by exploiting temporal information about the deals, we would be able to discover temporal sequences associated to specific categories of deals (e.g., successful/unsuccessful ones). Another direction regards the modeling of the network itself, since the detailed information about land deals available in the Land Matrix allows to build different networks by modeling complex relations among different commercial actors, such as investors, operating companies and top companies, going beyond the land trade network used in this work. As a final remark, this work will also pave the way for the application of text mining and network analysis techniques to extract additional information about LSLAs from heterogeneous sources, such as publicly available ones (e.g., newspapers, websites, social media) and data issued by investigative journalism (paradise papers, panama papers), which may help us to discover deeper relations between countries involved in land trade events.

References

1. Adbib, R.: Large scale foreign land acquisitions: neoliberal opportunities or neo-colonial challenges? (2012)
2. Anseeuw, W., Boche, M., Breu, T., Giger, M., Lay, J., Messerli, P., Nolte, K.: Transnational land deals for agriculture in the global south. Analytical report based on the land matrix database (2012)
3. Borras, S.M., Hall, R., Scoones, I., White, B., Wolford, W.: Towards a better understanding of land-grabbing. An editorial introduction. *J. Peasant. Stud.* **38**(2), 209–2016 (2011)

4. Borras, S.M., Hall, R., Scoones, I., White, B., Wolford, W.: The rush for land in Africa: resource grabbing or green revolution? *S. Afr. J. Int. Aff.* **20**(1), 159–177 (2013)
5. Byerlee, D., Garcia, A.F., Giertz, A., Palmade, V.: Poverty and the globalization of the food and agriculture system. In: *The Poorest and Hungry: Assessments, Analyses, and Actions: An IFPRI 2020 Book* (2007)
6. Byerlee, D., Garcia, A.F., Giertz, A., Palmade, V.: *Growing Africa. Unlocking the potential of agribusiness* (2013)
7. Cotula, L.: The international political economy of the global land rush: a critical appraisal of trends, scale, geography and drivers. *J. Peasant. Stud.* **39**, 649–680 (2012)
8. Cotula, L.: Human rights, natural resource and investment law in a globalised world: Shades of grey in the shadow of the law (2013)
9. FAO/UNIDO: Id3a, initiative pour le développement de l'agri-business et des agro-industries en afrique, page 38 (2010)
10. GRAIN: Accaparement des terres et souveraineté alimentaire en afrique de l'ouest et du centre. *A contre courant* (2012)
11. Hallam, D.: International investments in agricultural production (2009)
12. Konig, G., Da Silva, C.A., Mhlanga, N.: Enabling environments for agribusiness and agro-industries development. regional and country perspectives (2013)
13. Mechiche-Alami, A., Piccardi, C., Nicholas, K.A., Seaquist, J.W.: Transnational land acquisitions beyond the food and financial crises. *Environmental Research Letters*, page 18, in Press
14. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. *Science* **298**(5594), 824–827 (2002)
15. Nolte, K., Chamberlain, W., Giger, M.: International land deals for agriculture. Fresh insights from the land matrix: Analytical report ii (2016)
16. Ribeiro, P., Silva, F.: G-tries: an efficient data structure for discovering network motifs. In: *Proceedings of the 2010 ACM Symposium on Applied Computing, SAC 2010*. ACM, New York (2010)
17. Robertson, B., Pinstrup-Andersen, P.: Global land acquisition: neo-colonialism or development opportunity? *Food Secur.* **2**(3), 271–283 (2010)
18. Rulli, M.C., D'Odorico, P.: Food appropriation through large scale land acquisitions. *Environ. Res. Lett.* **9**, 8 (2014)
19. Seaquist, J.W., Johansson, E.L., Nicholas, K.A.: Architecture of the global land acquisition system: applying the tools of network science to identify key vulnerabilities. *Environ. Res. Lett.* **9**(11), 114006 (2014)
20. von Braun, J., Meinzen-Dick, R.: Land grabbing by foreign investors in developing countries. *Risks and opportunities* (2009)



A Network-Based Approach for Reducing Test Suites While Maintaining Code Coverage

Misael Mongiovi^(✉), Andrea Fornaia, and Emiliano Tramontana

Dipartimento di Matematica e Informatica, University of Catania, Catania, Italy
`{mongiovi,fornaia,tramontana}@dmi.unict.it`

Abstract. An effective test suite can be desirable for its ability to protect the code when introducing changes aiming at providing further functionalities, for refactoring, etc. An ample test suite, while taking considerable resources and time to execute, could be as effective as a smaller one having the same code coverage. This paper proposes an approach to reduce the number of test cases needed while ensuring the same code coverage of a large test suite. The approach is totally automatic, comprising test cases generation, code coverage computation, and reduction on the number of test cases. For finding the minimum subset of test cases, the code coverage problem has been formulated as a Set Cover problem. Our solution is based on Integer Linear Programming and on the properties of the Control Flow Graph. We evaluated the proposed solution on four sample software systems. The results show a drastic reduction on the number of tests, hence their execution time.

Keywords: Test cases · Static analysis · Graph analysis

1 Introduction

While a software system is being developed, good practices prescribe that an ample test suite is made available so that the correctness of the implementation can be checked [3, 4]. Test cases can be manually implemented, or can be generated by means of automatic tools. Test generation tools for Java that generate test code calling methods include Randoop [13], EvoSuite [8], and AgitarOne [1]. Other tools, such as e.g. Coffee4J [2], are available for generating input values for tests, given some parameters, and according to combinatorial test. Automatic test generation tools can be very handy in that developers can be freed from the task of implementing a test suite, hence shortening development time.

Randoop test generation creates sequences of methods to be called, selecting from available methods randomly [13]. EvoSuite generates test cases consisting of calls to methods, by using evolutionary computation over a population, where a population is a test suite [8]. While such a generation can be extremely useful to check contract violation and for regression tests, the number of test cases

generated can be very large, given the many possible ways in which available methods can be combined together, even for a software system consisting in a small number of methods. However, in general the number of generated test cases is not proportional to code coverage, and when the number of test cases increases, the percentage of code coverage, as well as their efficacy, could remain largely unchanged [16]. E.g. in our experiments, when generating tests using Randoop, 64 generated test cases ensured about 24% code coverage and 128 generated test cases a 25% for a small-size system (JUnit 4).

When following Agile practices, regression tests are executed very often, as soon as new code has been developed and integrated, hence for a collaborating team of a dozen people [7], tests could be run every couple of hours or so, requiring hardware resources and time. When generated test cases are of the order of thousands, the time needed to run them could be not compatible with the development practice of continuous integration, since the feedback from test runs becomes not timely. We can observe that from an ample amount of generated test cases, some subsets run the same code, e.g. because they use different values for parameters, or because several statements have been covered by multiple test cases. Hence, proper selection among tests could provide a more timely feedback to developers [16].

We propose a technique that automatically selects from an ample set of generated test cases a subset that can ensure the desired code coverage. E.g. we could set to have the maximum code coverage provided by the generated test cases and automatically select the minimum subset needed. We formulate the problem as an optimisation problem aimed at finding the minimum number of test cases that maintain the whole code coverage of a defined test suite. The approach can be straightforwardly generalised to cover an input defined set of code lines. In the remainder of this paper, we first describe our approach and framework (Sect. 2), then we describe our optimisation method (Sect. 3), and report experimental results (Sect. 4). Eventually, we discuss some related works and conclude the paper.

2 Proposed Approach and Framework

The proposed approach aims at selecting test cases, among a large suite, in order to run a minimum number of test cases while ensuring maximum coverage of code, considering the instruction-level. This is beneficial to reduce running time during regression testing.

Figure 1 shows a graph relating generated test cases and code coverage. As we can see, for toy examples having less than one-hundred lines when increasing the number of test cases the code coverage increases and goes to nearly 100%. For larger systems, having between 610 lines and 5216 lines (a chord client and junit4), code coverage increases very slowly when increasing the number of test cases. This lack of efficacy in terms of code coverage for each generated test case has been observed in other studies, and has often been related to a low level efficacy in terms of finding defects [16].

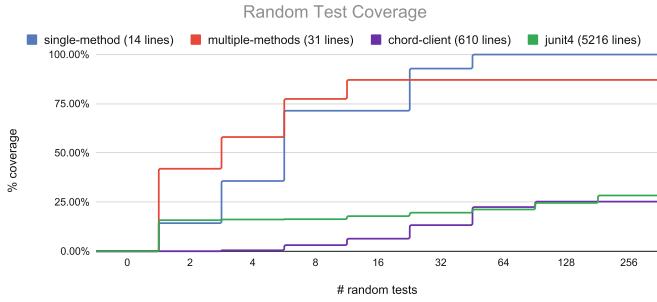


Fig. 1. Number of test cases and code coverage for four analysed systems.

A random testing tool, such as Randoop, automatically inspects the code of a Java class to generate unit tests. It tries to generate sequences of method calls in order to stress and violate the methods contracts, and check the related behaviour. Failing tests (i.e. with unchecked exceptions) may indicate potential errors that need to be checked by the developer. Passing tests (i.e. providing any method result) are instead used to augment a regression test suite capturing the current behaviour of the code, using the test execution feed-back to append result assertion to the test sequence.

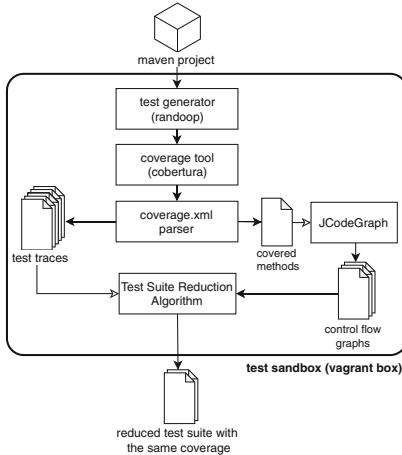


Fig. 2. The developed tool-chain consisting of test generation, code coverage assessment, control flow graph extraction, reduced test suite computation.

Figure 2 depicts the overall framework we created to automatically generate a test suite for a Java project. Starting from a maven project, a test generator tool, i.e. Randoop in our case, has been configured to create a large number of random test cases, hence potentially increasing code coverage. Such test cases

were then added to the maven project as Java unit tests. Since random tests can have unintended behaviour, to protect the execution environment, they have been executed in a properly created *sandbox* environment by means of a vagrant box¹, a declarative-configured virtual environment.

Test cases are then executed together with a tool tracing and reporting code coverage, i.e. cobertura², which has been included as a maven plugin for the reporting step. This has been used to collect the coverage of every single test case as a *test trace*. For each test trace, the related test case, the execution time and the covered lines were collected for supporting the further optimisation step. A list of covered methods has also been obtained (i.e. the methods called by at least one test case). This is given to a static analysis tool, called *JCodeGraph*, internally developed, to obtain the Control Flow Graph (CFG) for each covered method. Finally, both CFGs and the test traces are given as input to the reduction algorithm described in Sect. 3, which provides a reduced test suite with the same coverage rate of the initial one randomly generated by Randoop.

3 Problem Formulation and Method

Given a *test suite* $T = \{t_1, t_2, \dots, t_n\}$, where t_1, t_2, \dots, t_n are *test cases* (a test case is a specific configuration of the input parameters), we aim at finding a small (as small as possible) subset of test cases $T' \subseteq T$ that maintains *code coverage*, i.e. the set of instructions that are visited by executing all test cases in T . For each test case t we consider the set of instructions S_t which are executed by running test t (test trace). We say that a test case t *covers* an instruction i if instruction i is executed when t is run. We extends this concept to test suites and say that i is covered by a test suite T if i is covered by at least one test case in T . We call *universe* the set U_T of instructions covered by at least one test case in the test suite T , i.e. $U_T = \bigcup_{t \in T} S_t$.

The problem of optimising the test suite with no loss in coverage can be formulated as a Set Cover instance. Set Cover is a well-known problem in computer science. Given a family of sets, it calls for finding the minimum-size³ sub-family that covers all elements in the input family. An instance of Set Cover is given in Fig. 3. S_1, S_2, S_3 and S_4 compose the input family of sets, spanning a universe of 16 elements (black dots). In this example, the minimum-size sub-family that covers all elements consists of three sets ($\{S_1, S_2, S_3\}$), since there is no way to choose two sets that cover the whole universe (all 16 black dots).

To model our test optimisation problem, we take for each test case t a set S_t , which contains all instructions visited by test case t , and consider the family of sets corresponding to the whole test suite as an instance of Set Cover. We aim

¹ <https://www.vagrantup.com>.

² <https://cobertura.github.io/cobertura/>.

³ We consider the size of a family as its number of sets.

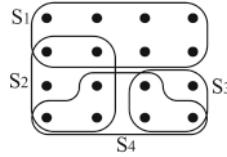


Fig. 3. An instance of Set-Cover. $\{S_1, S_2, S_3\}$ is the minimum-size subfamily that covers the whole universe of elements (black dots).

at finding the minimum-size subfamily that covers the whole set of instructions covered by the complete test suite. This approach guarantees the same code coverage of the original test suite while reducing considerably the number of tests. An exact solution of the Set Cover instance represents a minimal test set with coverage guarantee, where minimal indicates the impossibility to obtain the same coverage using a smaller number of sets drawn from the original test suite.

Set Cover is NP-complete and it has been proven not to have a constant factor approximation guarantee, i.e. the solution of an approximation algorithm cannot have an error within a fixed multiplicative factor of the returned solution. It can be solved by a greedy algorithm, which performs a factor- $\log(n)$ approximation, i.e. the error is bound by a logarithmic function of the solution [6].

A solution of the greedy algorithm could be employed as a smaller test suite representative of the whole test suite. Although its execution would perform no loss in coverage, the resulting test suite would not have any guarantee to be minimal and hence it would often require running more tests than necessary, therefore impacting the running time of the test execution. On the other hand, computing an optimal solution would be at best time consuming and often unfeasible, depending on the size of the test suite and the software project.

Our solution is somehow intermediate and is based on two hints. First, Set Cover can be formulated as an *Integer Linear Programming (ILP)* problem and solved with available open or commercial solvers, which perform a sequence of iterations aiming at improving the solution up to the optimal. If an optimal solution cannot be found in reasonable time, the system returns a solution within a certain error from the optimal. Second, we can take advantage of the *Control Flow Graph (CFG)* of the software project to reduce the size of the problem instance. Since the running time is over-linear on the size of the input, even a small reduction of the input size would produce a considerable improvement in terms of efficiency and accuracy of the Set Cover computation. In the remainder of this section we describe first the ILP solution, then the CFG-based optimisation.

3.1 ILP Solution for Test Optimisation

Set Cover can be formalised as an ILP problem as follows⁴:

$$\begin{aligned} & \text{minimise} \quad \sum_{t \in T} x_t, \\ & \text{subject to} \quad \sum_{t: i \in S_t} x_t \geq 1 \text{ for all } i \in U_T = \bigcup_{t \in T} S_t, \\ & \quad x_t \in \{0, 1\} \text{ for all } t \in T \end{aligned}$$

where T is the input test suite and $\{x_t\}$ are integer binary variables representing the set of chosen test cases. Specifically, the value of x_t is 1 if the test t is chosen, 0 otherwise. Solving the above ILP problem consists in assigning values to the $\{x_t\}$ variables in order to minimise the objective function (which is equivalent to choose as few sets as possible) and satisfy the constraints (which guarantee to have every instruction in the test suite T covered at least by one test).

Available solvers for ILP use a mixture of cutting-planes, branch-and-bound and dynamic programming techniques to attempt to find the exact solution and in any case return a solution with error within a bound. They use heuristics and relaxation to find lower bounds and upper bounds to the optimal solution and continue until both bounds coincide (an exact solution is found) or no further improvement is estimated to be accomplished in reasonable time. For this work we adopt GLPK⁵, one of the most known open-source solvers.

3.2 Reducing the Input Size by Means of CFG Analysis

A *Control Flow Graph (CFG)* is a static representation of all the possible execution flows of a program. We aim at exploiting the CFG for reducing the set of instructions to consider for the Set Cover input.

A CFG is a graph $G = (V_G, E_G)$ whose vertices (in V_G) are instructions of a program and two vertices are connected by a directed edge $(i, j) \in E_G$ if i and j can potentially be executed subsequently in some flow of the program. This includes contiguous instructions but also non-contiguous ones in the presence of conditional or loop statements. An example of CFG of a short program is shown in Fig. 4. The graph on the right represents the CFG of the program on the left. Vertices are labelled with the instruction line numbers. Node 4 corresponds to the first instruction; node 8 and 13 represent conditional statements and hence they have more than one outgoing edge. Nodes 21 and 22 describe a loop.

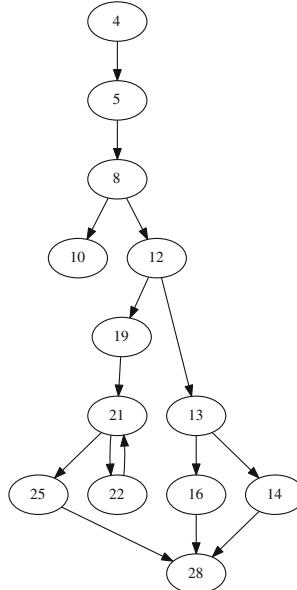
The underlying idea of the CFG-based input reduction is that we can guarantee that certain instructions are executed without including them in the Set Cover input. E.g., in the program shown in Fig. 4, if a test t covers instruction 10, we can be sure that test t also covers instructions 4, 5 and 8, since the only

⁴ For clarity of exposition we define the Set Cover ILP formalisation in the context of our test optimisation problem.

⁵ <http://www.gnu.org/software/glpk/>.

```

1 public class ShoppingCart {
2     double daily_discount = 0.20;
3
4     public double getTotal(int price, int qty,
5         int discount_level) {
6         double total = 0;
7
8         // input validation
9         if (price < 0 || qty < 0
10            || discount_level < 0 || discount_level >
11                3)
12             return -1;
13
14         if (price == 0) {
15             if (qty > 1) {
16                 total = -1;
17             } else {
18                 total = 0;
19             }
20         } else {
21             total = price * qty;
22
23             for (int i = 0; i < discount_level; i++) {
24                 total *= (1 - 0.10);
25             }
26
27             total += (1 - daily_discount);
28         }
29
30         return total;
31     }
32 }
```



```

1 @Test
2 public void test23() throws Throwable {
3     ShoppingCart shoppingCart0 = new ShoppingCart();
4     double double4 = shoppingCart0.getTotal((-1), 1, (-1));
5     double double8 = shoppingCart0.getTotal(1, 1, (int) (short) -1);
6     double double12 = shoppingCart0.getTotal(0, (int) (short) 0, 0);
7     java.lang.Class<?> wildcardClass13 = shoppingCart0.getClass();
8     // Regression assertions (captures the current behavior of the code)
9     org.junit.Assert.assertTrue("+"+double4+" != "+(-1.0d)+"", double4 == (-1.0d));
10    org.junit.Assert.assertTrue("+"+double8+" != "+(-1.0d)+"", double8 == (-1.0d));
11    org.junit.Assert.assertTrue("+"+double12+" != "+0.0d+"", double12 == 0.0d);
12    org.junit.Assert.assertNotNull(wildcardClass13);
13 }
```

Fig. 4. A sample class having several execution paths (on the left), with its control flow graph (on the right), and a Java unit test for it (on the bottom), automatically generated by Randoop, covering lines {4,5,8,10,12,13,16,28} of the ShoppingCart class.

path from the first instruction traverses such nodes. Therefore if instruction 10 is included in the Set Cover input, instructions 4, 5 and 8 can be excluded. Based on this idea we designed a simple algorithm, which iteratively removes instructions from an initial universe when its coverage can be guaranteed by a subsequent instruction in the CFG (see Algorithm 1).

The algorithm visits all edges (i, j) in the CFG G and removes sources of edges whose target is covered by the test suite and has exactly one incoming node. This guarantees that every path that reaches j must visit i and hence if j is covered i must necessarily be covered. Eventually the algorithm returns the

```

Result: Return a reduced input for the optimisation of test suite  $T$  based on CFG  $G$ 
 $U \leftarrow \bigcup_{t \in T} S_t;$ 
for every  $(i, j) \in E_G$  do
    if  $j$  belongs to  $U$  and has exactly one incoming node then
         $U \leftarrow U \setminus \{i\};$ 
    end
end
return  $\{S_t \cap U : t \in T \text{ and } S_t \cap U \neq \emptyset\};$ 

```

Algorithm 1: CFG-based input reduction for Set-Cover-based test

input family of sets for Set Cover cleaned by removing unnecessary elements from the sets and removing empty sets. In the example in Fig. 4, where we consider the universe U_T as composed by all vertices, the universe is reduced to $U = \{10, 19, 25, 22, 16, 14, 28\}$, since these vertices do not have a successor in U with exactly one incoming edge.

4 Experimental Results

In order to assess the performances of the proposed approach, we run our framework on a few samples software system in Java, whose details are given in Table 1.

We employed Randoop [13] for generating the initial test suite and JCode-Graph, a tool of ours based on JavaPDG [17] for computing the CFG. We set the number of tests to generate to 100 for every software project. We implemented the optimisation code in Java 10, employed GLPK⁶ as an ILP solver and made use of the library JGraphT⁷ for managing graphs.

Table 1 reports, from the left column the name of the sample software systems, the number of classes, the number of methods, the total number of executable lines of code, the coverage (as number of code lines that have been executed when running the test suite), the number of test cases in the initial test suite (generated by Randoop) and the number of test cases after optimisation. The size of the initial test suite is lower than 100 since some test cases are merged by Randoop. The proposed test optimisation manages to reduce considerably the number of test cases (e.g. from 55 to 2 in the best case, and from 73 to 53 in the worst case) while maintaining the same coverage.

In order to evaluate the impact of the CFG-based input reduction, we implemented two versions of our optimisation tool: the first, namely ILP, solves the ILP program without reducing the input, the second, namely ILP+CFG, employs the CFG-based input reduction and then solves the ILP program. Both versions find the optimal solution in fractions of seconds on all projects. Since with our data the running time cannot be compared, we compare the input size of both versions. Table 2 reports, for each version, the number of sets of the input family (i.e. the number of test cases), the number of elements of the universe (i.e. the covered instructions considered) and the input size. The input

⁶ <http://www.gnu.org/software/glpk/>.

⁷ <https://jgrapht.org/>.

Table 1. Coverage and number of test cases before and after the optimisation

Software	Classes	Methods	Lines	Coverage	Initial size	Reduced size
Single-method	1	2	14	14	55	2
Multiple-methods	3	9	31	28	59	4
Chord-client	11	61	610	143	73	25
JUnit4	347	1772	5216	1196	73	53

size for ILP is the total number of lines executed, i.e. the sum over all test cases of the number of statements executed by each test case (since a line of code – statement – happens to be executed multiple times by more than one test case). When we introduce the CFG optimisation, the input size is smaller since the total number of lines to be considered in the Set Cover computation is reduced, without affecting results. ILP+CFG performs a 68% reduction of the input size with respect to ILP on single-method and 42% reduction on multiple-method. For the two biggest software systems, chord-client, and junit4, the reduction is 66% and 56%, respectively.

Table 2. Sample software projects under analysis

Software	Sizes with ILP			Sizes with ILP+CFG			Reduction
	Family	Universe	Input size	Family	Universe	Input size	
Single-method	55	14	398	54	7	128	68%
Multiple-methods	59	28	577	59	17	336	42%
Chord-client	73	143	562	72	49	192	66%
JUnit4	73	1196	53423	73	545	23346	56%

5 Related Works

Randoop generates sequences of methods for a test case by selecting from available methods randomly [13]. This tool does not consider code coverage, hence generated tests could exercise the same portions of code. With respect to our contribution it is complementary, since it gives test cases while we select such test cases that cover the code without much repetition.

EvoSuite is a tool that achieves automatic generation of test cases, and their assertions, thanks to symbolic execution. It also aims at generating a test suite which ensures maximum coverage while minimising the number of test cases. The latter characteristic is the one giving EvoSuite more similarities to our work. However, each sequence of method calls generated by EvoSuite, which represents a population to be selected and bred if considered the most fit, is

executed in order to assess code coverage. This is time consuming; risky when having to interact with files, databases, and networks; and since their approach is evolutionary, the number of execution is not bounded nor small [8]. Our approach requires only one execution of the test cases, then the selection approach does not require any run.

In [15], the authors point out the need to attribute a priority to test cases, e.g. according to their ability to maximise code coverage. Our proposed solution is in that direction, and by using a reduced Set Cover, thanks to control flow knowledge, we manage to reduce the complexity cost when compared to the previous work.

The authors of [16] have shown that a small number of individual test cases, i.e. less than 1 over 5 test cases, are effective to find defects, when using test suite generators. Among other improvements, the authors highlight the need to further improve code coverage in order to execute faulty code. When employing our approach, generated test cases that do not cover further code are excluded, hence increasing the ratio of effective tests, and giving more opportunities to execute additional novel test cases in the same timeframe.

The authors in [12] firstly generate the test cases to cover all different subpaths in a program, then, since each test case can cover more than one subpath, they reduce the number of test cases by identifying which subpaths are covered by each test case. The test cases selected are found by a greedy solution, which finds in some cases the optimal solution. Compared to our approach, there are similarities on the coverage, though we consider the single line of code, instead of the subpath. As far as the solution is concerned, instead of using a simple greedy approach, we have proposed a solution that leverages on the knowledge of the control flow to reduce the number of code lines to consider, hence searching on a smaller solution space.

In [18], the proposed approach considers the requirements exercised by a test, and minimises the number of test cases by excluding a test case when requirements covered are also covered by other test cases. They introduce a solution called delayed-greedy consisting of some heuristics, such as checking whether a set of requirements is a superset or a subset of another, hence excluding the test cases having subsets of requirements; whether a requirement is covered by only one test case, then including such a test case; and whether the test case covers the maximum number of attributes, then including it. The main difference with our approach is that our approach focuses on code coverage, instead of requirements. Most development practices have only coarse-grained requirements, hence many portions of code are implemented for a requirement. As a result two (or more) test cases covering the same requirement could exercise different portions of code. Moreover, we automatically derive the code line covered by each test case, hence we need not rely on a manual mapping between test cases and requirements. Focusing on code coverage is harder due to the larger amount of data (code instructions) to handle. We take advantage of the properties of the Control Flow Graph to keep the problem feasible without penalising accuracy.

While in our approach code coverage has been used as the optimisation criteria, in [10] the authors mapped the Set Cover problem to a mutation testing problem. By using a greedy approach to find the minimum set of test cases that covers all the mutants killed by the original test suite, they were able to detect the same number of faults.

In [11] the authors propose a general ILP-based framework to support test-suite minimisation over multiple criteria, such as code coverage, test execution time and setup cost. Considered the generality of their approach, they don't make any assumption on the coverage data (Set Cover) to be used, which could be related either to covered requirements or statements. In our approach, we specifically leverage CFG information to reduce the number of statements to consider as a Set Cover.

The problem of reducing test suites have been considered by past work and often approached as a covering problem. Xu et al. [19] consider it as a variant of Set-Cover where costs are associated to sets (test cases) and the goal is to maintain the coverage of requirements by minimising the total cost. The authors apply a greedy algorithm to solve it. Although the approach could easily be extended to guarantee code coverage, in place of manually-curated requirements, it would penalise accuracy to guarantee feasibility. In contrast, we leverage on the properties of the CFG to feasibly solve the problem with high accuracy. Chen et al. [5] adopt an exact approach based on ILP and propose some heuristics to increase efficiency. Panda and Mohapatra [14] propose a an ILP-based approach for minimising the test suite without loosing coverage of modified sentences and meanwhile minimising a measure of cohesion of code parts covered by each test case. Again, this approach does not take advantage of the properties of the CFG for optimising the computation. FLOWER [9] describes a novel Maximum-Flow-based approach for reducing the test suite while maintaining requirement coverage and shows that this method is more effective than greedy approaches, is more efficient than exact ILP-based approaches and admits multi-objective optimisation. Although the described approaches can take into account code coverage, the methods are not optimised for it and therefore their employment for code coverage optimisation might be unfeasible or ineffective on large programs. To our knowledge no approach uses the CFG properties to optimise the test suite minimisation process.

6 Conclusions

Our proposed approach managed to find the minimum number of test cases that ensure the same code coverage as a large generated test suite. The implemented solution, based on a Set Cover formulation runs in a fraction of a second even for a representation of a software system (JUnit) having more than 5000 of lines, and more than 1700 methods. The reduction on the number of test cases has been up to 96%.

Thanks to our proposed chain-tool, which automates test generation, code coverage computation, and test suite reduction, we are able to analyse a software

system and generate the minimum amount of test cases that have the maximum possible code coverage. The said reduction is greatly beneficial for regression testing, since such tests are often executed several times a day. While the shorter time needed ensures timely feedback, there is also a gain in terms of reduced use of hardware resources.

Acknowledgement. This work is supported by the CLARA project, SCN 00451, funded by MIUR within the “Smart Cities and Communities and Social Innovation” initiative.

References

1. Agitar one - automated JUnit generation. www.agitar.com/solutions/products/automated_junit_generation.html
2. Coffee4j - combinatorial test and fault characterization framework. coffee4j.github.io
3. Beck, K., Gamma, E.: Extreme Programming Explained: Embrace Change. Addison-Wesley Professional, Reading (2000)
4. Calvagna, A., Tramontana, E.: Automated conformance testing of java virtual machines. In: Proceedings of IEEE Conference on Complex, Intelligent, and Software Intensive Systems (CISIS), Taichung, Taiwan (2013)
5. Chen, Z., Zhang, X., Xu, B.: A degraded ILP approach for test suite reduction. In: Proceedings of SEKE (2008)
6. Feige, U.: A threshold of $\ln n$ for approximating set cover. *J. ACM (JACM)* **45**(4), 634–652 (1998)
7. Fornaia, A., Mongiovì, M., Pappalardo, G., Tramontana, E.: A general powerful graph pattern matching system for data analysis. In: International Conference on Complex Networks and their Applications. Springer (2018)
8. Fraser, G., Arcuri, A.: EvoSuite: automatic test suite generation for object-oriented software. In: Proceedings of ACM SIGSOFT Symposium and European Conference on Foundations of Software Engineering (2011)
9. Gotlieb, A., Marijan, D.: Flower: optimal test suite reduction as a network maximum flow. In: Proceedings of ACM International Symposium on Software Testing and Analysis (2014)
10. Hsu, H.Y., Orso, A.: MINTS: a general framework and tool for supporting test-suite minimization. In: Proceedings of IEEE ICSE (2009)
11. Jatana, N., Suri, B., Kumar, P., Wadhwa, B.: Test suite reduction by mutation testing mapped to set cover problem. In: Proceedings of ACM International Conference on Information and Communication Technology for Competitive Strategies (2016)
12. Murphy, C., Zoomkawalla, Z., Narita, K.: Automatic test case generation and test suite reduction for closed-loop controller software. Technical report, University of Pennsylvania, Department of Computer and Information Science (2013)
13. Pacheco, C., Ernst, M.D.: Randoop: feedback-directed random testing for java. In: Proceedings of OOPSLA Companion (2007)
14. Panda, S., Mohapatra, D.P.: Regression test suite minimization using integer linear programming model. *Softw. Pract. Exp.* **47**(11), 1539–1560 (2017)
15. Rothermel, G., Untch, R.H., Chu, C., Harrold, M.J.: Prioritizing test cases for regression testing. *IEEE Trans. Softw. Eng.* **27**(10), 929–948 (2001)

16. Shamshiri, S., Just, R., Rojas, J.M., Fraser, G., McMinn, P., Arcuri, A.: Do automatically generated unit tests find real faults? An empirical study of effectiveness and challenges. In: Proceedings of IEEE/ACM ASE (2015)
17. Shu, G., Sun, B., Henderson, T.A., Podgurski, A.: JavaPDG: a new platform for program dependence analysis. In: Proceedings of IEEE International Conference on Software Testing, Verification and Validation (2013)
18. Tallam, S., Gupta, N.: A concept analysis inspired greedy algorithm for test suite minimization. ACM SIGSOFT Softw. Eng. Notes **31**(1), 35–42 (2006)
19. Xu, S., Miao, H., Gao, H.: Test suite reduction using weighted set covering techniques. In: Proceedings of IEEE International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (2012)

Structural Network Measure



Detection of Conflicts of Interest in Social Networks

Saadia Albane¹✉, Hachem Slimani¹, and Hamamache Kheddouci²

¹ LIMED Laboratory, Faculty of Exact Sciences, University of Bejaia, 06000 Bejaia, Algeria
saadialbane@gmail.com, haslimani@gmail.com

² Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, 69622 Lyon, France
hamamache.kheddouci@univ-lyon1.fr

Abstract. In this paper, we introduce a new approach to detect conflicts of interest (COIs) in social networks. We apply this approach to detect COIs in the review process of papers accepted in an international conference that is represented through a social network. This approach consists of extracting some special chains in the studied social network corresponding to conflict of interest cases where the source and target of each chain correspond to an author and a reviewer, respectively. To evaluate the proposed approach, we have conducted some experiments where a comparison with two methods in the literature has been done. The obtained results have shown some efficiency of the proposed approach.

Keywords: Conflicts of interest · Chains · Social network · Review process · International conference

1 Introduction

Conflict of Interest (COI) is a situation in which personal interests could compromise, or could have the appearance of compromising, the ability of an individual to carry out professional duties objectively [4]. Several factors can cause the COI for example family ties, business or friendship ties, and access to confidential information [1]. The conflict of interest exists in companies, medicine, review of scientific papers, etc. Undesirable effects are caused by COI such as violation of laws, illegal taking of interests, lack of transparency, and loss of confidence, etc. Examples of COIs with their effects can be found in [5, 7, 8]. In the field of social networks, Aleman-Meza et al. [1] integrated two social networks called *Knows* and *Co-author* extracted respectively from a FOAF (Friend-of-a-Friend) social network and the collaborative network in DBLP for determining a degree of conflict of interest. In fact, they used *semantic association* discovery techniques for identification of COI relationships between the reviewer and an author of a paper to be reviewed according to the weights of the implicated individual relationships. Moreover, Aleman-Meza et al. [2] extended their previous work by proposing an approach which measures the strength of collaboration between authors and reviewers, and includes heuristics that use information of friendship/affiliation for the detection of possible COIs. The collaboration strength of two authors adds a weight of $1/(n - 1)$ for each paper they co-authored together where n is the number of authors

in a paper. Furthermore, Khan [6] explored citation relationships as a potential indicator to identify different types of cognitive relationships between researchers. Moreover, Yan et al. [9] proposed an approach based on potential conflicts of interest between authors of submitted papers and reviewers for the assignment of these latter to papers. These potential relations are student-teacher, colleagues, co-authorship, etc.

In this paper, we propose a new approach, based on a type of graphs that we have called *(d-)chains*, to detect COIs in social networks. In particular, we study the problem of determining conflicts of interest during the review process of papers accepted in an international conference that is represented in the form of a social network. The *(d-)chains of COIs* are subgraphs extracted from the dataset/social network where the source (resp. the target) of each one corresponds to an author (resp. a reviewer). Our proposed approach takes into consideration other types of COIs such as the relations of co-editorship and the same affiliation contrary to the method of Aleman-Meza et al. [1]. Moreover, when there exist different COIs between a given couple of reviewer and author then our method takes into consideration only the highest one between them. This latter point is not verified by Aleman-Meza et al. [2]. In our case, a social network is a graph where each vertex can be a people, an organization, or an other social entity and each edge can be collaboration, affiliation, etc. Thus, we propose an approach for extracting, from a given social network, some special chains that correspond to COI cases without using the concept of weight because one collaboration is sufficient to say that it is a high level of COI. Furthermore, for the evaluation of the performance of our approach, we have compared it with the methods of Aleman-Meza et al. [1, 2]. Moreover, for the evaluation of its effectiveness, we have proposed a tool that we have designed and called *dataset generator* which can generate datasets. These latter have been extracted from papers accepted at 7th International Conference on Complex Networks and their Applications (ICCNA'2018) held on December 11–13, 2018 in Cambridge, United Kingdom.

The rest of the paper is structured as follows: we present in Sect. 2 some concepts and definitions concerning graph, arc, and chain that will be used in the sequel. In Sect. 3 we describe and give the different steps of our approach with its evaluation results. Finally, in Sect. 4, we finish by a conclusion and future work.

2 Preliminaries and Definitions

In this section, we present the definitions of graph, arc, and chain. For the need of our study, we introduce and define a new concept called *d-chain* that will be used in the sequel.

- A *graph* G is defined to be a pair (V, E) where $V = v_1, v_2, \dots, v_n$ is a set of elements called vertices, and $E = e_1, e_2, \dots, e_m$ is a set of 2-element subsets of V called edges. A graph can be (not) oriented and (not) labeled.
- A *one-sense arc* in the form of (x, y) , $x, y \in V$ is an oriented one-sense edge from x to y .

- A *chain* of length q in a graph G is a sequence $\mu = (u_1, u_2, \dots, u_q)$ of one-sense arcs such that each arc of the sequence has one extremity in common with the previous arc, and the other extremity in common with the next arc. The number of arcs in the sequence is the length of the chain μ .
- A *double-sense arc* in the form of $[x, y]$, $x, y \in V$ is an oriented double-sense edge, from x to y and from y to x . A double-sense edge $[x, y]$ has to be understood that there exist both the one-sense arcs (x, y) and (y, x) .
- A *d-chain* of length q is a chain that contains at least a double-sense arc, as illustrated in Fig. 1.

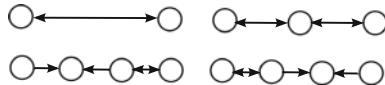


Fig. 1. Main examples of d-chains.

3 Approach of Detection of Conflicts of Interest in Social Networks

The aim of this paper is to define some special chains that correspond to conflicts of interest to be extracted in a social network representing the review process of papers accepted in an international conference.

3.1 Properties and Characteristics of the Considered Social Network

In the literature the properties of social networks depend mainly on the way of their presentations that can be graphs, matrices, etc. In our case, we represent the considered social network by a directed graph $G = (V, E_D, l_V, l_E)$ described as follows:

- V is the vertex set representing the different entities/objects implicated in the review process of the accepted papers in an international conference/journal;
- E_D is the set of one-sense and double-sense arcs;
- l_V is the set of vertex labels where each vertex can be labeled by an author, a reviewer, an accepted paper, an organization, a common collaborator, a collaborator, a conference, etc.;
- l_E is the set of one-sense and double-sense arcs labels where each arc can be labeled with co-authorship, friendship, co-editorship, at, in, etc.

Consequently, the detection of COIs, during the review process of papers accepted in a given conference, between reviewers and authors is to search from the social network if there exist some special chains (*i.e.* (d)-chain) between them that correspond to the associated conflicts of interest.

Since there exists any benchmark for evaluating our approach then the studied social networks are those used by Aleman-Meza et al. [1, 2] in the aim of comparing to their results. To construct their first social network, Aleman-Meza et al. [1] combined entities

(reviewers and authors of the 2004 International World Wide Web conference) and relationships from two independent social networks, called *Knows* and *co-author* extracted respectively from a FOAF social network and the collaborative network in DBLP bibliography for determining a degree of COIs. Moreover, to construct their second social network, Aleman-Meza et al. [2] improved their previous work [1] by including heuristics that use information of friendship/affiliation for the detection of possible COIs between reviewers and authors in various tracks, of the 2006 International World Wide Web conference.

According to Aleman-Meza et al. [1, 2], the existence of a foaf: knows relationship between two persons necessarily implies existence of co-author or co-editor relationships between them. For this, we have only used the DBLP bibliography (*i.e.* SwetoDblp) without the FOAF social network. Moreover, for taking into consideration other relations between reviewers and authors and detecting other COI cases, we have added data to those used by Aleman-Meza et al. [1, 2]. These data are organizations (*i.e.* laboratory, university, company, etc.), the collaborators of reviewers or authors, the organizations of these collaborators, etc. Unfortunately, it is impossible to know the reviewers of each paper of an author in a conference. For this, we suppose that all the papers are reviewed by each reviewer to detect the existence of COIs between the authors of these accepted papers and these reviewers.

3.2 Identifying Levels of Conflicts of Interest Basing on the Concept of (d)-chain

Graphically, the objective is to detect if there exists d-chains or chains, that correspond to conflicts of interest cases, between a given couple of reviewer and author. There exist four cases of COIs in the form of d-chain as given in Fig. 1 and one case of COIs in the form of chain. The forms of these (d)-chains, extracted from a social network, are given in Figs. 2(a), (b), and (c).

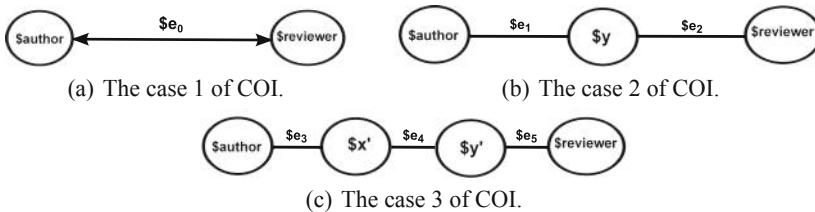


Fig. 2. Cases of COI.

The aim of searching conflicts of interest in a social network is to find if there are (d)-chains of the type given in Fig. 2(a), (b), or (c). Thus, the following levels of conflicts of interest are deduced:

- High level of COI (H) if the d-chain is of length = 1 (See Fig. 2(a)). This level is composed of two cases of d-chains:

1. c_1 : if the double-sense arc $\$e_0$ is labeled *co-authorship/friendship*. Thus, the obtained COI is H_{c_1} .
 2. c_2 : if the double-sense arc $\$e_0$ is labeled *co-editorship/friendship*. Thus, the obtained COI is H_{c_2} .
- Medium level of COI (M) if the (d-)chain is of length = 2 (See Fig. 2(b)). This level is composed of two cases of (d-)chains:
 1. c_3 : if the vertex $\$y$ is labeled *\$common collaborator*, and the double-sense arcs $[\$author, \$common collaborator]$, $[\$reviewer, \$common collaborator]$ are labeled with *co-authorship/friendship* in all the cases. Thus, the obtained COI is M_{c_3} ;
 2. c_4 : if the vertex $\$y$ is labeled *\$organization*, and the one-sense arcs $(\$author, \$organization)$, $(\$reviewer, \$organization)$ are labeled *at/in*. Thus, the obtained COI is M_{c_4} .
 - Low level of COI (L) if the d-chain is of length = 3 (See Fig. 2(c)). This level is constituted of two cases of d-chains:
 1. c_5 : if the vertices $\$x'$ and $\$y'$ are labeled respectively *\$organization* and *\$collaborator*, and the arcs $(\$author, \$organization)$, $(\$collaborator, \$organization)$, $(\$reviewer, \$collaborator)$ are labeled respectively with *at/in*, *at/in*, *co-authorship/friendship*. Thus, the obtained COI is L_{c_5} ;
 2. c_6 : if the vertices $\$x'$ and $\$y'$ are labeled respectively *\$collaborator* and *\$organization*, and the arcs $[\$author, \$collaborator]$, $(\$collaborator, \$organization)$, $(\$reviewer, \$organization)$ are labeled respectively with *co-authorship/friendship*, *at/in*, *at/in*. Thus, the obtained COI is L_{c_6} .

3.3 Experimental Results and Interpretation

For the evaluation of the proposed approach, it is worth mentioning that the field suffers from a clear lack of recent benchmark datasets for evaluation [2, 9]. Moreover, to the best of our knowledge the dataset of Aleman-Meza et al. [3], that is called *SwetoDblp dataset*, is the only publicly available one. For this, we have used this dataset that is available on the site <http://lsdis.cs.uga.edu/projects/semdis/swetodblp/> in RDF format. Furthermore, because of its large size, we have converted it to HDT (Header-Dictionary-Triples) and we have generated a SQLite database which eased our analysis. We have compared our method with two other concurrent approaches. For this, we have only taken two subsets of the SQLite database where the first subset is constituted of 15 reviewers and 5 accepted papers with their authors of the 2004 International World Wide Web conference (WWW2004), that constitutes our first studied *social network* (*i.e.* SN1). The second subset is constituted of 6 reviewers and 5 accepted papers with their authors of the 2006 International World Wide Web conference (WWW2006) in Data mining track, that constitutes our second studied *social network* (*i.e.* SN2). These two social networks (*i.e.* SN1 and SN2) are among those studied respectively by Aleman-Meza et al. [1] and [2] in order to compare to their obtained results. We have also supposed that each accepted paper is reviewed by all the reviewers. Thus, the vertices of the first social network (resp. second social network) have as labels the name of the conference WWW2004 (resp. WWW2006), reviewers, accepted papers with their authors, the organizations, the collaborator until 2004 (resp. 2006) of these reviewers

or authors, and the common collaborators until 2004 (resp. 2006) of these reviewers and authors. Moreover, we have proposed an algorithm which searches (d-)chains of type $c_i; i = 1, \dots, 6$ corresponding to the conflicts of interest (COIs) cases in the social networks. The proposed algorithm is constituted of three steps:

- *Step 1:* Search all the d-chains of type c_1 and c_2 that correspond respectively to the set C_1 and C_2 of high level of conflict of interest;
- *Step 2:* Search all the (d-)chains of type c_3 then c_4 , which correspond respectively to the sets C_3 and C_4 of medium level, that do not exist in C_1 and C_2 ;
- *Step 3:* Search all the d-chains of type c_5 then c_6 , that correspond respectively to the sets C_5 and C_6 of low level, which do not exist in C_1, C_2, C_3 , and C_4 .

Thus, this algorithm considers only the conflict of interest c_i with the smallest index i between a given couple of reviewer and author and ignores the other levels. For the evaluation of the performance of our approach, we have compared it with the methods of Aleman-Meza et al. [1,2]. Moreover, for the evaluation of its effectiveness, we have proposed a tool that we have designed and called *dataset generator* which can generate datasets. These datasets have been extracted from papers accepted at the 7th International Conference on Complex Networks and their Applications (ICCNA'2018) held on December 11–13, 2018 in Cambridge, United Kingdom.

3.3.1 Experimental Results and Comparison 1

Our approach (Algorithm) have been implemented using the Java language and 36 cases of COIs have been obtained where 7 cases are of the high level, 27 cases are of the medium level, and 2 cases are of the low level. Our results of different levels of COIs are obtained in the form of (d-)chains. Some examples of these COIs are presented in Fig. 3.

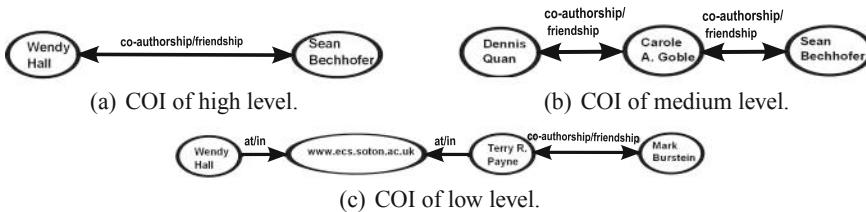


Fig. 3. Example of d-chains obtained from SN1.

Table 1 is established to summarize all the results of comparison of our approach and that of Aleman-Meza et al. [1] and present the different relations between the reviewers and the authors of the accepted papers where:

- H: High level of potential COI caused by previous co-authorship(s)/friendship between a given couple of reviewer and author.
- M: Medium level of potential COI caused by a previous low-to-medium co-authorship(s)/friendship between a given couple of reviewer and author.

- MR: Medium level of potential COI caused by a previous yet rare (*i.e.*, occasional) co-authorship(s)/friendship between a given couple of reviewer and author.
- LR: Low level of potential COI caused by previous yet very rare co-authorship(s)/friendship between a given couple of reviewer and author.
- kM (resp. kL): means that there exist k different COIs of type Medium level (resp. Low level) between a same couple of reviewer and author. To simplify things, if $k = 1$ then we put directly M instead of $1M$. It will be the same for the case of low level and high level.
- D: Definite COI because the reviewer is an author of the paper to be reviewed.
- The results given by Aleman-Meza et al. [1] are underlined, and the ours are not underlined.

In addition to what is studied in Aleman-Meza et al. [1], the proposed approach considers other relations such as same organizations and collaborators for detecting other COI cases without using the concept of weight because one collaboration between a given couple of reviewer and author is sufficient to say that it is a high level of COI. Note that all the COIs detected by Aleman-Meza et al. [1] are also detected by our approach. The only relation that was not detected by our method and that of Aleman-Meza et al. [1] is between the author *Daniel Schwabe* and the reviewer *Stefan Decker* because the used dataset did not have this information. In this part, we compare some COIs of different levels of our approach to those of Aleman-Meza et al. [1].

1. High level of COI: The relation between the author *Alon Y. Halevy* and the reviewer *Ian Horrocks* has a low level (*i.e.* LR) of potential COI because Aleman-Meza et al. [1] did not distinguish between the relation of co-authorship and co-editorship. In our approach, we have considered it as a high level of COI caused by co-editing relationship between them (*i.e.* d-chain of type c_2). Moreover, in Aleman-Meza et al.

Table 1. COI results and comparison 1

Accepted papers	Authors	Reviewers													
		Karl Aberer	Sean Bechhofer	Mark Burstein	Isabel Cruz	Stefan Decker	Addo Gangemi	Ramanathan V. Guha	Jeff Heflin	Ian Horrocks	Jane Hunter	Manolis Koubarakis	John Mylopoulos	Wolfgang Nejdl	Gaus Schreiber
Paper1	Dennis Quan		$3M_{c_3}$		M_{c_3}			M_{c_3}							M_{c_3}
Paper2	Sean Bechhofer	M_{c_3}	<u>D</u>		$3M_{c_3}$				H_c						$4M_{c_3}$
Paper3	Alon Y. Halevy			M_{c_3}	M_{c_3}				LR		$2M_{c_3}$	M_{c_3}	M_{c_3}		
Paper4	Wendy Hall	M_{c_3}	<u>MR</u> , H_{c_1}	$2L_{c_3}$	M_{c_3}			$2M_{c_3}$	M_{c_3}	M_{c_3}				MR , H_{c_1}	
	Leslie Carr	M_{c_3}	H_{c_1}	$2L_{c_3}$	M_{c_3}			$2M_{c_3}$	M_{c_3}					MR , H_{c_1}	
	Timothy Miles-Board		<u>M</u> , H_{c_1}		M_{c_3}			$2M_{c_3}$						$3M_{c_3}$	
	Christopher Bailey		<u>M</u> , $3M_{c_3}$											$4M_{c_3}$	
Paper5	Daniel Schwabe				<u>LR</u>					M_{c_3}					

[1], the relation between the author *Leslie Carr* and the reviewer *Sean Bechhofer* was not found, whereas, in our case, this relation is detected as high level because they have collaborated in 2001 (*i.e.* d-chain of type c_1).

2. Medium level of COI: Twenty six cases of COIs of type medium level are only detected by our approach such as the author *Leslie Carr* and the reviewer *Jane Hunter* because of a common collaborator *Carl Lagoze*. It is the same for the couple (*Leslie Carr, Stefan Decker*) which had as common collaborator *Carole A. Goble*. (*i.e.* d-chains of type c_3).
3. Low level of COI: Two cases of COIs of type low level are only detected by our approach such as the relation between the author *Wendy Hall* and the reviewer *Mark Burstein* such that this reviewer had two collaborators *Paul T. Groth* and *Terry R. Payne* of the same organization as *Wendy Hall*. (*i.e.* d-chain of type c_5).

3.3.2 Experimental Results and Comparison 2

In this part, Table 2 is established to compare our results to those of Aleman-Meza et al. [2]. This table summarizes all the results in Data Mining track and presents the different relations between the reviewers and the authors of the accepted papers. Note that the meaning of not underlined letters is the same to the first comparison. The significations of the underlined letters are defined as follows:

- Hc: High potential COI: due to previous co-authorship(s) between a given couple of reviewer and author.
- Mcc: Medium potential COI: due to common collaborator(s) between a given couple of reviewer and author.
- Ma: Medium potential COI: due to a same-affiliation for a given couple of reviewer and author.
- Me: Medium potential COI: due to previous co-editorship(s) between a given couple of reviewer and author.

The approach of Aleman-Meza et al. [2] is evaluated by analyzing separately the authors of the accepted papers and reviewers of most tracks of the 2006 International World Wide Web conference (WWW2006). In addition to the high and medium levels of COI, our approach takes into consideration the low level of COIs between reviewers and authors for detecting other COI cases from the social network of Aleman-Meza et al. [2]. In this part, we compare some COI cases of our approach to those of Aleman-Meza et al. [2].

1. Medium level of COI: Nine cases of medium level are only detected by our approach such as the author *Masaru Kitsuregawa* and the reviewer *Wei-Ying Ma* because of two common collaborators *Hongjun Lu* and *Xiaofang Zhou*. For this, we put directly $2M$ instead of M (*i.e.* d-chains of type c_3). It is the same for the author *Qiang Yang* and the reviewer *Philip S. Yu* which had three common collaborators *Xindong Wu*, *Jiawei Han*, and *Ke Wang*. For this, we put directly $3M$ instead of M and the level of COI is medium (*i.e.* these cases correspond to chains of type c_3), etc.
2. Low level of COI: Two cases of low level are only detected by our approach such as the author *ChengXiang Zhai* and the reviewer *Wei-Ying Ma* such that this author had

Table 2. COI results in Data Mining Track

WWW2006 Data Mining Track		Reviewers		Soumen Chakrabarti	Thomas Hofmann	Bing Liu	Wei-Ying Ma	Shinichi Morishita	Philip S. Yu
		Authors							
Paper1	Chao Liu								H_c, H_{c_1}
	ChengXiang Zhai			H_c, H_{c_1}	M_{c_3}	$2L_{c_6}$			M_{c_3}
Paper2	Dou Shen					H_c, H_{c_1}			
	Jian-Tao Sun					H_c, H_{c_1}			
	Qiang Yang				H_c, H_{c_1}	H_c, H_{c_1}	M_{c_3}	$3M_{c_3}$	
	Zheng Chen				M_{c_3}	H_c, H_{c_1}			
Paper3	Steven C. H. Hoi					H_c, H_{c_1}			
	Michael R. Lyu					H_c, H_{c_1}			M_{c_3}
Paper4	Masaru Kitsuregawa				M_{c_3}	$2M_{c_3}$	M_a, M_{c_4}	$3M_{c_3}$	
Paper5	Junghoo Cho	H_c, H_{c_1}				L_{c_6}			

two collaborators *Susan T. Dumais* and *Stephen E. Robertson* of the same organization as *Wei-Ying Ma*. For this, we put directly $2L$ instead of L (*i.e.* d-chains of type c_6).

3.3.3 Effectiveness Evaluation

Aleman-Meza et al. [2] (see also Yan et al. [9]) have affirmed that the field of detection of COIs during the review process suffers from a lack of available and recent benchmark datasets to evaluate approaches. Therefore, this fact raises a huge difficulty regarding the validation of approaches related to this latter field. Thus, to deal with this problem we have proposed a tool that we have called *dataset generator* which has several characteristics: first, it allows to avoid conflicts with authors or reviewers which are included in the dataset of the study. Second, it can generate different synthetic datasets. Third, it ensures dynamic parameter setting in order to generate datasets with characteristics as needed. In our case, for the construction of a social network (SN), we have taken all the authors (140) of the accepted papers at the 7th International Conference on Complex Networks and their Applications (ICCNA'2018) held on December 11–13, 2018 in Cambridge, United Kingdom, and selected 60 reviewers, 61 organizations from the affiliations of the considered authors and reviewers, and 50 collaborators selected randomly from the lists of the same authors and reviewers. Furthermore, we have considered that 30 fictitious papers are written by the 140 authors. Thus, using the previous data, we have constructed the social network SN, corresponding to a generated dataset by the dataset generator, as follows:

1. A node of SN can represent a fictif paper, its author, one of the 60 considered reviewers, one of the 50 collaborators, or one of the 61 organizations. Note that for each fictif paper its corresponding authors are affected randomly by the dataset generator as a subset from the whole set of 140 authors.
2. The dataset generator constructs the arcs representing the relations between the nodes of SN as follows:
 - For each fictif paper a number of 2 to 4 reviewers are affected randomly by the dataset generator,
 - For each author or reviewer a number of 1 to 6 collaborators are affected randomly by the dataset generator,
 - For each author, reviewer, or collaborator at most one organization is affected by the dataset generator.

We have implemented the dataset generator using Java language. To do our experimentation, we have generated 3 datasets. At each generation of each one, its ground truth that we have manually computed, is used as a benchmark, for evaluating the results of the proposed approach compared with the approach of Aleman-Meza et al. [2] after their application to the dataset. This process of computing is a challenge because for each studied dataset we have obtained a great number of COIs.

To evaluate the effectiveness of the proposed approach in term of the percentage of detected COIs, we have constructed three datasets by the dataset generator. This latter have taken into consideration a set of 30 fictif papers. Furthermore, up to 4 reviewers are assigned to review each of the papers. For the experimentations, we have applied the two approaches to each generated dataset and we have computed the percentage of the correct COIs found with each approach. Subsequently, we have also computed the average result for all the datasets. Figure 4 presents the overall results of the experimentation.

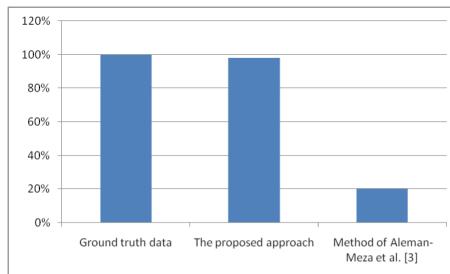


Fig. 4. The average result of the approaches.

From Fig. 4, it is clear that our average percentage result (which is about 98%) is practically similar to the ground truth result (*i.e.* 100%) and better than the result of the method of Aleman-Meza et al. [2] (which is around 20%). This difference is caused by two reasons: first, Aleman-Meza et al. [2] measure the collaboration strength (*i.e.* weight) by taking into consideration the number of authors in a paper and the

number of papers two people co-authored. However, our approach does not take into consideration this concept of weight. Second, contrary to our method, Aleman-Meza et al. [2] don't take into account the relations of low level of COIs. On the other hand, the small difference (*i.e.* 2%) between our method and the ground truth is caused by the fact that our approach does not take into consideration all the cases of COIs between a given couple of reviewer and author but only the highest one.

4 Conclusion

We have introduced an approach based on (d-)chains to detect COIs during the review process of papers accepted in an international conference that is represented in the form of a social network. The proposed approach takes into consideration the relations of co-editorship and affiliation and classifies them respectively as high level and medium level contrary to Aleman-Meza et al. [1]. On the other hand, in Aleman-Meza et al. [2], the relation of co-editorship is considered as medium level instead of high level, and the relation of low level between a given couple of reviewer and author is not considered. Moreover, contrary to Aleman-Meza et al. [2], our algorithm does not take into consideration all the levels of COIs between a given couple of reviewer and author but only the highest one. However, the results of our approach do not display all the information about the accepted papers, and the name of the conference. It is worth to note that for doing our experimentations and evaluation, the obtaining of the SwetoDblp dataset was hard, and the construction of the two social networks was a challenge because of the RDF format, and the difficulties of confirming and verifying the identity of each author and reviewer, and recovering for each one his/her affiliation and his/her collaborators to a specific date. Furthermore, we have implemented a dataset generator to generate several synthetic datasets. Moreover, we have evaluated the effectiveness of our approach on these datasets and we have compared it with the established ground truth and the method of Aleman-Meza et al. [2]. The obtained results have shown some efficiency of the proposed approach. As future work, it would be interesting to extend this approach by displaying more details about the conflicts of interest.

References

1. Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Sheth, A.P., Arpinar, I.B., Ding, L., Kolari, P., Joshi, A., Finin, T.: Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection. In: Proceedings of the 15th International Conference on World Wide Web, pp. 407–416. ACM (2006)
2. Aleman-Meza, B., Nagarajan, M., Ding, L., Sheth, A.P., Arpinar, I.B., Joshi, A., Finin, T.: Scalable semantic analytics on social networks for addressing the problem of conflict of interest detection. ACM Trans. Web (TWEB) **2**, 7 (2008)
3. Aleman-Meza, B., Hakimpour, F., Arpinar, I.B., Sheth, A.P.: SwetoDblp ontology of computer science publications. J. Web Semant. **5**(6), 151–155 (2007)
4. Biaggioni, I.: Conflict of interest guidelines: an argument for disclosure. Pharm. Ther. **322**, 324 (1993)
5. Frachon, I.: Mediator 150 mg: Sous-titre censuré, postface de Rony Brauman, éditions-dialogues.fr (2010)

6. Khan, M.S.: Exploring citations for conflict of interest detection in peerreview system. *Int. J. Comput. Inf. Syst. Ind. Manag. Appl.* **3** (2012)
7. Klitzman, R., Chin, L.J., Rifai-Bishjawish, H., Kleinert, K., Leu, C.S.: Disclosures of funding sources and conflicts of interest in published HIV/AIDS research conducted in developing countries. *J. Med. Ethics* **36**, 505–510 (2010)
8. Ross, J.S., Madigan, D., Hill, K.P., Egilman, D.S., Wang, Y., Krumholz, H.M.: Pooled analysis of rofecoxib placebo-controlled clinical trial data: lessons for postmarket pharmaceutical safety surveillance. *Arch. Intern. Med.* **169**, 1976–1985 (2009)
9. Yan, S., Jin, J., Geng, Q., Zhao, Y., Huang, X.: Utilizing academic-network-based conflict of interests for paper reviewer assignment. *Int. J. Knowl. Eng.* **3**, 65–73 (2017)



Comparing Spectra of Graph Shift Operator Matrices

Johannes F. Lutzeyer^(✉) and Andrew T. Walden

Imperial College London, London SW7 2AZ, UK
jl7511@ic.ac.uk, a.walden@imperial.ac.uk

Abstract. Typically network structures are represented by one of three different graph shift operator matrices: the adjacency matrix and unnormalised and normalised Laplacian matrices. To enable a sensible comparison of their spectral (eigenvalue) properties, an affine transform is first applied to one of them, which preserves eigengaps. Bounds, which depend on the minimum and maximum degree of the network, are given on the resulting eigenvalue differences. The monotonicity of the bounds and the structure of networks are related. Bounds, which again depend on the minimum and maximum degree of the network, are also given for normalised eigengap differences, used in spectral clustering. Results are illustrated on the karate dataset and a stochastic block model. If the degree extreme difference is large, different choices of graph shift operator matrix may give rise to disparate inference drawn from network analysis; contrariwise, smaller degree extreme difference results in consistent inference.

Keywords: Graph shift operator matrix · Spectrum · Clustering

1 Introduction

Networks can be represented in multiple ways via different graph shift operator matrices (GSOMs), typically by the adjacency matrix or normalised or unnormalised Laplacians. The degree d_i of the i^{th} vertex is defined to be the sum of the weights of all edges connecting this vertex with others and the degree matrix D is defined to be the diagonal matrix $D = \text{diag}(d_1, \dots, d_n)$. With the adjacency matrix denoted by A , the unnormalised graph Laplacian, L , is defined as $L = D - A$, and there are two commonly used normalised Laplacians defined as $L_{rw} = D^{-1}L$ and $L_{sym} = D^{-1/2}LD^{-1/2}$. The aim of this work is to compare the spectra (eigenvalues) of the GSOMs, A and L , and since the two normalised Laplacians are related by a similarity transform, so have identical eigenvalues, we choose to work with just one of these, namely L_{rw} .

The spectral properties of the GSOMs are utilised in many disciplines [4] for a broad range of different analysis methods, such as, the spectral clustering algorithm [16], graph wavelets [14], change point detection in dynamic networks [1] and the quantification of network similarity [7]. In particular, spectral clustering

has had a big impact in the recent analysis of networks, with active research on the theoretical guarantees which can be given for a clustering determined in this way [11, 13]. The spectral properties of the different GSOMs are also starting to be leveraged in graph learning, as for example in [10]. It is the recommendation to use whichever GSOM works best in a particular analysis or learning task [12]. A framework under which the GSOM spectra can be compared and an improved understanding of the magnitude of the spectral differences of the GSOMs is of real use in informing such a decision.

In the remainder of the introduction we motivate the main ideas of the paper, which are the use of affine transformation to enable the comparison of the GSOM spectra, the affine transformation parameter choice and the bounds on the individual eigenvalue difference of the GSOMs. In Sect. 2 we provide bounds on the eigenvalue differences of the GSOMs for general graphs and visualise the effects of the affine transformation on the GSOM spectra corresponding to Zachary’s karate social network. Then, in Sect. 3 we observe and explain the monotonicity of the eigenvalue bounds. In Sect. 4 we provide bounds on the difference of the normalised eigengaps (differences of successive eigenvalues) of the GSOMs and visualise them using a sample from a stochastic blockmodel. Section 5 highlights the impact of the GSOM choice on the inference drawn in graphical analysis by applying the spectral clustering algorithm using all three GSOMs to the karate network. In Sect. 6 we provide a summary and conclusions.

1.1 Motivation of the Affine Transformations

A direct comparison of the observed spectra is difficult. Primarily this is because for two of three comparisons to be made, the ordering of the eigenvalues is reversed; there is a rough correspondence of the *larger* end of the adjacency matrix spectrum to the the *smaller* end of the Laplacian spectra and vice versa for the other ends. Also, the supports of the three GSOM spectra are different [15, pp. 29, 64, 68]. (The upcoming Fig. 1 illustrates this.) For an insightful comparison it makes sense to relocate and scale the spectra (eigenvalues), $\mu_n \leq \dots \leq \mu_1$, say, via an affine transformation $g(\mu) = c\mu + b$, where $c, b \in \mathbb{R}, c \neq 0$. Ordering is preserved for $c > 0$, i.e., $g(\mu_1) \geq g(\mu_2)$ and reversed for $c < 0$, i.e., $g(\mu_1) \leq g(\mu_2)$. Eigengaps, relative to the spectral support, are preserved. Assume the domain of g is equal to the interval $[x_1, x_2]$, so g ’s image is equal to $[g(x_1), g(x_2)]$. Then normalised eigengaps are preserved:

$$\frac{\mu_1 - \mu_2}{x_2 - x_1} = \frac{c(\mu_1 - \mu_2) + b - b}{c(x_2 - x_1) + b - b} = \frac{g(\mu_1) - g(\mu_2)}{g(x_2) - g(x_1)}. \quad (1)$$

In the spectral clustering algorithms, eigengaps are used to determine the number of clusters in the network, while the eigenvalue ordering is used to identify which eigenvectors to consider [16]. Therefore, when comparing the impact of the choice of GSOMs on graphical analysis, we need to preserve or reverse eigenvalue ordering and preserve relative eigengap size.

1.2 Motivation of the Parameter Choice

All three GSOMs are related through the degree matrix D . Denote the degree extremes by d_{\min} and d_{\max} . For d -regular graphs, where $d_{\max} = d_{\min} = d$, the GSOMs are already exactly related by affine transformations and hence their eigenvalues are related by the same affine transformations. For general (non- d -regular graphs) we shall make parameter choices for the affine transformations such that eigenvalue differences, after affine transformation, can be bounded above in terms of the elements of D . These general parameter choices agree with the natural existing transformation parameters for d -regular graphs.

1.3 Motivation for Calculating a Bound on the Eigenvalue Differences

The motivation for bounding the eigenvalue difference is twofold. On the one hand, the structure of the bound and its dependency on network parameters allow us to infer for which graphs we are able to see larger differences in the GSOM spectra. On the other hand, the bounds allow us to add a sensible scale to observed eigenvalues; the bound represents the theoretically maximal distance the eigenvalues and eigengaps can differ by and therefore we are able to compare observed differences to the maximal possible differences.

2 GSOM Eigenvalue Differences

2.1 Transforms and Bounds

Theorem 1. *Consider a graph G with degree extremes d_{\min} and d_{\max} . Let the eigenvalues of the corresponding adjacency matrix A and unnormalised Laplacian matrix L be denoted by $\mu_n \leq \dots \leq \mu_1$ and $\lambda_1 \leq \dots \leq \lambda_n$, respectively. Then, there exists a transformation $f_1(\mu_i) = c_1\mu_i + b_1$, where $c_1 = -1$ and $b_1 = (d_{\max} + d_{\min})/2$, such that for all $i \in \{1, \dots, n\}$,*

$$|f_1(\mu_i) - \lambda_i| \leq \frac{d_{\max} - d_{\min}}{2} \stackrel{\text{def}}{=} e(A, L). \quad (2)$$

For d -regular graphs, (2) gives $e(A, L) = 0$, so that $f_1(\mu_i) = \lambda_i = c_1\mu_i + b_1 = d - \mu_i$, i.e., the eigenvalues are related by the required exact relation, as claimed in Sect. 1.2.

For general graphs, using (2), we can establish a rough correspondence, to within an affine transformation, between the eigenvalues of the adjacency matrix, A , and the unnormalised graph Laplacian, L , if the extremes of the degree sequence d_{\max} and d_{\min} are reasonably close.

Theorem 2. *Consider a graph G with degree extremes $d_{\min} > 0$ and d_{\max} . Let the eigenvalues of L and L_{rw} be denoted by $\lambda_1 \leq \dots \leq \lambda_n$ and $\eta_1 \leq \dots \leq \eta_n$, respectively. Then, there exists a transformation $f_2(\lambda_i) = c_2\lambda_i + b_2$, where $c_2 = 2/(d_{\max} + d_{\min})$ and $b_2 = 0$, such that for all $i \in \{1, \dots, n\}$,*

$$|f_2(\lambda_i) - \eta_i| \leq 2 \frac{d_{\max} - d_{\min}}{d_{\max} + d_{\min}} \stackrel{\text{def}}{=} e(L, L_{rw}). \quad (3)$$

Asymptotically, as $d_{\max} \rightarrow \infty$ with fixed $d_{\min} > 0$, the bound $e(L, L_{rw})$ tends to 2. The restricted range of d_{\min} is due to D^{-1} , the normalised Laplacian and consequently the bound $e(L, L_{rw})$ not being defined for $d_{\min} = 0$. For d -regular graphs, where $d_{\max} = d_{\min} = d$, the bound equals zero. Hence, the spectra of $f_2(L)$ and L_{rw} are equal and $\eta_i = c_2 \lambda_i = \lambda_i/d$, i.e., the eigenvalues are related by the required exact relation. Overall, the behaviour of this bound is similar to $e(A, L)$. The smaller the difference of the degree sequence extremes d_{\max} and d_{\min} , the closer the spectra of L and L_{rw} are related, signified by a smaller bound on the eigenvalue differences.

Theorem 3. *Consider a graph G with degree extremes $d_{\min} > 0$ and d_{\max} . Let the eigenvalues of A and L_{rw} be denoted by $\mu_n \leq \dots \leq \mu_1$ and $\eta_1 \leq \dots \leq \eta_n$, respectively. Then, there exists a transformation $f_3(\mu_i) = c_3 \mu_i + b_3$, where $c_3 = -2/(d_{\max} + d_{\min})$ and $b_3 = 1$, such that for all $i \in \{1, \dots, n\}$,*

$$|f_3(\mu_i) - \eta_i| \leq \frac{d_{\max} - d_{\min}}{d_{\max} + d_{\min}} \stackrel{\text{def}}{=} e(A, L_{rw}). \quad (4)$$

For d -regular graphs, $e(A, L_{rw})$ equals zero and hence the spectra of $f_3(A)$ and L_{rw} are equal and $\eta_i = 1 + c_3 \mu_i = 1 - (\mu_i/d)$, i.e., the spectra of L_{rw} and A are related by the required exact relation. For general graphs, $e(A, L_{rw})$ is small for small degree extreme differences and hence the spectra of the two GSOMs exhibit smaller maximal differences, as was the case for $e(A, L)$ and $e(L, L_{rw})$.

The proofs of Theorems 1, 2 and 3 directly follow from inequality (4.3) in [2, p. 46], choosing parameters c_1, b_2 and b_3 as given in the theorem statements to make the problem analytically solvable and by then analytically minimising the two norm. More detailed proofs of the theorems are available from the authors upon request.

2.2 Karate Dataset Example

The karate dataset [3] dates back to [18]. We work with a square matrix with 34 entries, ‘ZACHE,’ which is the adjacency matrix of a social network describing presence of interaction in a university karate club. This data set is popular as it naturally lends itself to clustering [5,8,9].

The untransformed and transformed eigenvalues of the GSOMs A , L and L_{rw} of Zachary’s karate dataset are shown in Fig. 1. From Fig. 1(a), (b) and (c) not much can be said about how the spectra compare. From Fig. 1(d), (e) and (f) it can be observed that each pair of spectra of the karate dataset cover a similar range after transformation and that clearly the spectra of A and L_{rw} are the most similar of the three. It can also be seen that the largest eigengaps occur at opposing ends of the spectra when comparing A and L_{rw} with L . This observation is only possible from the plots including the transforms, i.e., Fig. 1(d), (e) and (f). Clearly, the eigenvalue spectra become more comparable through utilising the proposed transformations.

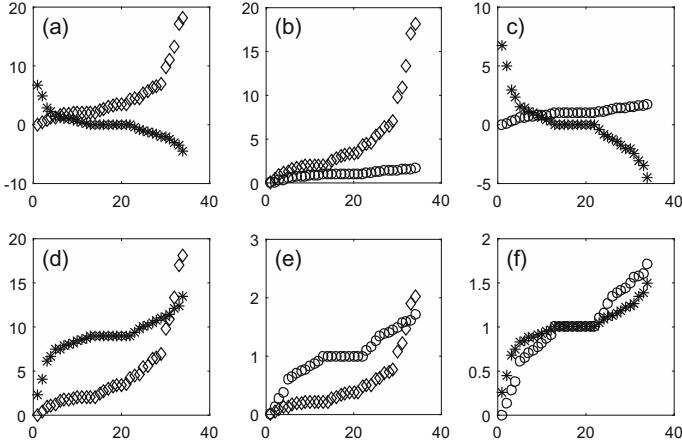


Fig. 1. Transformation results for Zachary's karate dataset. First row: (a) eigenvalues μ of A (stars) and λ of L (diamonds); (b) eigenvalues λ of L and η of L_{rw} (circles); (c) eigenvalues μ of A and η of L_{rw} . Second row: the first of the two eigenvalue spectra are transformed by their respective transformations f_1 , f_2 and f_3 .

Remark 1. [19, p. 32] shows that vertices with identical neighbourhoods generate eigenvalues equal to 0 and 1 in the adjacency and normalised Laplacian spectrum, respectively. Interestingly, we spot several such vertices in the karate network, which will be shown in Fig. 4, and several corresponding eigenvalues equal to 0 and 1 in Fig. 1(c). It is very nice to see that the equivalence of these eigenvalues is highlighted by our transformation mapping them exactly onto each other in Fig. 1(f). \triangleleft

In Fig. 2, we display a proof of concept of our bounds. The eigenvalue bounds are centred around the average value of each eigenvalue pair in order to highlight the maximal difference achievable by each individual eigenvalue pair under comparison. For the karate dataset the bound values $(e(A, L), e(L, L_{rw}), e(A, L_{rw}))$ are equal to $(8.00, 1.78, 0.89)$. The particular bounds displayed here are valid for all graphs with $d_{\min} = 1$ and $d_{\max} = 17$, i.e., all graphs in $\mathcal{C}_{1,17}$, where $\mathcal{C}_{j,k}$ is the class of graphs with $d_{\min} = j$ and $d_{\max} = k$. The bounds being almost attained in plot (a), and not attained in plots (b) and (c), is more a consequence of the structure of the graph given by the karate data set than tightness and quality of the bounds. Since the three bounds $e(A, L)$, $e(L, L_{rw})$ and $e(A, L_{rw})$ apply to entire classes $\mathcal{C}_{j,k}$ at a time, we can only achieve tightness on these classes—bounds being attained for some elements in $\mathcal{C}_{j,k}$ —and not on each individual element of them.

So for our particular social network, the karate dataset, firstly the bound being almost attained in plot (a) tells us that the spectra of A and L deviate almost as much as theoretically possible for a graph in $\mathcal{C}_{1,17}$. Secondly, from plot (c) we see that the spectra of A and L_{rw} are rather similar for the karate data set and could theoretically deviate significantly more for a different graph in

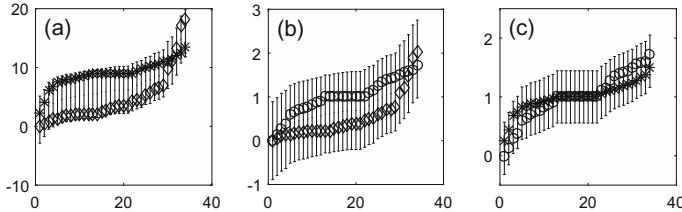


Fig. 2. Eigenvalue bounds on the karate eigenvalues. In plot (a) we display the bound $e(A, L)$ via the intervals together with the transformed eigenvalues of the adjacency matrix $f_1(\mu)$ (stars) and the eigenvalues of the Laplacian λ (diamonds). (b) the bound $e(L, L_{rw})$ is displayed via the intervals, the diamonds correspond to the transformed Laplacian eigenvalues $f_2(\lambda)$ and the circles are the eigenvalues of the normalised graph Laplacian η . (c) the bound $e(A, L_{rw})$ is displayed via the intervals, the stars correspond to the transformed adjacency eigenvalues $f_3(\mu)$ and the circles are the normalised Laplacian eigenvalues η .

$C_{1,17}$ having degree extreme difference 16. The bounds give us insight into how particular GSOM spectra behave for the karate dataset.

Table 1. Comparing bounds on eigenvalue differences of A , L and L_{rw} . The bound values are displayed as $(e(A, L), e(L, L_{rw}), e(A, L_{rw}))$.

d_{\min}						
		0	1	2	3	4
d_{\max}	1	$(0.5, \cdot, \cdot)$	$(0, 0, 0)$	*	*	*
	2	$(1, \cdot, \cdot)$	$(0.5, 0.67, 0.33)$	$(0, 0, 0)$	*	*
	3	$(1.5, \cdot, \cdot)$	$(1, 1, 0.5)$	$(0.5, 0.4, 0.2)$	$(0, 0, 0)$	*
	4	$(2, \cdot, \cdot)$	$(1.5, 1.2, 0.6)$	$(1, 0.67, 0.33)$	$(0.5, 0.29, 0.14)$	$(0, 0, 0)$
	5	$(2.5, \cdot, \cdot)$	$(2, 1.33, 0.67)$	$(1.5, 0.86, 0.43)$	$(1, 0.5, 0.25)$	$(0.5, 0.22, 0.11)$
	6	$(3, \cdot, \cdot)$	$(2.5, 1.43, 0.71)$	$(2, 1, 0.5)$	$(1.5, 0.67, 0.33)$	$(1, 0.4, 0.2)$

3 Relating the Spectral Bounds

We display some sample bound values in Table 1. The first column of Table 1 only contains values of $e(A, L)$ as for $d_{\min} = 0$ the normalised Laplacian and hence $e(L, L_{rw})$ and $e(A, L_{rw})$ are not well-defined. In practice this is of little consequence since disconnected nodes are commonly removed from the dataset as a preprocessing step. On the diagonal of Table 1, where $d_{\max} = d_{\min}$, i.e., for the d -regular graphs, we find $e(A, L) = e(L, L_{rw}) = e(A, L_{rw}) = 0$ due to the direct spectral relation of the GSOMs.

Remark 2. It is interesting to note, that since the spectral support of neither A , nor L , is bounded, the bound on their eigenvalue difference is also not bounded above. (This explains the large values in Fig. 2(a)). This does not apply to the other two bounds as we have $e(L, L_{rw}) \leq 2$ and $e(A, L_{rw}) \leq 1$. \triangleleft

Before discussing the structure of Table 1, we define connected components in a graph.

Definition 1. A path on a graph G is an ordered list of unique vertices such that consecutive vertices are connected by an edge. A vertex set S_k is called a connected component if there exists a path between any two vertices $v_i, v_j \in S_k$ and there exists no path from any $v_i \in S_k$ to any $v_j \notin S_k$. \triangleleft

Since all graphs in $\mathcal{C}_{j+1,k}$ can be extended to lie in $\mathcal{C}_{j,k}$ by adding one or more connected components, all spectra of graphs in $\mathcal{C}_{j+1,k}$ are subsets of spectra of graphs in $\mathcal{C}_{j,k}$ (the GSOM spectrum of a graph is the union of the spectra of its connected components [6, p. 7]). Therefore, the support of the spectra of graphs in $\mathcal{C}_{j,k}$ must be larger or equal to the support of spectra of graphs in $\mathcal{C}_{j+1,k}$. Hence, we expect the spectral bounds, $e(\cdot, \cdot)$, we derived to be decreasing or constant with increasing d_{\min} and constant d_{\max} . This phenomenon can be observed in Table 1 when traversing each row.

In similar fashion, any graph in $\mathcal{C}_{j,k}$, can be extended to be a graph in $\mathcal{C}_{j,k+1}$, by adding a connected component with all vertex degrees greater or equal to j and smaller or equal than $k+1$ with at least one node attaining degree $k+1$. Then spectra of graphs in $\mathcal{C}_{j,k}$ are subsets of spectra of graphs $\mathcal{C}_{j,k+1}$ and therefore the spectral support and hence the spectral bounds on $\mathcal{C}_{j,k+1}$ have to be greater or equal than the respective quantities for $\mathcal{C}_{j,k}$. So, any of the spectral bounds, $e(\cdot, \cdot)$, will be increasing or constant with increasing d_{\max} and constant d_{\min} , as seen in the columns of Table 1.

4 GSOM Normalised Eigengap Differences

Here we derive bounds on the normalised eigengap differences, where each eigengap is normalised by the spectral support of its corresponding GSOM.

Let \mathcal{M}_i denote the i^{th} eigengap of A , $\mathcal{M}_i = \mu_i - \mu_{i+1}$, \mathcal{L}_i denote the i^{th} eigengap of L , $\mathcal{L}_i = \lambda_{i+1} - \lambda_i$ and \mathcal{N}_i denote the i^{th} eigengap of L_{rw} , $\mathcal{N}_i = \eta_{i+1} - \eta_i$, for $i \in \{1, 2, \dots, n-1\}$. The spectral supports of A , L and L_{rw} are equal to $[-d_{\max}, d_{\max}]$, $[0, 2d_{\max}]$ and $[0, 2]$, respectively [15, pp. 29, 64, 68], so the lengths of the supports are $\ell(\mu) = \ell(\lambda) = 2d_{\max}$ and $\ell(\eta) = 2$.

From (1) we see that the normalisation of transformed eigengaps by the transformed spectral support is equal to the normalisation of the untransformed eigengaps by the untransformed spectral support. We will therefore start our analysis by considering the untransformed normalised eigengaps. Eigengap normalisation by the spectral support of the corresponding GSOM is crucial to be able to make a meaningful comparison of eigengap magnitudes.

Theorem 4. *Bounds on the normalised eigengap difference of the GSOMs are given by*

$$\begin{aligned} \left| \frac{\mathcal{M}_i}{2d_{\max}} - \frac{\mathcal{L}_i}{2d_{\max}} \right| &\leq \frac{d_{\max} - d_{\min}}{2d_{\max}} \stackrel{\text{def}}{=} g(A, L); \\ \left| \frac{\mathcal{L}_i}{2d_{\max}} - \frac{\mathcal{N}_i}{2} \right| &\leq 2 \frac{d_{\max} - d_{\min}}{d_{\max}} \stackrel{\text{def}}{=} g(L, L_{rw}); \\ \left| \frac{\mathcal{M}_i}{2d_{\max}} - \frac{\mathcal{N}_i}{2} \right| &\leq \frac{d_{\max} - d_{\min}}{d_{\max}} \stackrel{\text{def}}{=} g(A, L_{rw}). \end{aligned}$$

The proofs of the three inequalities in Theorem 4 directly follow by applying the triangle inequality and then varying the proofs of Theorems 1, 2 and 3 slightly to accommodate the normalisation of the spectra as transformation parameters. More detailed proofs of the theorems are available from the authors upon request.

4.1 Stochastic Blockmodel Example

We see that the eigengaps of the GSOMs are different, giving further evidence that the graphical structure captured in the different spectra differs significantly and possibly in a structured manner. Furthermore, we have been able to use our bounds to put the magnitude of the observed eigengaps into the broader perspective of all graphs with corresponding degree extremes.

The stochastic blockmodel, which is widely used in the networks literature [9, 11], allows us to encode a block structure in a random graph via different probabilities of edges within and between node-blocks. Our definition and parametrisation is adapted from [11].

Definition 2. *Consider a graph with node set $\{v_1, \dots, v_n\}$. Split this node set into K disjoint blocks denoted $\mathcal{B}_1, \dots, \mathcal{B}_K$. We encode block membership of the nodes via a membership matrix $M \in \{0, 1\}^{n \times K}$, where $M_{i,j} = 1$ if $v_i \in \mathcal{B}_j$ and $M_{i,j} = 0$ otherwise. Finally, we fix the probability of edges between blocks to be constant and collect these probabilities in a probability matrix $P \in [0, 1]^{K \times K}$, i.e., for nodes $v_i \in \mathcal{B}_l$ and $v_j \in \mathcal{B}_m$ the probability of an edge between v_i and v_j is equal to $P_{l,m}$.* \triangleleft

Hence, the parameters of the stochastic blockmodel are $M \in \{0, 1\}^{n \times K}$ and $P \in [0, 1]^{K \times K}$, where the number of nodes $n \in \mathbb{N}$ and the number of clusters $K \in \mathbb{N}$ are implicitly defined via the dimensions of M . We simulate graphs from this model by fixing these parameters and then sampling edges from Bernoulli trials. The Bernoulli parameter of the trial corresponding to the edge connecting v_i to v_j is given by the $(i, j)^{\text{th}}$ -entry of the matrix MPM^T .

Figure 3 contains eigengaps corresponding to the three GSOMs of a graph arising as a realisation of a stochastic blockmodel with parameters $n = 600$; $K = 6$; $M_{1,1} = \dots = M_{100,1} = M_{101,2} = \dots = M_{200,2} = M_{201,3} = \dots = M_{300,3} = M_{301,4} = \dots = M_{400,4} = M_{501,6} = \dots = M_{600,6} = 1$ and all other

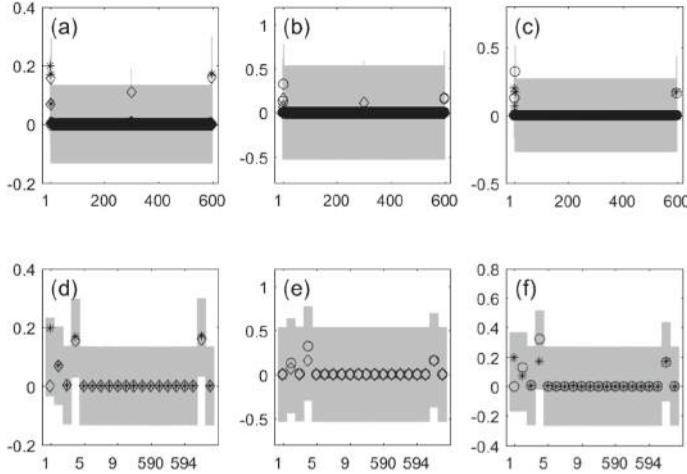


Fig. 3. Plots (a), (b) and (c) display all normalised eigengaps and plots (d), (e) and (f) display the first and last 10 normalised eigengaps of a graph sampled from a stochastic blockmodel. The stars correspond to eigengaps of A normalised by the spectral support of A , the diamonds display the corresponding quantity for L and the circles represent the corresponding quantity for L_{rw} . The filled light grey area delineates the three bounds, where $g(A, L)$ is displayed in plots (a) and (d), $g(L, L_{rw})$ is shown in plots (b) and (e) and $g(A, L_{rw})$ is displayed in plots (c) and (f).

entries of M equal 0;

$$P = \begin{pmatrix} 0.9 & 0.1 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0.9 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0.1 & 0.9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0.9 & 0.9 \\ 0 & 0 & 0 & 0.9 & 0.1 & 0.9 \\ 0 & 0 & 0 & 0.9 & 0.9 & 0.1 \end{pmatrix}.$$

Hence, we have a network of 6 blocks with 100 nodes per block. From observation of P , we recognise that all realisations from this model consist of at least *two connected components* since the probability of an edge between blocks $\{\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3\}$ and $\{\mathcal{B}_4, \mathcal{B}_5, \mathcal{B}_6\}$ is equal to 0.

In order to gain more insight into the block structure of this stochastic blockmodel we consider different definitions of clusters or blocks. [13] describe so called ‘*heterophilic*’ clusters, which they recover from the eigenvectors corresponding to the extreme negative eigenvalues of the normalised Laplacian. These heterophilic clusters have fewer connections within than between clusters and are exemplified by romantic relationship graphs where the genders are the two biggest clusters identifiable. Our blockmodel includes such heterophilic blocks in $\{\mathcal{B}_4, \mathcal{B}_5, \mathcal{B}_6\}$. While blocks $\{\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3\}$ form clusters in the more traditional sense with more edges within rather than between blocks, which will be referred to as ‘*homophilic*’ cluster structure for the remainder of this analysis.

In the analysis of Fig. 3, we interpret large eigengaps in the context of spectral clustering. Firstly, large eigengaps are understood to indicate the number of clusters discovered by the different GSOMs, i.e., a large K^{th} eigengap is under-

stood to indicate a clustering into K communities [16, p. 410]. Secondly, the size of the eigengap is understood to represent relative strength of evidence for a certain clustering [16, p. 407].

In Fig. 3 we find that all three spectra have a non-zero normalised eigengap with indices 2, 4 and 598. In addition, A has a large first normalised eigengap and L a large 301st normalised eigengap; with the exception of these two large eigengaps, in the context of the spectral clustering algorithms the eigengaps suggest searching for similar numbers of clusters.

From [16, pp. 397–8] we know that the nonzero second eigengap of the Laplacians in Fig. 3 encodes the separation of the network into two connected components. It is interesting to see that all GSOM spectra contain stronger evidence for the block structure within the two connected components of the stochastic blockmodel, encoded by the large 4th and 598th normalised eigengaps, than the clustering into the two connected components themselves, encoded by the large 2nd normalised eigengaps. It is especially interesting that for A and L the 4th and 598th normalised eigengaps are approximately equal, while for L_{rw} the 4th normalised eigengap is almost twice as large as the 598th normalised eigengap. Both components have symmetric parameters and therefore it is very interesting that the eigengaps in L_{rw} suggest stronger evidence for the homophilic cluster structure encoded by the 4th eigengap rather than the heterophilic cluster structure encoded by the 598th eigengap. A much larger simulation of stochastic blockmodels with varying parameters, which is beyond the scope of this work, could establish whether L_{rw} does indeed have a structured preference for homophilic over heterophilic blocks.

5 Application to Spectral Clustering

To illustrate the impact of the GSOM choice, we display the spectral clustering according to the first two eigenvectors of each of the three GSOMs corresponding to the karate data set. The choice of GSOM has a significant impact on the clustering outcome. We use the simplest form of spectral clustering by running the k -means algorithm on the rows of the first k eigenvectors of the different GSOMs [16, p. 399].

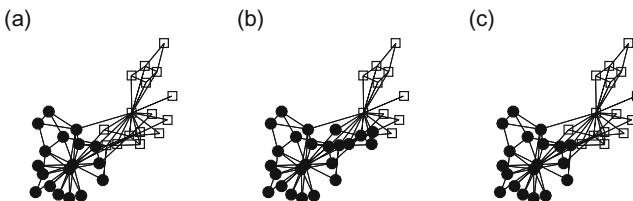


Fig. 4. Spectral clustering of the karate network according to the first two eigenvectors of the adjacency matrix A in (a), the unnormalised Laplacian L in (b) and the normalised Laplacian L_{rw} in (c).

Since for the karate dataset, we have a reasonably large degree extreme difference, $d_{\max} - d_{\min} = 16$, we expect to see deviating results in the spectral clustering according to the different GSOMs.

At first sight, all clusterings displayed in Fig. 4 seem sensible. The clustering according to the adjacency matrix extends the cluster marked with the unfilled, square nodes by one node, karate club member 3, in comparison to the clustering according to the normalised Laplacian. In contrast, the unnormalised Laplacian detects 5 nodes less (karate club members 2, 4, 8, 14 and 20) in the cluster marked by the unfilled square nodes, than does the normalised Laplacian.

We find the clustering according to the adjacency matrix A to agree with the ground truth clustering into social factions within the karate club as recorded by [18]. The normalised Laplacians misplace one out of 34 nodes, which is known to be difficult to cluster correctly in the literature [8]. The unnormalised Laplacian however, misplaces 6 nodes, only one of which is known to us to be commonly misclustered. Hence, the unnormalised Laplacian clustering is clearly outperformed by the other two GSOMs when using the first two eigenvectors to find two communities in the karate data set. In [17] the conditions under which spectral clustering using the unnormalised Laplacian converges are shown to be more restrictive than the conditions under which spectral clustering according to the normalised Laplacian converges. [17] hence advocate using the normalised Laplacian for spectral clustering over the unnormalised Laplacian. Our clustering results agree with this recommendation.

Remark 3. GSOM choice can clearly impact cluster inference results via spectral clustering. We suggest considering degree extreme difference as a parameter in graphical signal processing to infer potential impact of the choice of GSOM. \triangleleft

6 Summary and Conclusions

We have compared the spectra of the GSOMs: the adjacency matrix A , the unnormalised graph Laplacian L and the normalised graph Laplacian L_{rw} and found differences in the spectra corresponding to general graphs. For all three pairs of GSOMs the degree extreme difference, $d_{\max} - d_{\min}$, was found to linearly upper bound both the spectral differences, when transforming one of the two spectra by an affine transformation, and the normalised eigengap differences. We explained the monotonicity found in the eigenvalue bounds by partitioning the class of graphs according to their degree extremes and considering the addition/deletion of connected components to/from the graph. Our bounds were illustrated on Zachary's karate network and on a network sampled from a stochastic blockmodel with homophilic and heterophilic cluster structures. *We find that if the degree extreme difference is large, different choices of GSOMs may give rise to disparate inference drawn from network analysis; smaller degree extreme differences will result in consistent inference, whatever the choice of GSOM.* The significant differences in inference drawn from graphical analysis using the different GSOMs were illustrated via the spectral clustering algorithm applied to Zachary's karate network.

References

1. Aleardi, L.C., Salihoglu, S., Singh, G., Ovsjanikov, M.: Spectral measures of distortion for change detection in dynamic graphs. In: Aiello, L.M., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., Rocha, L.M. (eds.) Complex Networks and Their Applications VII, pp. 54–66. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-03414-4_5
2. Bai, Z., Demmel, J., Dongarra, J., Ruhe, A., van der Vorst, H.: Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide. SIAM (2000)
3. Batagelj, V., Mrvar, A.: Pajek datasets (2006). <http://vlado.fmf.uni-lj.si/pub/networks/data/>. Accessed 23 Nov 2016
4. Cvetkovic, D., Gutman, I.: Applications of Graph Spectra: An Introduction to the Literature. *Zbornik radova*, vol. 14, pp. 9–34 (2011)
5. Chen, P.Y., Hero, A.O.: Deep community detection. *IEEE Trans. Signal Process.* **63**, 5706–5719 (2015). <https://doi.org/10.1109/TSP.2015.2458782>
6. Chung, F.R.K.: Spectral graph theory. American Mathematical Society (1997)
7. Crawford, B., Gera, R., House, J., Knuth, T., Miller, R.: Graph structure similarity using spectral graph theory. In: Cherifi, H., Gaito, S., Quattrociocchi, W., Sala, A. (eds.) Complex Networks & Their Applications V, pp. 209–221. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-50901-3_17
8. Fortunato, S.: Community detection in graphs. *Phys. Rep.* **486**, 75–174 (2010). <https://doi.org/10.1016/j.physrep.2009.11.002>
9. Karrer, B., Newman, M.E.J.: Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107 (2011). <https://doi.org/10.1103/PhysRevE.83.016107>
10. Kumar, S., Ying, J., de M. Cardoso, J.V., Palomar, D.P.: A unified framework for structured graph learning via spectral constraints. [arXiv:1904.09792](https://arxiv.org/abs/1904.09792) [stat.ML] (2019)
11. Lei, J., Rinaldo, A.: Consistency of spectral clustering in stochastic block models. *Ann. Stat.* **43**, 215–237 (2015). <https://doi.org/10.1214/14-AOS1274>
12. Ortega, A., Frossard, P., Kovacevic, J., Moura, J.M.F., Vandergheynst, P.: Graph signal processing: overview, challenges, and applications. *Proc. IEEE* **106**, 808–828 (2018). <https://doi.org/10.1109/JPROC.2018.2820126>
13. Rohe, K., Chatterjee, S., Yu, B.: Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Stat.* **39**, 1878–1915 (2011). <https://doi.org/10.1214/11-AOS887>
14. Tremblay, N., Borgnat, P.: Graph wavelets for multiscale community mining. *IEEE Signal Process. Mag.* **62**, 5227–5239 (2014). <https://doi.org/10.1109/TSP.2014.2345355>
15. van Mieghem, P.: Graph Spectra for Complex Networks. Cambridge University Press, Cambridge (2011)
16. von Luxburg, U.: A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007). <https://doi.org/10.1109/TSP.2015.2458782>
17. von Luxburg, U., Belkin, M., Bousquet, O.: Consistency of spectral clustering. *Ann. Stat.* **36**, 555–586 (2008). <https://doi.org/10.1214/009053607000000640>
18. Zachary, W.W.: An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977). <https://doi.org/10.1086/jar.33.4.3629752>
19. Zumstein, P.: Comparison of Spectral Methods Through the Adjacency Matrix and the Laplacian of a Graph. Ph.D. thesis, ETH Zürich (2005)



Induced Edge Samplings and Triangle Count Distributions in Large Networks

Nelson Antunes^{1,2(✉)}, Tianjian Guo³, and Vladas Pipiras³

¹ Center for Computational and Stochastic Mathematics, University of Lisbon,
Avenida Rovisco Pais, 1049-001 Lisbon, Portugal

² University of Algarve, Faro, Portugal
nantunes@ualg.pt

³ Department of Statistics and Operations Research, University of North Carolina,
CB 3260, Chapel Hill, NC 27599, USA
Tianjian.Guo@mccombs.utexas.edu, pipiras@email.unc.edu

Abstract. This work focuses on distributions of triangle counts per node and edge, as a means for network description, analysis, model building and other tasks. The main interest is in estimating these distributions through sampling, especially for large networks. Suitable sampling schemes for this are introduced and also adapted to the situations where network access is restricted or streaming data of edges are available. Estimation under the proposed sampling schemes is studied through several methods, and examined on simulated and real-world networks.

Keywords: Triangles · Random sampling · Distribution estimation · Static and streaming graphs · Power laws

1 Introduction

Triangles formed by edges are the most fundamental topological structures of networks. The number of triangles enters network metrics, such as clustering coefficient or transitivity ratio (e.g. Newman [11]), and serves in network description, analysis and model building. Applications exploiting triangle counts include spam detection (e.g. Beachetti et al. [4]), discovering common topics on the web (Eckmann and Moses [6]), query plan optimization in databases (Bar-Yossef et al. [3]), community detection (Palla et al. [12]), and others.

Triangle counting has also spurred much theoretical research on suitable counting algorithms for various contexts and goals, for example, to scale well with increasing network size, to adapt to streaming data of edges, to work on networks with restricted access, and many others. The recent review paper by Hasan and Dave [1] describes well the many developments in this research direction. Algorithms based on various sampling schemes have also been studied actively, especially for large networks, with the goal of estimating the number of triangles space-efficiently, even if not quite exactly. See, for example, Jha et al. [8], Stefani et al. [13] and the references in [1].

In this work, our focus is also on sampling methods and triangles in large networks. But we go beyond the number of triangles as in much of the earlier literature and consider instead the distributions of triangle counts, per both node and edge. We are interested in estimating these distributions through suitable sampling schemes. Note that our object of study involves a range of counts (i.e. the number of nodes or edges participating in 0, 1, ... triangles) and not just a single count as the total number of triangles. For this reason perhaps, the considered distributions of triangle counts have thus far received relatively little attention in the literature. The few related exceptions are the works by Stefani et al. [13] and Lim et al. [10] who estimate the number of triangles for each node. However, it is yet to be seen whether these estimates translate successfully to obtain the distributions of triangle counts. These distributions are nevertheless informative, showing how triangles are distributed over the network (scattered or concentrated) with respect to nodes and edges. For example, as we discuss below (Sect. 2.3), these distributions might have power-law tails, suggesting the existence of multiple hub-like nodes/edges participating in large numbers of triangles. For instance, in social networks, the number of triangles of a user is used to identify its social role in the network, and provides a good metric in assessing the content quality provided by the user.

Our contributions are several-fold. First, we propose a new sampling scheme tailored for estimating triangle count distributions, which we call *node induced edge subgraph* (NIES) sampling. Some available classical sampling schemes such as induced and incident subgraph sampling (e.g. [9], Ch. 5.3.1), often do not carry enough information for proper inference about triangle count distributions (see Remark 1 below). Other sampling schemes that seem natural for these distributions can be computationally over-expensive (see Remark 2 below). NIES sampling balances these issues and is also the scheme that is easily amenable to theoretical analysis. The latter is our second major contribution, consisting of two main components: estimation through both inversion and asymptotic methods are studied in the paper. Third, we adapt the NIES sampling approach to the situations of restricted access to networks, where we use random walks, and to streaming data of edges, where we employ hashing. Fourth, we examine our proposed methods on simulated and real-world networks.

The rest of the paper is organized as follows. Section 2 presents our sampling framework and estimation approaches based on inversion and asymptotics. Section 3 details our sampling algorithms. Section 4 concerns data applications. Section 5 concludes and discusses directions for future work.

2 Sampling and Estimation Approaches

2.1 Sampling Framework and Quantities of Interest

We represent a network under study as a graph $G = (V, E)$, where V is the set of nodes (vertices) and E is the set of edges. The graph is assumed to be unweighted, undirected and without loops or multiple edges (simple). Let $N = |V|$ and $M = |E|$ denote the numbers of nodes and edges, respectively, which are assumed

to be known for simplicity. The numbers of triangles of a node $w \in V$ and an edge $(u, v) \in E$ are defined by

$$T_w = |\{(u, v) \in E : (u, w), (v, w) \in E\}|, \quad w \in V, \quad (1)$$

$$T_{(u,v)} = |\{w \in V : (u, w), (v, w) \in E\}|, \quad (u, v) \in E. \quad (2)$$

Let $N_t = \max\{T_v\}$ and $M_t = \max\{T_{(u,v)}\}$ be the maximum numbers of triangles per node and edge, respectively. Let also

$$t_v(j) = \frac{1}{N} \sum_{w \in V} 1_{\{T_w=j\}} \quad (3)$$

be the probability that a node v selected at random participates in j triangles, and

$$t_e(j) = \frac{1}{M} \sum_{(u,v) \in E} 1_{\{T_{(u,v)}=j\}} \quad (4)$$

be the probability that an edge e selected at random participates in j triangles. We are interested in the distribution of triangle counts per node $\mathbf{t}_v = (t_v(0), \dots, t_v(N_t))'$ and the distribution of triangle counts per edge $\mathbf{t}_e = (t_e(0), \dots, t_e(M_t))'$. In principle, the largest possible value for M_t in a simple network is $N - 2$ and we assume that in practice, N_t and M_t may be known, or set large enough.

We introduce and focus on the following sampling scheme for estimating the distributions of triangle counts of interest. We follow the idea of Buriol et al. [5] where the basic components to estimate the (total) number of triangles are node-edge pairs selected uniformly to check if they form a triangle. In our setting, we suppose that nodes are sampled at random with probability p_1 and edges with probability p_2 (alternatively, n nodes and m edges are selected, again through simple random sampling without replacement), and then we check for triangles for all pairs of selected nodes and edges. More specifically, random samples of nodes and edges are selected according to Bernoulli trials from V and E , which yields the sets of *sampled nodes* V^* and *sampled edges* E^* , respectively. Additionally, edges are supposed to be observed for all $w \in V^*$ and $(u, v) \in E^*$ such that $(w, u), (w, v) \in E$, yielding with V^* and E^* the subgraph $G^s = (V^s, E^s)$. We refer to the procedure to select G^s as *node induced edge subgraph* (NIES) sampling.

The measured sample analogues of (1) and (2) are

$$T_w^s = |\{(u, v) \in E^* : (u, w), (v, w) \in E\}|, \quad w \in V^*, \quad (5)$$

representing the number of triangles of node $w \in V^*$ (connecting through edges in E) to sampled edges in E^* , where $T_w^s = 0$ if $w \notin V^*$, and

$$T_{(u,v)}^s = |\{w \in V^* : (u, w), (v, w) \in E\}|, \quad (u, v) \in E^*, \quad (6)$$

representing the number of triangles that share an edge $(u, v) \in E^*$ connected to node $w \in V^*$, where $T_{(u,v)}^s = 0$ if $(u, v) \notin E^*$. Note that the quantities (5)

and (6) are not those in (1) and (2) defined for NIES sampling: for example, a triangle formed just by the edges of E^* (with no nodes in V^*) would not be counted in either (5) or (6). Our objective is to estimate \mathbf{t}_v and \mathbf{t}_e from the observed data of T_w^s and $T_{(u,v)}^s$.

Remark 1. For clarity, we also note the following. When $p_1 = 0$, observe that only edges are sampled and that the NIES sampling is the so-called incident subgraph sampling (e.g. [9], Ch. 5.3.1). But we note again that the quantities T_w^s and $T_{(u,v)}^s$ of interest are not triangle counts on the resulting subgraph (since no nodes are sampled when $p_1 = 0$, $T_w^s = 0$ and $T_{(u,v)}^s = 0$ for all nodes w and edges (u, v)). In fact, for incident or induced subgraph sampling, if we considered the triangle counts per node (in the resulting subgraph), this setting would not be amenable to theoretical analysis, in the sense that we could not relate the count distributions per node in the subgraph to those in the original graph (as in Sects. 2.2 and 2.3 for the NIES sampling).

Remark 2. We comment here further on our choice of the sampling scheme. For example, for estimating \mathbf{t}_v , a natural and simple sampling scheme could consist of sampling nodes with probability p_1 and then for each sampled node, calculating the exact number of triangles (by checking all pairs of its neighbors and seeing how many of these also connect) and then just using the empirical distribution of these as an estimate of \mathbf{t}_v . This sampling scheme, however, may not be practical, especially for large networks with heavy-tailed degree distributions, since checking all pairs of neighbors becomes prohibitively expensive for nodes with large degrees. Furthermore, this scheme would not work for sampling a streaming graph in one pass which is also an aim of this work.

The flexibility to have $p_2 < 1$ in our NIES sampling mitigates the aforementioned computational issue. On the other hand, for estimating \mathbf{t}_e , note that taking $p_1 = 1$ is not as big of an issue, since for each sampled edge, it is less likely that both of its nodes have high degrees. The case $p_1 = 1$ can also be considered when estimating \mathbf{t}_v .

2.2 Inversion Problem

In this section, we present a basic approach to estimate \mathbf{t}_v and \mathbf{t}_e from sampled data of T_w^s and $T_{(u,v)}^s$. The approach is based on inversion and is common in similar problems (e.g. [2, 14, 16]).

Distribution of Triangle Counts Per Node. Similarly to (3), let

$$t_v^s(i) = \frac{1}{N} \sum_{w \in V} 1_{\{T_w^s=i\}} \quad (7)$$

be the probability that a node v chosen at random participates in i sampled triangles after a generic NIES sampling. For NIES sampling, we can write

$$t_v^s(i) = \sum_{j=0}^{N_t} P_v(i, j) t_v(j), \quad (8)$$

where $P_v(i, j)$ is the probability that a node with j triangles participates in i sampled triangles, given by

$$P_v(i, j) = (1 - p_1)1_{\{i=0\}} + p_1 \binom{j}{i} p_2^i (1 - p_2)^{j-i}. \quad (9)$$

When $j = i$ and $p_2 = 1$, the last term in (9) should be interpreted as p_1 . In a matrix form, Eq. (8) can be expressed as

$$\mathbf{t}_v^s = \mathbf{P}_v \mathbf{t}_v, \quad (10)$$

where $\mathbf{t}_v^s = (t_v^s(0), \dots, t_v^s(N_t))'$, $\mathbf{P}_v = (P_v(i, j))$ and $\mathbf{t}_v = (t_v(0), \dots, t_v(N_t))'$. A naive estimator for the distribution of triangles per node is

$$\hat{\mathbf{t}}_v = \mathbf{P}_v^{-1} \hat{\mathbf{t}}_v^s, \quad (11)$$

where $\hat{\mathbf{t}}_v^s$ is defined as \mathbf{t}_v^s but using the observed sampled quantities T_w^s .

Distribution of Triangle Counts Per Edge. This distribution can be considered similarly as above. Similarly to (4), let

$$t_e^s(i) = \frac{1}{M} \sum_{(u,v) \in E} 1_{\{T_{(u,v)}^s = i\}} \quad (12)$$

be the probability that an edge e chosen at random participates in i sampled triangles under a generic NIES sampling. Then,

$$t_e^s(i) = \sum_{j=0}^{M_t} P_e(i, j) t_e(j), \quad (13)$$

where $P_e(i, j)$ is the probability that an edge with j triangles participates in i sampled triangles, given by

$$P_e(i, j) = (1 - p_2)1_{\{i=0\}} + p_2 \binom{j}{i} p_1^i (1 - p_1)^{j-i}. \quad (14)$$

When $j = i$ and $p_1 = 1$, the last term in (14) should be interpreted as p_2 . In a matrix form,

$$\mathbf{t}_e^s = \mathbf{P}_e \mathbf{t}_e, \quad (15)$$

where $\mathbf{t}_e^s = (t_e^s(0), \dots, t_e^s(M_t))'$, $\mathbf{P}_e = (P_e(i, j))$ and $\mathbf{t}_e = (t_e(0), \dots, t_e(M_t))'$. A naive estimator for the distribution of triangles per edge is

$$\hat{\mathbf{t}}_e = \mathbf{P}_e^{-1} \hat{\mathbf{t}}_e^s, \quad (16)$$

where $\hat{\mathbf{t}}_e^s$ is the observed analogue of \mathbf{t}_e^s .

Remark 3. Other variations of Eqs. (9) and (14) are possible, such as when a random set of n vertices and m edges are selected from V and E without replacement, respectively, in which case

$$P_v(i, j) = \left(1 - \frac{n}{N}\right) \mathbf{1}_{\{i=0\}} + \frac{n}{N} \frac{\binom{j}{i} \binom{M-j}{m-i}}{\binom{M}{m}}, \quad (17)$$

and similarly for $P_e(i, j)$ replacing (N, n, M, m) by (M, m, N, n) .

Remark 4. If $p_1 = 1$, then \mathbf{P}_e^{-1} (see (14)) can be shown to have a closed form with the non-zero elements

$$P_e^{-1}(0, 0) = 1, \quad P_e^{-1}(0, j) = \frac{1-p_2}{p_2}, \quad P_e^{-1}(i, i) = \frac{1}{p_2}. \quad (18)$$

In this case, the resulting estimator (16) (except $\hat{t}_e(0)$) is just a (scaled) empirical distribution of the true triangle counts per sampled edges.

Regularization Approach. The formulations (10) and (15) are ill-posed for smaller values of p_1 and p_2 , in the sense that the matrices \mathbf{P}_v and \mathbf{P}_e are ill-conditioned. Regularization is a common method used to solve ill-posed problems. For shortness sake, we write $\hat{\mathbf{t}}^s$ for $\hat{\mathbf{t}}_v^s$ or $\hat{\mathbf{t}}_e^s$, \mathbf{P} for \mathbf{P}_v or \mathbf{P}_e , K_t for N_t or M_t , and L_t for N or M . We consider a penalized weighted least-squares approach for the inversion problem with constraints

$$\operatorname{argmin}_{\mathbf{t}} (\hat{\mathbf{t}}^s - \mathbf{Pt})' \mathbf{W}^{-1} (\hat{\mathbf{t}}^s - \mathbf{Pt}) + \lambda \phi(\mathbf{t}) \quad (19)$$

subject to $t(i) \geq 0$, $i = 0, 1, \dots, K_t$, $\sum_{i=0}^{K_t} t(i) = L_t$, where \mathbf{W} is a matrix representing suitable weights taken here to be a diagonal matrix with entries \hat{t}^s , $\phi(\mathbf{t})$ refers to the function regularizing \mathbf{t} , and $\lambda > 0$ is a scalar penalty, that sets the degree of regularization, to be determined separately. When $\lambda = 0$, $\mathbf{W} = \mathbf{I}$ and the constraints are ignored, the minimization (19) yields the inversion estimators (11) and (16). A convenient regularization method uses the quadratic smoothing function

$$\phi_{\text{quad}}(\mathbf{t}) = \sum_{i=0}^{K_t-1} (t(i+1) - t(i))^2. \quad (20)$$

This works well as a reconstruction method when the original distribution is smooth. How to deal with non-smooth distributions is left for future work. The optimization problem (19)–(20) can be written as a quadratic program and solved for \mathbf{t} using standard software. However, the inversion approach generally performs poorly at the distribution tail, especially when the latter is heavy. Estimation in the tail is addressed in the next section.

The parameter λ is chosen based on SURE (Stein's unbiased risk estimation) method proposed by Eldar [7]. Due to the space limitations, we refer the reader to [16] for the details and implementation of the method in a similar context (degree distribution).

2.3 Asymptotic Analysis

In this section, we relate the tails of the distributions of triangle counts per node and edge with the respective tails of sample distributions. Since the latter is observable, the relation can be used to estimate the former. We also consider the case when the original distribution has a power-law tail.

Distribution of Triangle Counts Per Node. The triangle count of a sampled node can be expressed as

$$T_v^s = \sum_{i=1}^{T_v} A_i, \quad v \in V^*, \quad (21)$$

where A_i is equal to 1 if the i th triangle of the node is sampled and 0 otherwise under NIES sampling. Since $T_v^s = p_2 T_v + \sum_{i=1}^{T_v} (A_i - p_2)$ and the second term can be thought as approximately Gaussian with standard deviation of the order $\sqrt{T_v} \ll T_v$ for large T_v , we expect that if v is sampled, then

$$T_v^s \approx p_2 T_v \quad (22)$$

and also

$$\mathbb{P}(T_v \leq t) \approx \mathbb{P}(T_v^s \leq t p_2 | v \in V^*), \quad (23)$$

for large t . Viewing (22) and (23) as relations for continuous random variables,

$$\mathbb{P}(T_v = t) \approx p_2 \mathbb{P}(T_v^s = t p_2 | v \in V^*) \approx p_2 p_1^{-1} \mathbb{P}(T_v^s = t p_2). \quad (24)$$

which allows to estimate the original tail through the r.h.s. of (24). If the random variable T_v has a power-law tail with parameter β , that is, $\mathbb{P}(T_v = t) \approx c\beta t^{-\beta-1}$, for large t , where $\beta > 0$ and $c > 0$, then $\mathbb{P}(T_v^s = t) \approx p_1^{-1} p_2^{-\beta} c\beta t^{-\beta-1}$ has a power-law tail as well with the same parameter.

Distribution of Triangle Counts per Edge. The triangle count of a sampled edge is given by

$$T_e^s = \sum_{i=1}^{T_e} B_i, \quad e \in E^*, \quad (25)$$

where B_i is equal to 1 if the i th triangle of the edge is sampled and 0 otherwise under NIES sampling. As above, we expect that if e is sampled,

$$T_e^s \approx p_1 T_e \quad (26)$$

and also, for large t ,

$$\mathbb{P}(T_e \leq t) \approx \mathbb{P}(T_e^s \leq t p_1 | e \in E^*), \quad (27)$$

$$\mathbb{P}(T_e = t) \approx p_1 \mathbb{P}(T_e^s = t p_1 | e \in E^*) \approx p_1 p_2^{-1} \mathbb{P}(T_e^s = t p_1). \quad (28)$$

If $\mathbb{P}(T_e = t) \approx c\gamma t^{-\gamma-1}$, for large t , where $\gamma > 0$ and $c > 0$, then $\mathbb{P}(T_e^s = t) \approx p_2^{-1} p_1^{-\gamma} c C \gamma t^{-\gamma-1}$ has also the same power-law parameter.

When random sets of n vertices and m edges are selected from V and E without replacement, the analysis above holds with $p_1 = n/N$ and $p_2 = m/M$.

Relations Between Heavy Tails Exponents. The reference to and relevance of power-law tails above should not surprise the reader. On one hand, examples of such real networks appear in Sect. 4 below. On the other hand, such tails are also expected for the following reason. It is well known (e.g. [11], Ch. 10) that the degree distributions of many real networks have power-law tails. That is, if D_v denotes the degree of a randomly selected node, then

$$\mathbb{P}(D_v = d) \approx c_0 \alpha d^{-\alpha-1} \quad (29)$$

for large d , where $c_0 > 0$, $\alpha > 1$. Furthermore, for many real networks, one commonly finds the clustering coefficient of a node v , that is, $T_v / \binom{D_v}{2}$ to be roughly ξ_v where ξ_v varies over a limited range. Then, $T_v \approx \xi_v D_v^2$ and, by conditioning on ξ_v and using (29),

$$\mathbb{P}(T_v = t) \approx \mathbb{P}\left(D_v = \frac{t^{1/2}}{\xi^{1/2}}\right) \approx c_0 \alpha (\mathbb{E} \xi_v^{-\alpha/2-1}) t^{-\alpha/2-1}, \quad (30)$$

for large t . In particular, note that the tail exponent β of the distribution of triangle counts per node relates to α as

$$\beta = \alpha/2. \quad (31)$$

This is also what we typically observe for real networks. Relationship between the tail exponent γ of the distribution of triangle counts per edge and the tail exponent α appears to be more delicate, and will not be discussed here.

3 Algorithms

When a node or edge of a network can be chosen at random, NIES sampling can be carried out as described in Sect. 2.1. Such full access to network, however, may not be available in other scenarios, for example, when networks can only be crawled or when dealing with streaming edges. In this section, we describe two algorithms for NIES sampling to select G_s from G in these sampling scenarios.

3.1 Sampling Static Graphs with Restricted Access

Many real-world networks can only be crawled, i.e. we can only explore the neighbors of the visited node. In this context, sampling procedures are commonly based on random walks. It is assumed that access to one initial node is available and the network is connected or the largest giant connected component covers the majority of the network so that the disconnected parts can be ignored. Two independent random walks could be used to carry out NIES sampling:

Algorithm 1: Random walk algorithm: NIES sampling

Data: static graph G ; **Result:** T_w^s and $T_{(u,v)}^s$ from G^s, E^*, V^* .

- 1 **Initialization:** $(u, v) \leftarrow \text{rand}(V)$, $V^s \leftarrow \{u\}$, $E^s, V^*, E^* \leftarrow \emptyset$;
- 2 **While** $|E^*| < m$ **do**
- 3 **If** $(u, v) \notin E^*$ **then**
- 4 $E^* \leftarrow E^* \cup \{(u, v)\}$;
- 5 **If** $v \notin V^s$ (resp. $(u, v) \notin E^s$) **then** $V^s \leftarrow V^s \cup \{v\}$ (resp. $E^s \leftarrow E^s \cup \{(u, v)\}$);
- 6 **ForEach** $k \in \text{Neighbor}(u) \cap \text{Neighbor}(v)$ **do**
- 7 **If** $k \notin V^*$ (resp. V^s) **then** $V^* \leftarrow V^* \cup \{k\}$ (resp. $V^s \leftarrow V^s \cup \{k\}$);
- 8 **If** (u, k) (resp. (v, k)) $\notin E^s$ **then** $E^s \leftarrow E^s \cup \{(u, k)\}$ (resp. $\{(v, k)\}$);
- 9 $u \leftarrow v$, $v \leftarrow \text{rand}(\text{Neighbor}(u))$;

Algorithm 2: Stream algorithm: NIES sampling

Data: streaming graph G ; **Result:** T_w^s and $T_{(u,v)}^s$ from G^s, E^*, V^* .

- 1 **Initialization:** $V^s, E^s, V^*, E^* \leftarrow \emptyset$;
- 2 **ForEach** edge (u, v) from G **do**
- 3 **If** $\text{hash}_1(u)$ (resp. $\text{hash}_1(v)$) $< p_1$ **and** u (resp. v) $\notin V^*$ **then**
- 4 $V^* \leftarrow V^* \cup \{u\}$ (resp. $\{v\}\}$);
- 5 **If** $\text{hash}_2(u, v) < p_2$ **then** $E^* \leftarrow E^* \cup \{(u, v)\}$;
- 6 **If** $\text{hash}_1(u) < p_1$ **or** $\text{hash}_1(v) < p_1$ **or** $\text{hash}_2(u, v) < p_2$ **then**
- 7 **If** $u \notin V^s$ (resp. $u \notin V^s$) **then** $V^s \leftarrow V^s \cup \{u\}$ (resp. $\{v\}\}$);
- 8 $E^s \leftarrow E^s \cup \{(u, v)\}$;

one performing a basic version of random walk sampling (i.e. selecting a node uniformly at random among the neighbors of the visited node) to sample edges at random (E^*); and the other performing a random walk to sample nodes at random (V^*) using the Metropolis-Hastings algorithm with the edges that connect them to E^* . Alternatively, we propose here a scheme for a single random walk on edges that emulates NIES sampling with $p_1 = 1$; see Algorithm 1. The main idea is that for each sampled edge through the random walk in E^* , the common neighbors of the incident nodes of the edge are observed and added to V^* and V^s , with the edges that connect them to these neighbors also added to E^s (see lines 6–8 of Algorithm 1).

3.2 Sampling Streaming Graphs

Many networks in the online world naturally evolve over time, as new edges/nodes are added to the network. A natural representation of such networks (or streaming graphs) is in the form of a stream of edges. We shall describe how to perform NIES sampling to select G_s from G in one pass when G is presented as a stream of edges in no particular order. Two uniform random hash functions on $[0, 1]$ are used to sample nodes and edges at random with probabilities p_1 and p_2 , respectively; see Algorithm 2 (lines 3–5). We note that in the streaming scenario if a node is sampled (i.e. $\text{hash}_1(u) < p_1$) we need to add its edges to

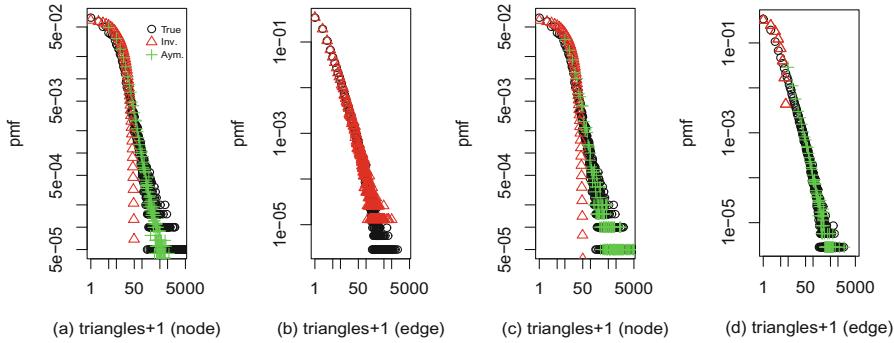


Fig. 1. Power-law network: (a)–(b) Algorithm 1 and (c)–(d) Algorithm 2.

the subgraph G^s as they arrive (lines 6–8 of Algorithm 2), in order to be able to compute the quantities of interest (5) and (6) from the sampled graph. The additional edges that do not enter into the NIES sampling can be deleted at the end of the stream but we omit this step in the algorithm. If, alternatively, we are interested in sampling n nodes and m edges at random without replacement from the stream, we could use reservoir sampling [15] to keep the n nodes and m edges with the minimum hash values in the reservoir. Finally, the algorithm also applies to the case of a static graph with full access, as a one-pass algorithm through the list of edges.

4 Data Applications

We first assess the proposed estimation and sampling methods for the triangle count distributions on a synthetic network. We consider the Chung-Lu model (e.g. [11], Ch. 10), which has the power-law degree distribution $\mathbb{P}(D_v = d) \approx d^{-2.5}$, with 20000 nodes and 350000 edges. In Fig. 1, (a)–(b), the network is sampled through Algorithm 1 where m is equal to 20% of the edges. (Sampling rates to estimate network distributions, e.g. the degree distribution, are in the range of 10%–30%; [16].) Plot (a) shows the true distribution of triangle counts per node and its estimates based on the inversion and asymptotic approaches (all the results presented have been averaged over 20 runs). For the inversion, the penalized estimator (19) allows to recover well only the bulk of the distribution due to the heavy tail of the distribution and a small sampling rate. The estimation in the tail can be recovered through the scaling of the empirical sample distribution of T_v^s as in (24) (r.h.s). We also note that the estimated power-law exponent of the true distribution is $\beta = 0.73$ (using the maximum likelihood estimation in the formula (10.9) of [11]) while for the degree distribution, the true exponent is $\alpha = 1.5$; this agrees with (31). Plot (b) depicts estimation of the triangle count distribution per edge when using inversion. Since $p_1 = 1$ with Algorithm 1, the matrix \mathbf{P}_e is not ill-conditioned with the inverse described by (18). In this case, the naive estimator (16) can be used. We omit the asymptotic

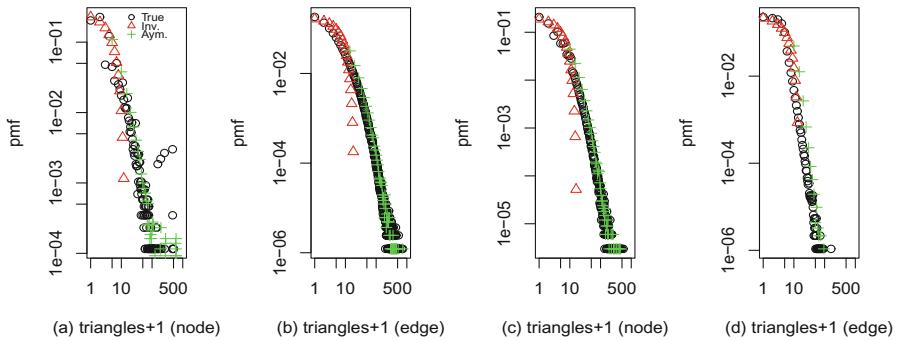


Fig. 2. (a) Arxiv HEP-TH (Algorithm 1), (b) gemsec-Facebook (Algorithm 2), (c) and (d) com-Amazon (Algorithms 1 and 2, respectively).

estimation (28) in the plot since it coincides with the inversion approach in view of (18).

For Fig. 1, (c)–(d), the power-law network is converted into a stream of edges taken in random order. The stream of edges is then sampled through Algorithm 2 with $p_1 = p_2 = 0.2$. The estimation of the distribution per node (plot (c)) using the penalized estimator (19) performs similarly when compared to that in plot (a). The asymptotic estimation (24) though is slightly more accurate in the tail in plot (c). This is due to the fact that with Algorithm 2, edges and nodes are effectively sampled at random, while in the random walk (Algorithm 1) edges are sampled at random only in the stationary limit. The estimation of the distribution per edge is given in plot (d), where the inversion works only with the penalized estimator (19) since $p_1 \neq 1$ is small.

We also consider several real-world networks from SNAP database¹: a collaboration network from the e-print high energy physics-theory (Arxiv HEP-TH) with $N = 8638$ and $M = 24806$; a Facebook social network (gemsec-Facebook) with $N = 50515$ and $M = 819090$; and an Amazon product co-purchasing network (com-Amazon) with $N = 334863$ and $M = 925872$. Figure 2 shows the estimation of the distributions per node and edge for the two sampling algorithms with sampling rate of 20%. The two estimation approaches show that the true triangle count distributions can be recovered quite accurately which agrees with the results for synthetic networks. For the relations between power-law exponents, e.g., we found $\beta = 1.24$ and $\alpha = 2.28$ for the com-Amazon network.

5 Conclusions

In this work, we focused on triangle count distributions in networks, and their estimation through a newly introduced sampling scheme. The scheme can be

¹ <https://snap.stanford.edu/data/index.html>.

emulated via random walks on restricted access networks or hashing in the setting of streaming edges. The proposed estimation methods were based on inversion and asymptotics, with good performance on several synthetic and real-world networks.

Several open questions for future work were already raised above, for example, concerning the relation between the power-law exponents of the degree distribution and the triangle count distribution per edge (see Sect. 2). In other directions, one could possibly consider graphs with repeated edges (multigraphs) or directed graphs, and count distributions per node and edge for higher-order graphical structures other than triangles such as k -cliques.

References

1. Al Hasan, M., Dave, V.S.: Triangle counting in large networks: a review. *WIREs Data Mining Knowl. Discov.* **8**(2), e1226 (2018)
2. Antunes, N., Pipiras, V.: Estimation of flow distributions from sampled traffic. *ACM Trans. Model Perform. Eval. Comput. Syst.* **1**(3), 11:1–11:28 (2016)
3. Bar-Yossef, Z., Kumar, R., Sivakumar, D.: Reductions in streaming algorithms, with an application to counting triangles in graphs. In: Proceedings of the 13th Annual ACM-SIAM SODA, pp. 623–632 (2002)
4. Bechetti, L., Castillo, C., Donato, D., Baeza-YATES, R., Leonardi, S.: Link analysis for web spam detection. *ACM Trans. Web* **2**(1), 2:1–2:42 (2008)
5. Buriol, L.S., Frahling, G., Leonardi, S., Marchetti-Spaccamela, A., Sohler, C.: Counting triangles in data streams. In: Proceedings of the 25th ACM SIGMOD-SIGACT-SIGART PODS, pp. 253–262 (2006)
6. Eckmann, J.-P., Moses, E.: Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proc. Natl. Acad. Sci.* **99**(9), 5825–5829 (2002)
7. Eldar, Y.C.: Generalized SURE for exponential families: applications to regularization. *IEEE Trans. Signal Process.* **57**(2), 471–481 (2009)
8. Jha, M., Seshadhri, C., Pinar, A.: A space-efficient streaming algorithm for estimating transitivity and triangle counts using the birthday paradox. *ACM Trans. Knowl. Discov. Data* **9**(3), 15:1–15:21 (2015)
9. Kolaczyk, E.D.: Statistical Analysis of Network Data. Springer, New York (2009). <https://doi.org/10.1007/978-0-387-88146-1>
10. Lim, Y., Jung, M., Kang, U.: Memory-efficient and accurate sampling for counting local triangles in graph streams: from simple to multigraphs. *ACM Trans. Knowl. Discov. Data* **12**(1), 4:1–4:28 (2018)
11. Newman, M.: Networks: An Introduction, 2nd edn. Oxford University Press Inc., New York (2018)
12. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**(7043), 814–818 (2005)
13. Stefani, L.D., Epasto, A., Riondato, M., Upfal, E.: TriEst: counting local and global triangles in fully dynamic streams with fixed memory size. *ACM Trans. Knowl. Discov. Data* **11**(4), 43:1–43:50 (2017)
14. Tune, P., Veitch, D.: Fisher information in flow size distribution estimation. *IEEE Trans. Info. Theory* **57**(10), 7011–7035 (2011)

15. Vitter, J.S.: Random sampling with a reservoir. *ACM Trans. Math. Softw.* **1**(1), 37–57 (1985)
16. Zhang, Y., Kolaczyk, E.D., Spencer, B.D.: Estimating network degree distributions under sampling: an inverse problem, with applications to monitoring social media networks. *Ann. Appl. Stat.* **9**(1), 166–199 (2015)



Spectral Vertex Sampling for Big Complex Graphs

Jingming Hu, Seok-Hee Hong^(✉), and Peter Eades

School of Computer Science, University of Sydney, Sydney, Australia
`{jihu2855, seokhee.hong, peter.eades}@sydney.edu.au`

Abstract. This paper introduces a new vertex sampling method for big complex graphs, based on the *spectral sparsification*, a technique to reduce the number of edges in a graph while retaining its structural properties. More specifically, our method reduces the number of *vertices* in a graph while retaining its structural properties, based on the high *effective resistance values*. Extensive experimental results using graph sampling quality metrics, visual comparison and shape-based metrics confirm that our new method significantly outperforms the random vertex sampling and the degree centrality based sampling.

1 Introduction

Nowadays many big complex networks are abundant in various application domains, such as the internet, finance, social networks, and systems biology. Examples include web graphs, AS graphs, Facebook networks, Twitter networks, protein-protein interaction networks and biochemical pathways. However, analysis and visualization of big complex networks is extremely challenging due to scalability and complexity.

Graph sampling methods have been used to reduce the size of graphs. Popular graph sampling methods include random vertex sampling, random edge sampling, random path sampling and random walk. However, previous work to compute graph samples based on *random sampling* techniques often fails to preserve the connectivity and important global skeletal structure in the original graph [16, 18].

In this paper, we introduce a new method called the *Spectral Vertex* (SV) sampling for computing graph samples, using an approach based on the *spectral sparsification*, a technique to reduce the number of *edges* in a graph, while retaining its structural properties, introduced by Spielman *et al.* [15]. Roughly speaking, we select *vertices* with high *effective resistance values* [15].

Using real-world benchmark graph data sets with different structures, extensive experimental results based on the graph sampling quality metrics, visual comparison with various graph layouts and shape-based metrics confirm that our new method SV significantly outperforms the *Random Vertex* sampling (RV)

Research supported by ARC Discovery Project.

and the *Degree Centrality* (DC) based sampling. For example, Fig. 1 shows comparison between the graph samples of *gN1080* graph with 25% sampling ratio, computed by RV, DC, and SV methods. Clearly, our SV method retains the structure of the original graph.

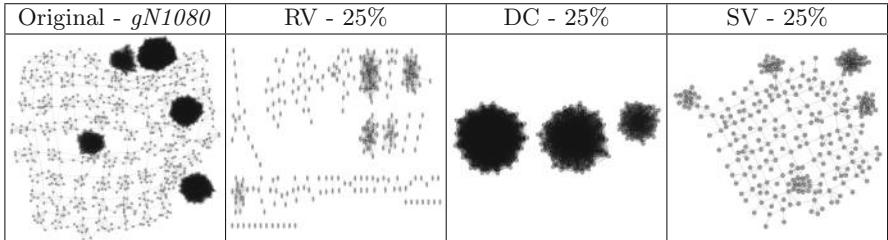


Fig. 1. Comparison of sampling of *gN1080* graph with 25% sampling ratio, computed by RV, DC and SV.

2 Related Work

2.1 Spectral Sparsification Approach

Spielman and Teng [15] introduced the *spectral sparsification*, which samples edges to find a subgraph that preserves the structural properties of the original graph. More specifically, they proved that every n -vertex graph has a spectral approximation with $O(n \log n)$ edges, and presented a stochastic sampling method using the concept of *effective resistance*, which is closely related to the commute distance. The commute distance between two vertices u and v is the average time that a random walker takes to travel from vertex u to vertex v and return [15].

More formally, a sparsification G' is a subgraph of a graph G where the edge density of G' is much smaller than that of G . Typically, sparsification is achieved by stochastically sampling the edges, and has been extensively studied in graph mining. Consequently, there are many stochastic sampling methods available [8, 10]. For example, the most common stochastic sampling is random vertex (resp., edge) sampling (RV) (resp., RE): each vertex (resp., edge) is chosen independently with a probability p .

Furthermore, stochastic sampling methods have been empirically investigated in the context of visualization of large graphs [13, 16]. However, previous work to compute graph samples based on random sampling techniques often fails to preserve the connectivity and important structure in the original graph [16].

2.2 Graph Sampling Quality Metrics

There are a number of quality metrics for graph sampling [8–10]; here we explain some of most popular quality metrics used in [18].

- (i) *Degree Correlation (Degree)* associativity is a basic structural metric [11]. It computes the likelihood that vertices link to other vertices of similar degree, called positive degree correlation.
- (ii) *Closeness Centrality (Closeness)* is a centrality measure of a vertex in a graph, which sums the length of all shortest paths between the vertex and all the other vertices in the graph [5].
- (iii) *Largest Connected Component (LCC)* determines the size of the biggest connected component of a graph. It measures the variation on a fraction of vertices in the largest connected component upon link removals [7].
- (iv) *Average Neighbor Degree (AND)* is the measure of the average degree of the neighbors of each vertex [1]. It is defined as $\frac{1}{|N(v_i)|} \sum_{v_j \in N(v_i)} d(v_j)$, where $N(v_i)$ is the set of neighbors of vertex v_i and $d(v_j)$ is the degree of vertex v_j .

2.3 Shape-Based Metrics for Large Graph Visualization

The *shape-based metrics* [4], denoted as Q in this paper, is a new quality metrics specially designed for big graph visualization. The aim of these metrics is to measure how well the *shape* of the visualization represents the graph. Examples of shape graphs are proximity graphs such as the *Euclidean minimum spanning tree*, the *Relative neighbourhood graph*, and the *Gabriel graph*.

More specifically, the metric computes the *Jaccard similarity* indexes between a graph G and the proximity graph P , which was computed from the vertex locations in a drawing $D(G)$ of a graph G . In particular, the shape-based metrics are shown to be effective for comparing different visualizations of big graphs; see [4].

3 Spectral Vertex Sampling

In this section, we introduce a new method for computing graph samples using *spectral vertex* (SV) sampling; effectively we sample vertices with high *effective resistance values* [15]. Let $G = (V, E)$ be a graph with a vertex set V ($n = |V|$) and an edge set E ($m = |E|$). The *adjacency matrix* of an n -vertex graph G is the $n \times n$ matrix A , indexed by V , such that $A_{uv} = 1$ if $(u, v) \in E$ and $A_{uv} = 0$ otherwise. The *degree matrix* D of G is the diagonal matrix with where D_{uu} is the degree of vertex u . The *Laplacian* of G is $L = D - A$. The *spectrum* of G is the list $\lambda_1, \lambda_2, \dots, \lambda_n$ of eigenvalues of L .

Suppose that we regard a graph G as an electrical network where each edge e is a $1-\Omega$ resistor, and a current is applied. The *effective resistance* $r(e)$ of an edge e is the voltage drop over the edge e . Effective resistance values in a graph can be computed from the Moore-Penrose inverse of the Laplacian [15].

We now describe our new method called SV (Spectral Vertex) for computing spectral sampling $G' = (V', E')$ of $G = (V, E)$. More specifically, we define two variations of the *effective resistance value* $r(v)$ and $r_2(v)$ of a vertex v , as in Algorithms *SV* and *SV2* below. Let $\deg(v)$ represents a degree of a vertex v (i.e., the number of edges incident to v), and E_v represents a set of edges incident to a vertex v .

1. Algorithm SV:

- (a) Compute $r(v) = \sum_{e \in E_v} r(e)$.
- (b) Let V' consist of the n' vertices with largest values of $r(v)$.
- (c) Then G' is the subgraph of G induced by V' .

2. Algorithm SV2:

- (a) Compute $r_2(v) = \frac{\sum_{e \in E_v} r(e)}{\deg(v)}$.
- (b) Let V' consist of the n' vertices with largest values of $r_2(v)$.
- (c) Then G' is the subgraph of G induced by V' .

The running time of the algorithm is dominated by computing effective resistance values, which can be implemented in near linear time [15].

4 Comparison with Random Vertex Sampling

The *main hypothesis* of our experiment is that *SV sampling method performs better than RV sampling method in three ways: better sampling quality metrics, better shape-based metrics, and visually better preserve the structure of the original graph. In particular, we expect a much better performance of the SV algorithm over RV, especially with small sampling ratio.*

To test this hypothesis, we implemented the two spectral vertex sampling algorithms SV and SV2, the random vertex (RV) sampling method, and sampling quality metrics in Python. The Jaccard similarity index and the shape-based metrics Q were implemented in C++. For graph layouts, we used *Visone* Backbone layout [2, 12] and *Yed* Organic layout [17]. We ran the experiments on a Mac Pro 13 laptop, with 2.4 GHz Intel Core i5, 16 GB memory and macOS High Sierra.

Table 1. Data sets.

Graph	V	E	Type
can_144	144	576	grid
G_15	1785	20459	scalefree
G_2	4970	7400	grid
G_3	2851	15093	grid
G_4	2075	4769	scalefree
mm_0	3296	6432	grid
nasa1824	1824	18692	grid
facebook01	4039	88234	scalefree
oflights	2939	15677	scalefree
soc_h	2426	16630	scalefree

(a) Benchmark graphs

Graph	V	E
graph_1	5452	118404
graph_2	1159	6424
graph_3	7885	427406
graph_4	5953	186279
graph_5	1748	13957
graph_6	1785	20459
graph_7	3010	41757
graph_8	4924	52502

(b) GION graphs

Graph	V	E
cN377	377	4790
cN823	823	14995
cN1031	1031	22638
gN285	285	2009
gN733	733	62509
gN1080	1080	17636
gN4784	4784	38135

(c) Black-hole graphs

Table 1 shows the details of the data sets. The first data set consists of *Benchmark* graphs of two types (scale-free graphs and grid graphs); these are real world

graphs from Hachul’s library [3] and the network repository [14]. The second data set is the GION data set, which are RNA sequence graphs with distinctive shapes [4]. The third one is the Black-hole data set, which consist of synthetic locally dense graphs, which are difficult to sample.

4.1 Graph Sampling Quality Metrics

We used the most popular quality metrics used in [18]: Degree correlation (Degree), Closeness centrality (Closeness), Largest Connected Component (LCC) and Average Neighbor Degree (AND). More specifically, we used the *Kolmogorov-Smirnov* (KS) distance value to compute the distance values between two Cumulative Distribution Functions (CDFs) [6]. The KS distance value is between 0 to 1, where the lower KS value means the better result. Namely, the KS distance value closer to 0 indicates higher the similarity between the CDFs.

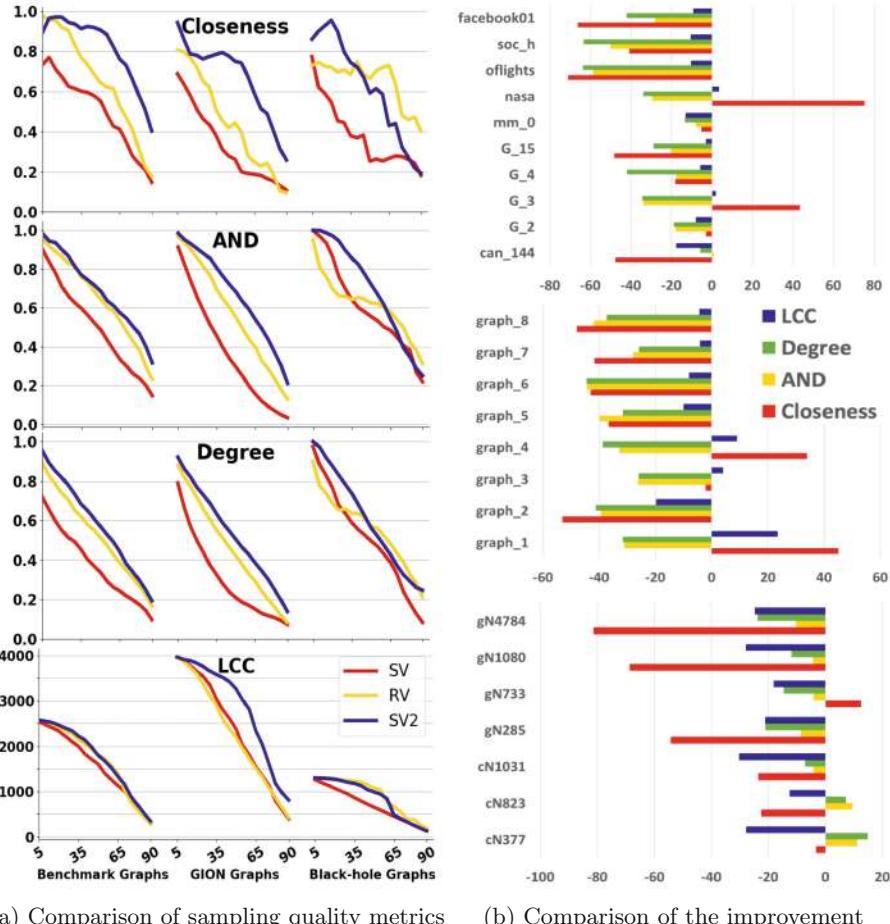
We computed the KS distance value to analyze how well the sampling metrics (Closeness, AND, Degree, and LCC) perform for graph samples computed by SV and RV. Specifically, we computed the average of the KS distance values with four metrics over the three types of data sets with sampling ratios from 5% to 90%.

Figure 2(a) shows the average (for each data set) of the KS distance values of the graph samples computed by SV (red), SV2 (blue), and RV (yellow), with four sampling quality metrics (Closeness, AND, Degree, and LCC) and sampling ratios 5% to 90%. Clearly, the SV method performed consistently the best, as we expected, especially for the Benchmark graphs.

On the other hand, the SV2 method performed a bit worse than RV, which is unexpected. This is due to the fact that in fact many real world graphs have important vertices with high degrees (i.e., hubs). However, the resistance value of such vertex was averaged by the degree of the vertex; therefore these vertices have low resistance values and were not selected. Therefore, for the rest of our experiments, we mainly compare performance between the SV and RV methods.

For detailed analysis per data sets, we computed the *improvement* by the SV method over RV method (i.e., $(KS(SV)/KS(RV)) - 1$), based on sampling metrics. The percentage difference of the average KS distance was computed. The more negative value indicates that the SV method performs better than the RV method.

Figure 2(b) shows the detailed performance improvement of the SV method over the RV method for each data set (Benchmark, GION and Blackhole graphs). We can see that overall for most of data sets and most of sampling quality metrics, the SV method consistently performs better than the RV method, confirming our hypothesis. More specifically, for the Benchmark data set, the SV method produces significantly better results over RV around 35% improvement for all metrics (up to 70% for Closeness). Similarly, for the GION data set, the SV method showed around 30% improvement over RV for all metrics (up to 50% for Closeness). For the Black-hole data set, the SV method showed around 25% improvement for all metrics (up to 80% for Closeness).



(a) Comparison of sampling quality metrics

(b) Comparison of the improvement

Fig. 2. (a) Comparison of sampling quality metrics: the KS values of the sampling quality metrics (Closeness, AND, Degree, and LCC) for graph samples computed by SV (red), SV2 (blue), and RV (yellow), averaged over each data set with sampling ratios from 5% to 90%. The lower KS value means the better result. (b) Comparison of the improvement: *Improvement* by SV over the RV method (i.e., $(KS(SV)/KS(RV)) - 1$), based on sampling metrics using all data sets. The percentage difference of the average KS distance was computed. The more negative value indicates that the SV method performs better than the RV method.

Overall, for Benchmark graphs, the SV method gives significantly better improvement on the sampling metrics, especially for scale-free graphs. For the GION and Black-hole graphs, the SV method also showed significant improvement on all sampling metrics. *In summary, our experimental results with sampling metrics confirm that the SV method shows significant (30%) improvement over the RV method, confirming our hypothesis.*

4.2 Visual Comparison

We experimented with a number of graph layouts available from various graph layout tools including *Yed* [17] and *Visone* [2]. We found that many graph layouts gave similar results for our data sets, except the *Backbone* layout from Visone, which was specifically designed to untangle the hairball drawings of large and complex networks by Nocaj *et al.* [12]. Therefore, we report two graph layouts that gave different shapes: the *Backbone* layout from Visone and the *Organic* layout from Yed, which gave similar shapes to other layouts that we considered. We further observed that the Backbone layout shows better structure for Benchmark graphs (i.e., real world graphs), esp., scale-free graphs, and the Organic layout produces better shape for Black-hole graphs (i.e., synthetic graphs).

We conducted a visual comparison of graph samples computed by the SV method and the RV method for all data sets, using both Organic and Backbone layouts. Figures 3 and 4 show the visual comparison of graph samples computed by the SV and the RV methods. We used the Backbone layout for Benchmark graphs, and the Organic layout for the GION graphs and the Black-hole graphs. Overall, we can see that the SV method produces graph samples with significantly better connectivity with the similar structure to the original graph than the RV method.

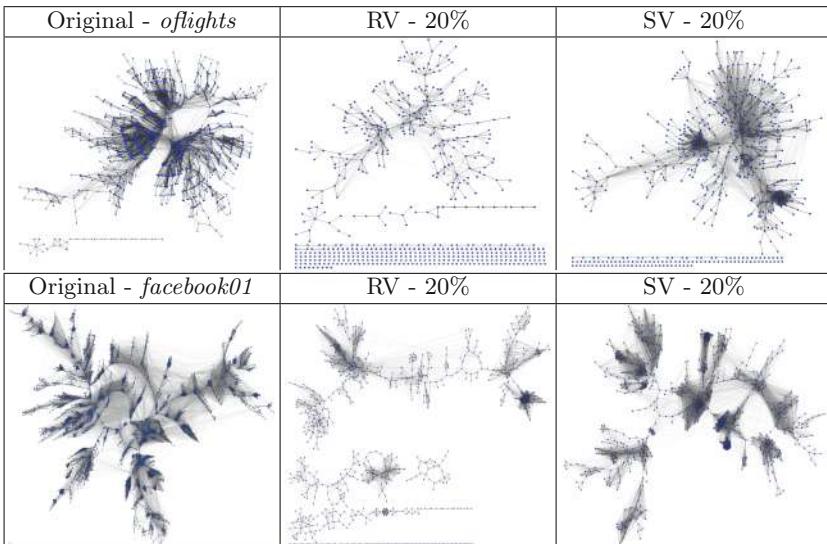


Fig. 3. Samples of *oflights* and *facebook01*: Backbone layout.

In summary, visual comparison of samples computed by the SV method and the RV method using the Backbone and Organic layouts confirms that SV produces samples with significantly better connectivity and similar visual structure to the original graph than RV.

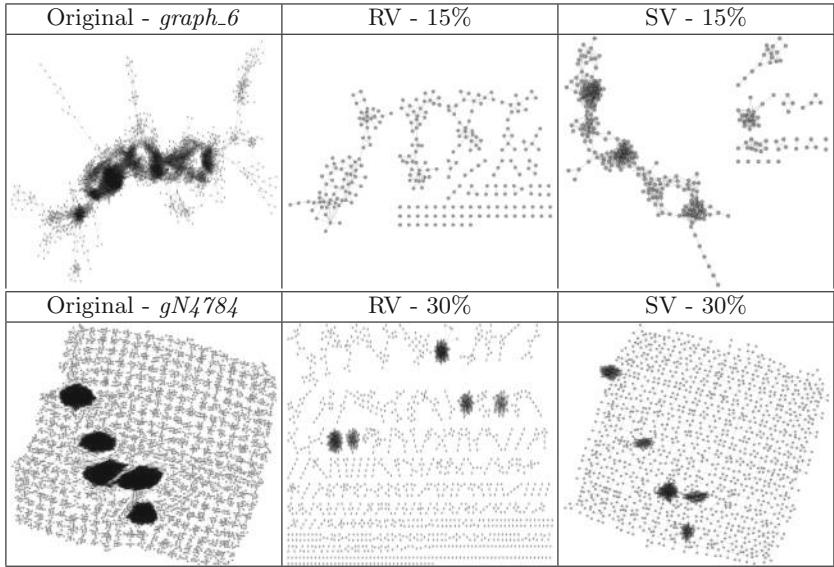


Fig. 4. Samples of *graph_6* and *gN4784*: Organic layout.

4.3 Shape-Based Metrics Comparison

To compare the quality of sampling for visualisation, we used the shape-based metric Q of graph samples computed by the SV (i.e. Q_{SV}) and RV method (i.e. Q_{RV}) using the three data sets. We expect that the shape-based metrics values increase as the sampling ratios increase.

To better compare the shape-based metrics computed by the SV and the RV methods, we use the shape-based metric ratio Q_{SV}/Q_{RV} (i.e., values over 1 means that SV performs better than RV).

Figure 5 shows the comparison of the *average* of shape-based metric ratios Q_{SV}/Q_{RV} per data sets, using the Backbone (red) and Organic (blue) layouts. The x -axis shows the sampling ratios from 5% to 90%, and the y -axis shows the

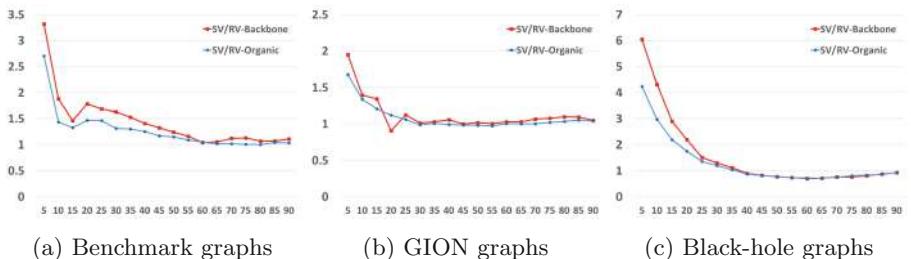


Fig. 5. Average of shape-based metric ratios Q_{SV}/Q_{RV} per data sets with sampling ratios 5% to 90%: Backbone (red) layout and Organic (blue) layout.

value of the ratio. As we expected, for most of the data sets, the shape-based metric ratio is significantly above 1, especially for 5% to 25% sampling ratios. We can observe that overall the Backbone layout shows better performance than the Organic layout, consistent with all the data sets (Benchmark graphs, GION graphs, and Black-hole graphs).

In summary, the SV method performs significantly better than RV in shape-based metrics, esp., when the sampling ratio is low, confirming our hypothesis. The Backbone layout performs significantly better than the Organic layout in terms of the improvement in shape-based metrics by the SV over the RV method.

5 Comparison with Degree Centrality Based Sampling

In this section, we present experiments on comparison of the SV method with the *Degree Centrality* based sampling method (DC), using the shape-based metrics and visual comparison of graph samples.

Degree centrality is one of the simplest centrality measurements in network analysis introduced by Freeman [5]. More specifically, we compute the degree centrality for each vertex, and select the vertices with the largest degree centrality values to compute the graph sample.

The main hypothesis of our investigation is that the SV method perform better than the DC method in two ways: better shape-based metrics, and visually better preserving the structure of the original graph. In particular, we expect a much better performance of the SV method over the DC method, especially with small sampling ratio.

5.1 Visual Comparison

We conducted a visual comparison of graph samples computed by the SV and the DC methods for all data sets, using both Organic and Backbone layouts. We used the Backbone layout for Benchmark graphs and the Organic layout for the GION graphs and the Black-hole graphs.

Figures 6 and 7 show the visual comparison of graph samples computed by the SV and the DC methods with sampling ratio 25% and 30%. Overall, experiments confirm that the SV method produces graph samples with significantly better connectivity and the globally similar structure to the original graph than the DC method. We observe that the DC method produces graph samples with locally dense structure of original graphs rather than the global structure, especially for Black-hole graphs and GION graphs.

In summary, visual comparison of graph samples computed by the SV and the DC methods using Benchmark, GION, and Black-hole data sets with the Backbone and Organic layouts confirms that the SV method produces graph samples with significantly better connectivity structure as well as globally similar visual structure to the original graph than the DC method.

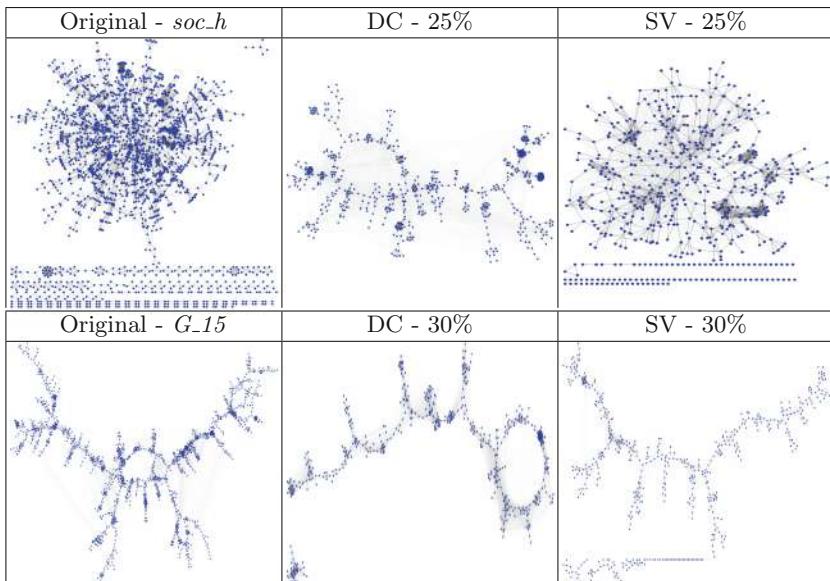


Fig. 6. Graph samples of *soc_h* and *G_15*: Backbone layout.

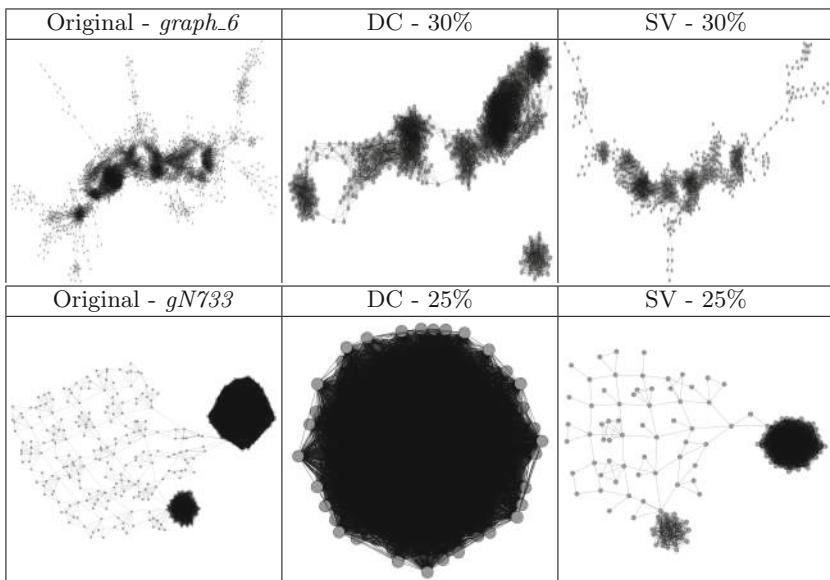


Fig. 7. Graph samples of *graph_6* and *gN733*: Organic layout.

5.2 Shape-Based Metrics Comparison: SV, DC and RV

We now present the overall comparison between the SV, RV, and DC methods using the shape-based metrics. Figure 8 shows a summary of the average of shape-based metric ratios Q_{SV}/Q_{RV} (red) and Q_{DC}/Q_{RV} (blue) per data sets, using the Backbone layout. The y-axis values above 1 means that the SV (resp., DC) method performs better than RV.

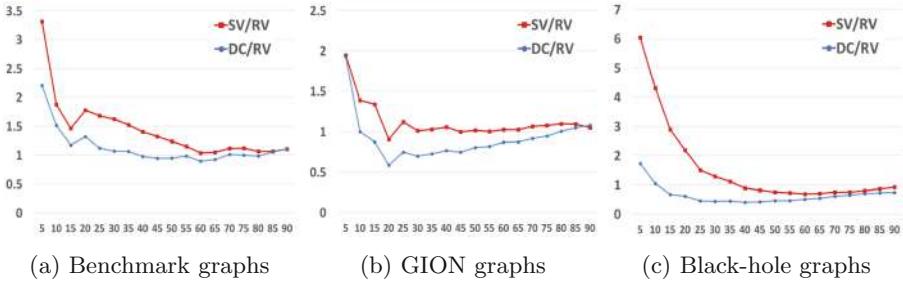


Fig. 8. Summary: average shape-based metric ratios per data sets by SV, DC and RV: SV over RV (i.e., Q_{SV}/Q_{RV}) (red), and DC over RV (i.e., Q_{DC}/Q_{RV}) (blue).

We observe that overall the SV method performs significantly better than the DC method, consistent with all the data sets, especially with small sampling ratio. The DC method performs slightly better than the RV method for the Benchmark graphs when the sampling ratio is small, but mostly worse than the RV method for the GION graph and the Black-hole graphs.

In summary, experiments with shape-based metrics showed that the SV method outperforms the DC and RV methods, and the DC method performs slightly worse than the RV method.

6 Conclusion and Future Work

In this paper, we present a new spectral vertex sampling method SV for drawing large graphs, based on the effective resistance values of the vertices.

Our extensive experimental results using the graph sampling quality metrics with both benchmark real world graphs and synthetic graphs show significant improvement (30%) by the SV method over the RV (Random Vertex) sampling and the DC (Degree Centrality) sampling. Visual comparison using the Backbone and Organic layouts show that the SV method significantly better preserves the structure of the original graph with better connectivity, esp., for low sampling ratio. Our experiments using shape-based metrics also confirm significant improvement by SV over RV and DC methods, esp., for scale-free graphs.

For future work, we plan to improve the performance of the SV method, by integrating graph decomposition techniques and network analysis methods.

References

1. Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **101**(11), 3747–3752 (2004)
2. Brandes, U., Wagner, D.: Analysis and visualization of social networks. In: *Graph Drawing Software*, pp. 321–340 (2004)
3. Davis, T.A., Hu, Y.: The University of Florida sparse matrix collection. *ACM Trans. Math. Softw.* **38**(1), 1–25 (2011)
4. Eades, P., Hong, S.-H., Nguyen, A., Klein, K.: Shape-based quality metrics for large graph visualization. *J. Graph Algorithms Appl.* **21**(1), 29–53 (2017)
5. Freeman, L.C.: Centrality in social networks conceptual clarification. *Soc. Netw.* **1**(3), 215–239 (1978)
6. Gammon, J., Chakravarti, I.M., Laha, R.G., Roy, J.: *Handbook of Methods of Applied Statistics* (1967)
7. Hernández, J.M., Van Mieghem, P.: Classification of graph metrics. Delft University of Technology: Mekelweg, The Netherlands, pp. 1–20 (2011)
8. Hu, P., Lau, W.C.: A survey and taxonomy of graph sampling. arXiv preprint [arXiv:1308.5865](https://arxiv.org/abs/1308.5865) (2013)
9. Lee, S.H., Kim, P.-J., Jeong, H.: Statistical properties of sampled networks. *Phys. Rev. E* **73**(1), 16102 (2006)
10. Leskovec, J., Faloutsos, C.: Sampling from large graphs. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 631–636 (2006)
11. Newman, M.E.J.: Mixing patterns in networks. *Phys. Rev. E* **67**(2), 26126 (2003)
12. Nocaj, A., Ortmann, M., Brandes, U.: Untangling the hairballs of multi-centered, small-world online social media networks. *J. Graph Algorithms Appl.* **19**(2), 595–618 (2015)
13. Rafiei, D.: Effectively visualizing large networks through sampling. In: *VIS 05. IEEE Visualization*, pp. 375–382 (2005)
14. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: *AAAI 2015 Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 4292–4293 (2015)
15. Spielman, D.A., Teng, S.-H.: Spectral sparsification of graphs. *SIAM J. Comput.* **40**(4), 981–1025 (2011)
16. Wu, Y., Cao, N., Archambault, D.W., Shen, Q., Qu, H., Cui, W.: Evaluation of graph sampling: a visualization perspective. *IEEE Trans. Visual Comput. Graph.* **23**(1), 401–410 (2017)
17. yWorks GmbH: yEd graph editor (2018). <https://www.yworks.com/yed>
18. Zhang, F., Zhang, S., Wong, P.C., Edward Swan II, J., Jankun-Kelly, T.: A visual and statistical benchmark for graph sampling methods. In: *Proceedings IEEE Visual Exploring Graphs Scale*, October 2015



Attributed Graph Pattern Set Selection Under a Distance Constraint

Henry Soldano^{1,2(✉)}, Guillaume Santini², and Dominique Bouthinon²

¹ Nukkai Lab, Paris, France

henry.soldano@nukk.ai

² LIPN, Université Paris-Nord UMR-CNRS, 7030 Paris, France

Abstract. Pattern mining exhaustively enumerates patterns that occur in some structured dataset and each satisfy some constraints. To avoid redundancy and reduce the set of patterns resulting from the enumeration, it is necessary to go beyond the individual selection of patterns and select a pattern subset which, as a whole, contains relevant and non redundant information. This is particularly useful when enumerating bi-patterns, which represent pairs of attribute patterns describing for instance subnetworks in two-mode attributed networks. We present and experiment a general greedy algorithm performing pattern set selection on attributed graphs.

Keywords: Closed pattern mining · Core subgraph · Attributed network · Bi-pattern mining · Pattern set selection

1 Introduction

Pattern mining exhaustively enumerates attribute patterns that each occurs in part of the entries of some structured dataset and satisfies some constraints. However, individually selecting patterns may still result in a large number of patterns. This is often due to a large redundancy in the pattern set, and a natural way to reduce the pattern set size and redundancy is to apply some *pattern set selection* process. This may be partially performed during the search, for instance by only selecting *closed patterns*: a closed pattern is the most specific pattern among those occurring in the same entries of the dataset. However, adding a post-processing subset selection, guided by measures of *pattern interestingness*, allows a more accurate control on the pattern set selection process [17].

The general context of this work is the investigation of attributed networks, i.e. networks whose vertices are described according to attributes values, which is a difficult task, including subtasks as community detection [18] and *attributed graph mining*. In the latter the mining process searches for subnetworks whose vertices share common attribute values and which satisfy topological requirements [6, 12]. We consider here attributed graph mining through enumeration of clore closed patterns in which each pattern is associated to a core subgraph [13]. We are in particular interested in a potentially highly redundant pattern mining

process called *bi-pattern* mining designed to mine attributed directed networks or two-mode networks [14].

In this article we will propose $g\beta$ a simple and general post-processing pattern subset selection scheme that selects within a pattern set P a pattern subset S such that (i) in S pairwise distances between patterns all exceed some threshold β and (ii) S maximizes the sum of the individual interestingness g of its patterns. This supposes that we have some distance definition on patterns together with some positive interestingness measure that allows to order the patterns in P . Overall the methodology is as follows:

1. Exhaustively search for a set P of patterns each satisfying a set of constraints.
2. Order patterns in P according to an interestingness measure g .
3. Apply $g\beta$ pattern set selection to the pattern set P and return S .

We first introduce our $g\beta$ greedy algorithm. We then describe how to apply pattern set selection to single pattern mining and bi-pattern mining of attributed networks. Finally we describe various experiments on real attributed networks.

2 Problem Statement and Greedy $g\beta$ Algorithm

We consider a set P together with a distance d between elements. Let S be a subset of P and β a distance threshold. S is called a *candidate* when it satisfies two conditions: (C1) For all pairs $(x, y) \in S$, we have $d(x, y) > \beta$, and (C2) S is maximal: there is no larger subset $S' \supset S$ that satisfies condition C1. Let g be a positive mapping on P and f be defined as $f(S) = \sum_{x \in S} g(x)$. We consider then the $g\beta$ *set selection* problem as follows:

- Find $S^* \subseteq P$ such that S^* is a candidate and $f(S^*)$ is maximum

Example 1. Figure 1 displays a set $P = \{1, 2, 3, 4, 5\}$ in which only (x, y) pairs such that $d(x, y) \leq \beta$ are $(1, 4)$ and $(3, 5)$. There are various candidates such as $S = \{1, 2, 3\}$, $S = \{2, 4, 5\}$, $S = \{1, 2, 5\}$. For instance $S = \{1, 2, 3\}$ is a candidate because we cannot add to S neither 4 (too close to 1) nor 5 (too close to 3). When assuming that for any element i , we have $g(i) = i$, there is one single solution $S^* = \{2, 4, 5\}$, with maximum g value 11.

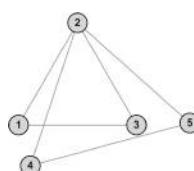


Fig. 1. Two candidates (figured as triangles) in the $g\beta$ set selection of Example 1

When considering a graph whose vertices are the elements of P and whose edges relate vertices x, y such that $d(x, y) \leq \beta$, a candidate S is a maximal stable. The $g\beta$ selection problem is then equivalent to the Maximum Weight Stable problem with g as the weight function and is consequently NP-complete in the general case [5]. We provide below a greedy algorithm $g\beta$ that returns an approximate solution S to our problem. The greedy $g\beta$ algorithm has worst case complexity $\mathcal{O}(|P||S|)$ both in number of comparisons and number of distances to compute:

```
Greedy  $g\beta(P)$ 
//  $P$  is ordered by decreasing  $g$  values
 $F \leftarrow P; S \leftarrow \emptyset$ 
while  $F \neq \emptyset$ 
    Find first  $x$  in  $F$  such that for any  $y$  in  $S$ ,  $d(x, y) > \beta$ .
    if  $x$  has been found
        Remove  $x$  from  $F$  as well as all his predecessors in  $F$ .
         $S \leftarrow S \cup \{x\}$ 
    else
         $F \leftarrow \emptyset$ 
    endIf
endWhile
return  $F$ 
end
```

Example 2. The $g\beta$ greedy algorithm runs our Example 1 problem as follows:

- (1) After the initialisation step $F = [5, 4, 3, 2, 1]$ and $S = []$.
- (2) After iterations (1) and (2) we have $F = [3, 2, 1]$ and $S = [5, 4]$,
- (3) At iteration (3) the next element in F with distance to elements of S above β is 2 (as $d(3, 5) \leq \beta$), which results in $F = [1]$ and $S = [5, 4, 2]$.
- (4) The algorithm stops after iteration (4) in which 1, the only element in F , is too close to 4 to be added to S , thus resulting in $F = []$ and $S = [2, 4, 5]$.

We will apply $g\beta$ set selection to pattern sets P resulting from attributed networks mining introduced hereunder.

3 Single and Bi-pattern Mining

3.1 Core Closed Single Pattern Mining

In standard closed itemset mining, we have a set V of objects each described as an *itemset*, i.e. a subset of a set of items I . A pattern q is an itemset as well. The *support set* $e = \text{ext}(q)$ of pattern q represents the objects v in which pattern q occurs, i.e. such that $q \subseteq v$. Given some pattern q , the most specific pattern c with support set $\text{ext}(q)$ is the representative of the class of patterns with same support set as q and is called a *closed pattern*. It is obtained as $c = f = \text{int} \circ \text{ext}(q)$

where the operator int applies to a set of itemsets and returns their intersection. *Core closed pattern mining* [13] follows from the remark that applying an interior operator¹ p to $\text{ext}(q)$ we define a coarser equivalence relation. The most specific pattern c of the class of patterns with same *core support set* $e = p(\text{ext}(q))$ as pattern q is obtained as:

$$c = f(q) = \text{int} \circ p \circ \text{ext}(q) \quad (1)$$

More precisely, let V be the vertex set of a graph G , $W \subseteq V$ be a vertex subset and $C = p(W)$ be the *core* of the subgraph G_W induced by W according to some core definition [4]. It has been shown that the operator p is then an interior operator [13]. A core definition is always associated to a *core property* about a vertex v within a vertex subset W . As an example, the k -core of the subgraph G_W is the largest vertex subset $C \subseteq W$ such that all vertices in C have degree at least k in G_C [11]. Consider now that G is an attributed graph whose vertices are described as itemsets. Let then W be the support set of some pattern q , then C is its core support set, G_C the *pattern q core subgraph* and $c = \text{int}(C)$ the associated *core closed pattern* with same core subgraph as q . We display Fig. 2 an attributed network together with a 2-core pattern subgraph and its associated core closed pattern.

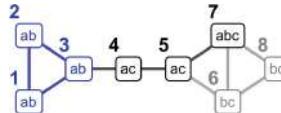


Fig. 2. An attributed network whose vertex labels are subsets of abc . The pattern a occurs in vertices 123457 so inducing the pattern a subgraph represented as bold vertices and edges. Within the pattern a 2-core subgraph, represented in blue, each vertex has degree (at least) 2. The labels of its vertices, namely vertices 123, have in common ab which is therefore a 2-core closed pattern.

3.2 Core Closed Bi-pattern Mining

Bi-pattern mining has been introduced in [14]. Let I_1 and I_2 be two sets of items and V_1 and V_2 be two vertex sets. Bi-pattern mining of an attributed graph considers pattern pairs $q = (q_1, q_2)$ where q_1 and q_2 are respectively subsets of I_1 and I_2 . This way we define $\text{ext}(q_1, q_2)$ as a pair of support sets, as well as $\text{int}(W_1, W_2)$ as the bi-pattern obtained by intersecting description of vertices from W_1 on one hand and description of vertices from W_2 in an other hand. Bi-pattern mining relies on applying to a vertex subset pair (W_1, W_2) an operator that reduces both components in such a way that a *bi-core* property is satisfied. More precisely, given a pair of vertex subset (W_1, W_2) we first consider the subgraph induced by (W_1, W_2) , i.e. the subgraph G_{W_1, W_2} made of the edges

¹ Interior operators p are monotonic, idempotent and such that for any X , $p(X) \leq X$.

relating W_1 and W_2 , and then we reduce this subgraph by removing vertices until the bi-core property is satisfied. We give hereunder the bi-core definition we use for directed networks:

- The $h\text{-}a$ BHA bi-core (H, A) of the directed subgraph G_{W_1, W_2} is such that in the subgraph $G_{H,A}$ induced by the directed edges relating H to A , all vertices in H have outdegree at least h and all vertices in A have indegree at least a .

The associated bi-core operator p is an interior operator such that $(H, A) = p(W_1, W_2)$. We then find back Eq. 1 and define accordingly the core closed bi-pattern c as the most specific bi-pattern (considering both components) whose core support set pair is the same as the bi-pattern q core support set pair.

We display Fig. 3 the bi-core subgraphs associated to two core closed bi-patterns of a two-mode network. The two mode network is considered as a directed network with edges directed from blue vertices towards red vertices.

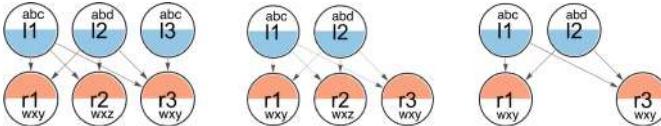


Fig. 3. Two 2-2 BHA bi-cores of a two-mode network. Blue vertices are described as subsets of $abcd$ while red vertices are described as subsets of $wxyz$. The leftmost part displays the whole network. In the middle we have its 2-2 BHA-bi-core associated with the closed bi-pattern (ab, wx) : all its blue vertices labels contain ab while all red vertices labels contain wx . The rightmost part of the figure displays the 2-2 BHA bi-core associated with the bi-pattern (ab, wxy) .

In a BHA bi-core vertices from H are called *hubs* while vertices from A are called *authorities*. This is a reference to the hubs and authorities indexes as introduced by Kleinberg [7]. Directed networks are defined with a single vertex set V , consequently the components C_1 and C_2 of the bi-core may intersect and a bi-pattern is made of two subsets of the same set of items I .

When performing single pattern mining, we will use the $h\text{-}a$ HA-core C of a single pattern subgraph G_W [15]. C is then obtained as the union $H \cup A$ of the bi-core components of $G_{W,W}$.

4 Attributed Graph Pattern Set Selection

To apply $g\beta$ to pattern set selection we first need a distance d between patterns. In order to compare two different pattern set selections of the same pattern set P , we will also define a distance between pattern subsets. To apply $g\beta$ we will also need interestingness measures, first to perform individual selection prior to pattern set selection and then to define the decreasing ordering of P according to a mapping g .

4.1 Distances

In the single pattern mining case, the core closed pattern q is associated to a vertex subset, its core support set $p \circ \text{ext}(q)$. As a distance $d(q, q')$ between patterns q and q' we will use the Jaccard distance between their core support sets. Recall that the Jaccard distance between two subsets X and X' of some set has range $[0,1]$ and is defined as $d_J(X, X') = 1 - \frac{|X \cap X'|}{|X \cup X'|}$. We then have:

$$d(q, q') = (d_J(W, W'))$$

where W is the core support set of q and W' is the core support set of q' .

Regarding bi-patterns we compute the distances between their pattern components and take the least value. This is a conservative choice: when bi-pattern q is selected, to remove bi-pattern q' both components of q' have to be at distance less than β from q . We have then:

$$d(q, q') = \max(d_J(H, A), d_J(H', A'))$$

where (H, A) is the core support set pair of bi-pattern q and (H', A') is the core support set pair of bi-pattern q' .

Regarding pattern subsets we define the distance d_s between two subsets S_1 and S_2 of some set P as $d_s(S_1, S_2) = \max(m_1, m_2)$ where

$$\begin{aligned} d_1(q) &= \min_{q_2 \in S_2} d(q, q_2), d_2(q) = \min_{q_1 \in S_1} d(q, q_1), \\ m_1 &= \frac{\sum d_1(q_1)}{|S_1|} \text{ and } m_2 = \frac{\sum d_2(q_2)}{|S_2|}. \end{aligned}$$

In this definition, we first consider the distance of each element of S_1 to its closest element in S_2 and compute the average m_1 of such minimal distances. In the same way we compute the average minimal distance m_2 of elements of S_2 to elements of S_1 . $d_s(S_1, S_2)$ returns the greatest of these two values.

Computation of $d_s(S_1, S_2)$ needs $\mathcal{O}(|S_1 * |S_2|)$ operations (i.e. comparisons, additions and divisions). When applied to $g\beta$ selections d_s has the following properties:

Proposition 1. Let S_1 and S_2 be two set selections of P with parameters (g_1, β_1) and (g_2, β_2) , then

- $0 \leq d_s(S_1, S_2) \leq \max(\beta_1, \beta_2)$
- If $\beta_1 = \beta_2 = 0$ then $d_s(S_1, S_2) = 0$
- If $\beta_1 = \beta_2 = 1$ then $S_1 = \{q_1\}$, $S_2 = \{q_2\}$ and $d_s(S_1, S_2) = d(q_1, q_2)$.

4.2 Selecting and Ordering Patterns

Consider an interestingness measure μ and a pattern set P . At the individual level, selecting pattern q means requiring that some constraint as $\mu(q) \geq \alpha$ is satisfied. Any such measure may also be used to order the patterns. Our general methodology consists in first selecting a pattern subset whose patterns all

satisfy a set of constraints, related to interestingness measures, then ordering this pattern set by decreasing order of one interestingness measure. The resulting pattern list is the input of the algorithm $g\beta$ that performs the pattern set selection process. The pattern ordering may have a high impact on the pattern set we obtain. We use the following measures in our experiments:

- *Local modularity* $\text{lm}(q) \geq m$. Local modularity of an induced subgraph is the contribution of its vertex subset to the modularity associated to a partition of a network [10]. Local modularity beyond 0 means more internal edges in the subgraph than expected. $\text{lm}(q) \geq m$ was used as a selection criteria in [1, 2].
- *Inhomogeneity* $\text{ih}(q) \geq m$. It is the Jaccard distance between the support sets of the two components of a bi-pattern (see Sect. 3.2) when mining one-mode networks [14]. It is used to focus on bi-patterns we cannot observe with single pattern mining.
- *Deviation to expected core size* $\text{sd}(q) \geq m$. Given a pattern q with support size S we may compare the size x of the core of its subgraph to the expected size \hat{x} of the core of a subgraph induced by x vertices randomly drawn from V . The value of $\text{sd}(q)$ is expressed in number of standard deviation as $(x - \hat{x})/\sigma_x$ [15]. Patterns with high $\text{sd}(q)$ values are indicative of network homophily [9].

5 Experiments

5.1 Datasets, Measures and Orderings

We mainly use two datasets described below:

Lawyers Advice network. This dataset concerns a network study of corporate law partnership that was carried out in a Northeastern US corporate law firm from 1988 to 1991 in New England [8]. It concerns 71 attorneys (partners and associates) of this firm². In the Lawyer Advice network a directed edge $x \rightarrow y$ links two lawyers whenever x tends to go to y for basic professional advice. There are 892 such edges in the network. The vertices are labelled with the lawyers description as attribute values regarding status (associate or partner), gender, office location (Boston, Hartford, Providence), age and seniority in the firm.

LastFM. LastFM, which was used in a work of Galbrun and co-authors [6], is a social network of last.fm community where individuals are described by the artists or groups they have listened. It is made of 1892 vertices related by 12717 undirected edges. vertex labels are subsets of items representing 17625 artists and groups. The average itemset size is 18.

Single and bi-pattern mining is performed using the minerLC software³. In our experiments we considered four orderings of the pattern set. Our default ordering is the Dev ordering that favors unexpected patterns:

- “Dev”: Decreasing deviations $\text{sd}(q)$ from the q subgraph expected core size.

² https://www.stats.ox.ac.uk/~snijders/siena/Lazega_lawyers_data.htm.

³ <https://lipn.univ-paris13.fr/MinerLC/>.

- “ISupp”: Increasing core supports $|p \circ \text{ext}(q)|$
- “DSupp”: Decreasing core supports $|p \circ \text{ext}(q)|$
- “DLM”: Decreasing local modularities of the pattern q core subgraphs.

To analyse our experiments we use various measures of pattern set selection:

- The *covering* $\text{cov}(S)$ represents the number of vertices *covered* by S , i.e. which belongs to the core support set of at least one pattern from S .
- The *redundancy ratio* is $\text{qred}(S) = \text{red}(S)/|S|$ where $\text{red}(S)$ is the ratio of the total number of vertices in the core support sets of patterns from S to the covering of S . $\text{qred}(S)$ is the average proportion of patterns in S to which a vertex covered by S belongs.
- The *mean core support* $\overline{\text{cs}}(S)$ of patterns in S .

In the experiments we denote by S_β^O pattern subsets selected with ordering O and distance threshold β . When using the default ordering Dev, we simply write S_β .

5.2 Pattern Set Selection in Core Closed Single Pattern Mining

We first consider the LastFM network. In this online network there are many vertices, each described by a small number of items from a large set. Applying a 4-core constraint we obtain a set of 61560 core closed patterns. We report Table 1 the number of selected patterns in S_β selections together with the values of various measures. β ranges from 0.4 to 0.9 and the pattern set is ordered following Dev. The redundancy ratio seems to be stable on the whole β range: a vertex is covered in average by less than 3/100 of the selected patterns. The mean core support of S_β selections is stable (between 18 and 25 vertices) and smaller than the mean core support of P , except for $\beta = 0.9$.

Table 1. Measures on pattern sets S_β from the LastFM dataset 4-4 HA closed patterns

β	#Patterns	Covering	Redundancy ratio	Mean core support
0	61560	1211	0.032631	39.5161
0.4	3135	1211	0.0188593	22.8386
0.5	1594	1211	0.0173271	20.9831
0.6	843	1211	0.0175879	21.2989
0.7	402	680	0.0273229	18.5796
0.8	157	1211	0.0209965	25.4268
0.9	51	1211	0.0328363	39.7647

We then consider the Lawyers Advice network. In [15], MinerLC was run on 4-4 HA cores on this network and resulted in 930 patterns. We represent Table 2

Table 2. Measures on pattern sets S_β from the Lawyers dataset 4-4 HA closed patterns

β	#Patterns	Covering	Redundancy ratio	Mean core support
0	930	70	0.298971	20.928
0.3	164	70	0.239547	16.7683
0.4	100	70	0.222143	15.55
0.5	56	69	0.214545	14.8036
0.6	37	70	0.215444	15.0811
0.7	22	69	0.195652	13.5
0.8	10	57	0.201754	11.5
0.9	6	70	0.254762	17.8333

the results of S_β selections using the Dev ordering and β values ranging from 0.3 to 0.9.

We observe that the redundancy ratio is stable on the whole range while the covering is near the maximum value 70 for β values up to 0.7. Note that the Dev ordering tends to favor patterns with small (global) support but still non empty cores (see [15]). $S_{0.9}$ selects six patterns, one of which is q_0 the most general one (found at the end of the ordering) with core support 70. The mean core support slowly decreases (as more unexpected patterns are selected) except regarding $S_{0.9}$ that contains q_0 . When comparing Table 2 to Table 1 which investigates the LastFM we observe that the redundancy ratio is as stable but at a level one order of magnitude lower in the LastFM table.

We also report Table 3 distances d_s between S_β^{Dev} and selections obtained with various orderings. As suspected Dev produces selections close to ISupp and, accordingly distant from DSupp. Overall, all distances increase with β values as less patterns are selected.

Table 3. Distance between pattern sets S_β obtained with various orderings.

β	Dev vs. ISupp	Dev vs. DLM	Dev vs. DSupp
0.3	0.0385637	0.116855	0.123451
0.4	0.0779205	0.161736	0.184607
0.5	0.177322	0.274963	0.314344
0.6	0.229075	0.320703	0.307019
0.7	0.221967	0.488141	0.550112
0.8	0.276306	0.584953	0.581629
0.9	0.490625	0.705413	0.482323

5.3 Pattern Set Selection in Core Closed Bi-pattern Mining

Bi-pattern Set Selection with No Previous Individual Selection. We investigate now core closed bi-pattern selection. Our experiments concern the Lawyers network using 4-4 BHA bi-cores. Recall that whenever we consider a bi-pattern with identical components (q, q) the 4-4 BHA-bi-core is the same as the 4-4 HA-core of pattern q . So, basically the core constraint is the same as in Lawyers network experiments Sect. 5.2. We report Table 4 some measures on the S_β bi-pattern selections with β ranging from 0.3 to 0.7. When comparing these results to those on the 4-4 HA core closed single patterns displayed Table 2 we observe that the decrease with increasing β values is much steeper: At $\beta = 0.3$ we select 9055 bi-patterns, i.e. about 3% of the 293490 bi-patterns while we select 164 single patterns, i.e. about 17% of the 930 single patterns. These results were expected as bi-patterns tends to have more neighbours than single patterns which make selection more useful. The redundancy ratios and mean core supports are similar to those of single patterns selections.

Table 4. Measures on S_β from the Lawyers dataset 4-4 BHA closed bi-patterns

β	#Bi-pattern	Covering	Redundancy ratio	Mean core support
0	293490	70	0.382115	26.748
0.3	9055	70	0.269848	18.8893
0.4	3258	70	0.249974	17.4982
0.5	1093	70	0.237159	16.6011
0.6	522	70	0.212178	14.8525
0.7	190	70	0.196541	13.7579

Bi-pattern Set Selection with Previous Individual Selection. In this section we first apply an individual selection process to the bi-patterns resulting from the mining process. We used various criteria, starting from selecting inhomeogenous bi-patterns. Inhomeogenous bi-patterns are those in which the two patterns are strongly different. In the Lawyers Advice network we search for bi-patterns representing lawyers that ask for advice from lawyers that differs from them (for instance older, with a different law practise or a higher position in the firm). In what follows we use our default Dev ordering.

When selecting bi-patterns with homogeneity at most 0.1 we obtain a selection P_0 of 29 186 inhomogeneous bi-patterns among the 293 490 4-4 BHA bi-patterns obtained in [14]. From P_0 we select those whose deviation to expected core sizes were above 3 standard deviations (i.e. the first bi-patterns in the Dev order) resulting in a subset P_1 of 16 217 bi-patterns. Finally we apply to P_1 a moderate local modularity constraint requiring pattern core subgraphs to have local modularity at least 0.03, leading to a subset P_2 of 2 000 bi-patterns. We report Table 5 the selections S_β from P_2 for various β values.

Table 5. Measures on S_β from the P_2 subset of Lawyers 4-4 BHA closed bi-patterns

β	#Bi-pattern	Covering	Redundancy ratio	Mean core support
0	2000	69	0.343906	23.729
0.5	37	68	0.310413	21.108
0.7	10	61	0.306557	18.700
0.8	8	59	0.311441	18.375
0.9	4	54	0.319444	17.250

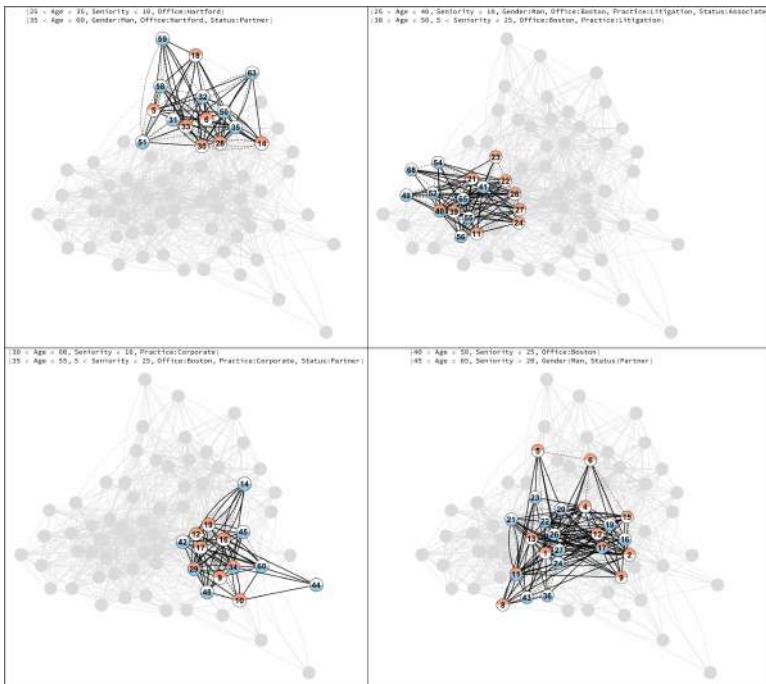


Fig. 4. Core subgraphs of the $S_{0.9}$ bi-patterns selected from P_2 . The top-left subgraph represents a group of lawyers from Hartford: young lawyers with seniority at most 10 (in blue) ask for advice from older, male partners (in red). The top-right subgraph represents a group of lawyers from Boston who practise Litigation Law: young male associates with seniority at most 10 years (in blue) ask for advice from lawyers with seniority higher than 5 years (in red). The bottom-left subgraph represents a group of lawyers who practise corporate law: lawyers with seniority at most 10 (in blue) ask for advice from partners from Boston with seniority at least 5 years (in red). In the bottom-right subgraph middle-aged lawyers ask for advice from older male partners from Boston with seniority in the firm beyond 20 years. Within a group some lawyers ask for advice as well as they are asked for and are then represented as blue and red vertices. For instance associates 29 and 34 have both out (blue) and in (red) roles in the bottom left subgraph.

The four bi-patterns found in $S_{0.9}$ and displayed Fig. 4, represent subgroups in which lawyers with low seniority ask for advice from lawyers with higher seniority. The bi-patterns differ in the location (Hartford or Boston) and law practised (Litigation or Corporate).

5.4 Comparison with KRIMP on Standard Closed Itemset Mining

KRIMP relies on the *Minimum Description Length* (or MDL) principle: a pattern is considered as *redundant* with respect to a pattern subset when it does not reduce the cost associated to representing the data using this pattern subset. KRIMP produces a set of codes representing patterns representative of the whole pattern set. Similarly to $g\beta$ it uses a particular ordering of the pattern set. As it is, KRIMP cannot be applied to core closed patterns or bi-patterns compression and it is still an open problem to find a proper way to apply MDL for core closed patterns and bi-patterns.

We compared $g\beta$ to KRIMP [17] on the pattern subset selection on standard closed pattern mining. We consider four datasets previously investigated with KRIMP (see Table 6) and processed as follows: first from the set of closed patterns we remove the empty set (if present), the remaining subset P is then processed using both KRIMP and $g\beta$.

On one hand KRIMP computes K , the representative itemsets of the code table that compresses the pattern set P . On the other hand P is ordered according to the order used by KRIMP (first by decreasing size of support sets then by decreasing itemsets sizes). We then search for a β threshold such that $g\beta$ outputs a pattern set S_P of size close to the size of the set K extracted from the KRIMP results. Finally we compute the pattern subset S_K obtained by $g\beta$ selection on K , with same β value. Table 6 displays various measures computed on these results. In these datasets β values are high indicating high redundancy in the pattern set. Pattern subsets distances $d_s(S, K)$ varies from 0.49 to 0.85 indicating a variable agreement between the two methods. Also note that applying $g\beta$ to K still leads to a reduction of the pattern set (about from one third to half of K is selected).

Table 6. $g\beta$ and KRIMP selections from various datasets: the columns respectively represent the number $\#O$ of objects in the database, the number $\#I$ of items, the number $\#P$ of closed patterns submitted to $g\beta$ and KRIMP, the size $\#K$ of the KRIMP output, the distance threshold β used for greedy selection, the number $\#S$ of patterns selected by $g\beta$ on P , the distance $d_s(S, K)$ between pattern set S and pattern set K , the covering cov(S) of the object set by S , the covering cov(K) of the object set by K , and the size $\#S_K$ of the pattern set S_K selected by $g\beta$ from K with the β value associated to the dataset.

Name	#O	#I	#P	#K	β	#S	$d_s(S, K)$	cov(S)	cov(K)	# S_K
led7	3200	24	7037	176	0.9	190	0.73	1	0.97	104
breast	699	16	641	40	0.7	41	0.49	1	1	22
pima	768	38	3203	88	0.9	100	0.62	1	0.94	35
chess (kr-k)	28056	58	180864	1733	0.995	2210	0.85	1	0.94	891

6 Conclusion

Pattern set selection is an important issue in data mining, addressed in particular using compression techniques as in the KRIMP algorithm [17]. The general $g\beta$ subset selection method introduced here is an alternative and complementary simple way to perform pattern set selection. Its main parameter is a distance threshold, which allows working at finer to coarser levels. Distances criteria were also used in a different way in [3] for non overlapping community detection. In the latter work the authors use approximation techniques, such as bottom-k sketches for approximating Jaccard distances. Such efficient approximations should be mandatory for scalability of $g\beta$, in particular when support sets are vertex subsets of large networks. In this work we have applied $g\beta$ to attributed graph single pattern and bi-pattern mining which relies on core definitions. In this context the prior individual pattern selection step according to interestingness measures is of particular importance. For instance, focussing on high local modularity vertex subsets expresses our interest in subgraphs with less external links than expected, while reducing pattern subgraphs to pattern core subgraphs comes with deviation to core size as a natural unexpectedness measure. Other interestingness measures related to unexpectedness may be considered, as it was the case in a recent work about multi-relational pattern mining [16]. Furthermore, when considering bi-patterns we may be interested in those with low homogeneity, i.e. those whose subgraphs display links between vertices that have different attribute values, therefore focussing in network heterophily. Note that the preliminary ordering of patterns using a decreasing interestingness order is an important step of the $g\beta$ selection process, and from this point of view, the greedy behaviour of the algorithm appears as a strength rather than as a weakness. In our experiments we have applied the methodology to two attributed networks of different nature and size with interesting results, in particular regarding quantities, as the redundancy ratio, which seems stable through a wide range of β values. However, our general guess is that more work is needed to relate the attributed network structure to the mining parameters (which core definitions and parameters to consider?) as well as to the pattern set selection β parameter. Overall our intended contribution is to add to attributed networks investigation tools allowing to avoid redundancy and to focus on relevant criteria.

References

1. Atzmueller, M., Doerfel, S., Mitzlaff, F.: Description-oriented community detection using exhaustive subgroup discovery. *Inf. Sci.* **329**, 965–984 (2016)
2. Atzmueller, M., Soldano, H., Santini, G., Bouthinon, D.: MinerLSD: efficient local pattern mining on attributed graphs. In: Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE Press (2018)
3. Baroni, A., Conte, A., Patrignani, M., Ruggieri, S.: Efficiently clustering very large attributed graphs. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 369–376. ASONAM 2017. ACM, New York (2017)

4. Batagelj, V., Zaversnik, M.: Fast algorithms for determining (generalized) core groups in social networks. *Adv. Data Anal. Classif.* **5**(2), 129–145 (2011)
5. Brandstädt, A.: On robust algorithms for the maximum weight stable set problem. In: Freivalds, R. (ed.) *Fundamentals of Computation Theory*, pp. 445–458. Springer, Heidelberg (2001)
6. Galbrun, E., Gionis, A., Tatti, N.: Overlapping community detection in labeled graphs. *Data Min. Knowl. Discov.* **28**(5–6), 1586–1610 (2014)
7. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *J. ACM (JACM)* **46**(5), 604–632 (1999)
8. Lazega, E.: *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press, Oxford (2001)
9. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* **27**(1), 415–444 (2001)
10. Newman, M.E.J.: Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**(23), 8577–8582 (2006)
11. Seidman, S.B.: Network structure and minimum degree. *Soc. Networks* **5**, 269–287 (1983)
12. Silva, A., Meira Jr., W., Zaki, M.J.: Mining attribute-structure correlated patterns in large attributed graphs. *Proc. VLDB Endow.* **5**(5), 466–477 (2012)
13. Soldano, H., Santini, G.: Graph abstraction for closed pattern mining in attributed networks. In: Schaub, T., Friedrich, G., O’Sullivan, B. (eds.) *European Conference in Artificial Intelligence (ECAI). Frontiers in Artificial Intelligence and Applications*, vol. 263, pp. 849–854. IOS Press (2014)
14. Soldano, H., Santini, G., Bouthinon, D., Bary, S., Lazega, E.: Bi-pattern mining of attributed networks. *Appl. Network Sci.* **4**(1), 37 (2019)
15. Soldano, H., Santini, G., Bouthinon, D., Lazega, E.: Hub-authority cores and attributed directed network mining. In: *International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1120–1127. IEEE Computer Society, Boston, 7–9 November 2017
16. Spyropoulou, E., Bie, T.D., Boley, M.: Interesting pattern mining in multi-relational data. *Data Min. Knowl. Discov.* **28**(3), 808–849 (2014)
17. Vreeken, J., van Leeuwen, M., Siebes, A.: Krimp: mining itemsets that compress. *Data Min. Knowl. Disc.* **23**(1), 169–214 (2011)
18. Yang, J., McAuley, J., Leskovec, J.: Community detection in networks with node attributes. In: *2013 IEEE 13th International Conference on Data Mining*, pp. 1151–1156, December 2013



On the Relation of Edit Behavior, Link Structure, and Article Quality on Wikipedia

Thorsten Ruprechter^(✉), Tiago Santos, and Denis Helic

Graz University of Technology, Graz, Austria

{ruprechter, dhelic}@tugraz.at, tsantos@iicm.edu

Abstract. When editing articles on Wikipedia, arguments between editors frequently occur. These conflicts occasionally lead to destructive behavior and diminish article quality. Currently, the relation between editing behavior, link structure, and article quality is not well-understood in our community, notwithstanding that this relation may facilitate editing processes and article quality on Wikipedia. To shed light on this complex relation, we classify edits for 13,045 articles and perform an in-depth analysis of a 4,800 article subsample. Additionally, we build a network of wikilinks (internal Wikipedia hyperlinks) between articles. Using this data, we compute parsimonious metrics to quantify editing and linking behavior. Our analysis unveils that controversial articles differ considerably from others for almost all metrics, while slight trends are also detectable for higher-quality articles. With our work, we assist online collaboration communities, especially Wikipedia, in long-term improvement of article quality by identifying deviant behavior via simple sequence-based edit and network-based article metrics.

Keywords: Wikipedia · Edit behavior · Link structure · Article quality · Edit wars · Semantic edit types · Multi-label classification · Network analysis

1 Introduction

Editing behavior on Wikipedia has been a widely studied subject in previous research [3, 13, 19, 33, 36]. In particular, past studies investigated the misbehavior on Wikipedia including, among others, vandalism [1, 20], conflict or controversy [2, 35], and so-called edit wars [30, 34]. Edit wars are a type of behavior in which two or more opposing editors (or editor groups) override each others content due to the differences in their opinions on a given subject. Prominent examples of such behavior include the Wikipedia pages on Evolution¹ or Nikola Tesla². Studying edit wars and controversial articles on Wikipedia has been and still is

¹ <https://en.wikipedia.org/wiki/Evolution>.

² https://en.wikipedia.org/wiki/Nikola_Tesla.

an important endeavor from a scientific perspective, as several studies show the connection between article quality and editing behavior on Wikipedia [11, 27]. As for practical purposes, this research may help complement other prevalent conflict prediction methods, such as those utilizing mutual reverts [14, 34].

In this paper, we extend previous work on Wikipedia editing behavior with an investigation on the relation between editing and linking across article quality categories—a relation which, to the best of our knowledge, has not been analyzed in depth previously. We are particularly interested in answering the following research questions: (i) How can we characterize editing behavior on Wikipedia? (ii) Is there a relation between editing behavior, article quality, and wikilink (internal Wikipedia hyperlinks) network topology on Wikipedia? (iii) If such a relation exists, how strong is it?

To answer these research questions, we: (i) classify edit actions for 13,045 Wikipedia articles using a state-of-the-art machine learning approach, (ii) compute relative frequencies of edit actions and build first-order Markov chains from edit sequences for each of our Wikipedia articles, (iii) perform statistical significance tests on these results to characterize differences in editing behavior, (iv) extract a wikilink network for our selection of articles, and (v) compute and compare standard network metrics of selected articles given the article’s quality.

We find that there is significant difference in editing behavior between edit wars or similar controversially edited content and higher quality articles, which corroborates previous results on edit wars [30, 36]. Adding to those studies, we find that editors of conflicted articles significantly more often make meaning-changing edits while performing less formatting and potentially less link-editing operations, thus rendering factual content especially contested. Consequentially, edit war articles are clear outliers in several standard network measures. These articles in particular have, on average, significantly higher in-degree, out-degree, PageRank, and k -core in contrast to a lower reciprocity and clustering coefficient. In addition, the distribution of those quantities differs substantially from other quality categories in both moments and shape, which takes on non-typical forms for such a Web-based editing processes.

With our work, we provide practical contributions for Wikipedia’s community. First, with our observation of a clear trend in semantic edit intentions and the occurring disparities in the underlying link structure of articles from different quality categories, we open up possibilities to use link metrics as automated indicators of controversy. Moreover, our results may inform further development of existing Wikipedia content-assessment tools such as Huggle³, Contropedia [2], and ORES [15]. Furthermore, through our findings for semantic edit labels, we facilitate new solutions for problems such as editor role identification in Wikipedia or other online collaboration systems [32]. Our base methods can be readily applied to similar domains, and we make our code available on GitHub⁴.

³ <https://en.wikipedia.org/wiki/Wikipedia:Huggle>.

⁴ <https://github.com/ruptho/editlinkquality-wikipedia>.

2 Related Work

Controversy and conflicts on Wikipedia can take on different forms. In order to quantify conflicts in articles, authors previously introduced metrics to measure controversy [2, 13, 35]. Although conflict and controversy do not inherently resemble destructive behavior [19], the resulting acts of vandalism or edit wars can. Automatically detecting vandalistic contributions and performing counter-vandalism actions on articles presents a well-developed research area [1, 15, 20]. Similarly, prediction and prevention of conflicts as well as edit wars were researched broadly in the past [14, 30, 34, 36].

Multiple edit label taxonomies were formerly proposed, considering both semantic and syntactic changes. Early works differentiate edits which either change (“Text-Base”) or preserve (“Surface”) the meaning of texts [12]. Later, more sophisticated taxonomies were introduced, adapting to the context of Wikipedia [7, 32]. In 2016, the Wikimedia foundation deployed an experimental three-level taxonomy for article edits⁵, which is structured into 14 semantic intentions, 18 syntactic elements, and three editor actions [33]. For our work, we adapt the 14 semantic labels specified by this recent taxonomy for our classification method.

To this day, many diverse network analysis approaches have been carried out for Wikipedia. Firstly, researchers frequently analyzed collaboration and social structure in editor networks, regularly also deriving information about article quality [3, 8, 21, 22, 24]. As far as hyperlink networks between article pages are concerned, studies often utilized such link structures to leverage semantic [23], topical [6], or categorical [29] information. Regarding link success, authors found that users seem to frequently choose links leading to target Wikipedia articles less prominent than the source article, or to one with a similar topic [9]. As for centrality, featured articles are “more central” than others in specific Wikipedia language editions, depicted by their lower clustering coefficient [17]. Moreover, other authors suggested that out-degree correlates with quality, in-degree with popularity, and PageRank with importance for the Portuguese Wikipedia—although correlation was generally weaker for quality and importance, while being moderate for popularity [16]. In 2009, a study provided information about general network metrics for Wikipedia, claiming that median values for article in- and out-degree were 4 and 12, while average degree was 20.63 [18]. On a completely different note, article pages not only contain text links but also links generated by templates. To solely focus on user-created links, datasets such as *WikiLinkGraphs* [5] were proposed, which only include wikilinks in article texts and exclude automatically generated links. In this work, we extend the previously mentioned studies by combining semantic editing behavior, wikilink network structure, and article quality.

⁵ https://en.wikipedia.org/wiki/Wikipedia:Labels/Edit_types/Taxonomy.

3 Materials and Methods

3.1 Background and Preliminaries

Revisions, Edits, Wikilinks. Wikipedia articles are products of a vast number of revisions, performed by either registered or unregistered editors [26]. A single revision may contain multiple edits to an article, such as insertion, modification, or deletion of content. Through such edits, editors can create hyperlinks to both external pages as well as other Wikipedia articles (i.e., wikilinks). While wikilinks add useful context for readers, they also produce structural information by means of the emerging article network. Therefore, the resulting link structure enables readers to follow the flow of topical or categorical information on Wikipedia.

Content Assessment on Wikipedia. Wikipedia establishes article ratings using a well-defined content assessment system⁶. This system allows for assessment of articles according to two factors: quality and importance. For this work, we focus on article quality as a distinguishing factor. Wikipedia defines concise guidelines for rating article quality. Quality ratings range from highest to lowest, including: Featured (FA), A-class, Good (GA), B-Class, C-class, Start, and Stub. In addition, Featured List and List pages exist, which must also follow Wikipedia content policies. The highest-quality content on Wikipedia, such as GA and FA, must be approved through a specific review process, while articles in other classes can have multiple ratings. These different quality ratings are individually assigned by users of separate WikiProjects⁷. The final assessment for an article is determined by its best rating over all WikiProjects. As of June 2019, out of the 6,377,691 pages existing in the article namespace (“ns0”) of English Wikipedia, there are 6,705 FA, 1,874 A, 32,759 GA, 127,107 B, 308,621 C, and 5,900,625 otherwise classified articles.

3.2 Dataset

We utilize the Wikimedia API *RevScoring*⁸ to process revision histories for 2,542 FA-, 873 A-, 6,339 GA-, 1,189 B- and 1,161 C-category articles in the English Wikipedia. The Start and Stub categories are ignored, due to their low quality and in many cases short revision history. We also omit Featured List and List pages, because of their different content structure in comparison to regular articles. In addition to these officially rated pages, we collect revision histories of articles deemed to be especially controversial or contain edit wars. For this, we gather a total of 941 articles from the following sources: 401 of the “most conflicted” articles [13], 94 of the “most controversial” articles [35], 396 articles from Wikipedia’s *List of Controversial Issues*⁹, and 50 articles from Wikipedia’s list of *Lamest Edit Wars*¹⁰. We name this article category *CombinedWars* (CW).

⁶ https://en.wikipedia.org/wiki/Wikipedia:Content_assessment.

⁷ <https://en.wikipedia.org/wiki/Wikipedia:WikiProject>.

⁸ <https://github.com/wikimedia/revscoring>.

⁹ https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues.

¹⁰ https://en.wikipedia.org/wiki/Wikipedia:Lamest_edit_wars.

If there is an overlap between CW and other quality categories, articles are assigned to the CW class. For our analysis, we randomly sample and investigate 800 articles having at least 50 revisions for each of the six categories—a total of 4,800 of the 13,045 retrieved articles.

3.3 Labeling Edit Actions on Wikipedia

Edit Actions. To make automatic labeling of article edits more robust, we derive three super-labels from the 14-label taxonomy deployed by Yang et al. [32]: Content, Format, and WikiContext. First, Content captures all edits aiming towards modifying actual information on the article page. It combines fact-updates, simplification, elaboration, clarification, increasing verifiability, and establishing a neutral point of view. Secondly, Format describes all editing actions not changing the meaning of texts, facts, or contained information. This includes refactoring, copy-editing, manipulating wikilinks and wiki markup, as well as link disambiguation. Lastly, WikiContext captures all wiki-specific interactions such as modifying processing tags, vandalism, counter-vandalism, or other intentions.

Classification of Edit Actions. Alongside their taxonomy, Yang et al. [33] released a manually labeled dataset of 5,777 multi-labeled revisions. We transform this dataset to fit our three-category taxonomy through aggregation of edit labels. Furthermore, we collect additional samples for the WikiContext category via retrieving edits performed by the anti-vandalism bot ClueBotNG¹¹. The bot’s edits are categorized as examples of counter-vandalism, while their parent revisions are classified as vandalism (0.1% false positive rate). By combining these samples with the initial dataset, we accumulate 6,670 multi-label revisions which contain 3,497 Format, 2,346 Content, and 1,641 WikiContext edits.

We utilize most of the feature framework published parallel to the 14-label dataset [33] to retrieve and build the feature set for our multi-label classifier. Per revision, we use *RevScoring* to retrieve 163 base features. This feature base mostly consists of text features such as differences in words, punctuation, and numbers as well as trivial editor information. Further processing these revision features produces 207 final features for classification, extending the initial feature set with, for example, features about revision comments, stemmed text, and similarity to the parent revision [33]. Finally, we utilize scikit-learn¹² to train a multi-label Random Forest classifier on the training set. Through a 80–20 train/test split and grid search with 10-fold validation, we find a feasible configuration (weighted F1-score of 0.8153). The parameter setting for this configuration is: 750 estimators, a maximum depth of 25, and 50% of features considered for finding the best split.

¹¹ https://en.wikipedia.org/wiki/User:ClueBot_NG.

¹² <https://scikit-learn.org>.

3.4 Modeling Edit Action Sequences

Relative Edit Label Frequency. We compute relative edit label frequency for each category by macro-averaging the relative frequencies of their articles. Accordingly, we first calculate relative frequency of the automatically created edit labels for all articles. After that, averaging the article values in each quality category produces the per-category macro-averages.

Edit Label Transition Probability. To investigate label transitions, we build first-order Markov chains out of the automatically labeled article revision histories. Similar to the computation of relative label frequencies explained above, we accumulate macro-averaged transition probabilities for all quality categories. First, transition probabilities are computed from edit label sequences in article revision histories. Subsequently, we average article results for each quality category to generate per-category label transition probabilities.

Characterizing Differences in Categories. We perform pair-wise permutation tests [4, 31] for quality categories to assess statistical significance of differences between categories. These tests compare the distance of actual category means to that of means where articles were randomly exchanged between categories. The null hypothesis H_0 for our test states that values for article subsets drawn from different categories stem from the same probability distribution.

3.5 Network of Wikilinks

To compare article network properties, we generate a wikilink graph by employing the framework used to create the *WikiLinkGraphs* dataset [5]. In addition, we execute our own post-processing pipeline, which removes nodes with a single outgoing link (i.e., redirects) and resolves leftover duplicate article titles in the original dataset. We then use the Python package *NetworkX*¹³ to compute an article graph for June 2019 solely from wikilinks in article texts. In this dataset, each article is a node and wikilinks between articles are edges. When navigating a Web network, users typically browse and search links, and we therefore compute corresponding network metrics [10]. These include empirical complementary cumulative distribution functions (CCDF) for in-degree, out-degree, PageRank, reciprocity, clustering coefficient, and k -core of the 4,800 articles in our examined subset. We could not find links for 35 articles (20 A, 2 B, 1 C, 12 CW) in the computed *WikiLinkGraphs* dataset, due to discrepancies we yet have to identify.

4 Results

4.1 Relative Edit Label Frequency

We show results for category-wise relative label frequencies in Fig. 1a. Although multi-label classification was applied, we exclude results for any label combination besides Content and Format, due to their extremely low relative frequency

¹³ <https://networkx.github.io>.

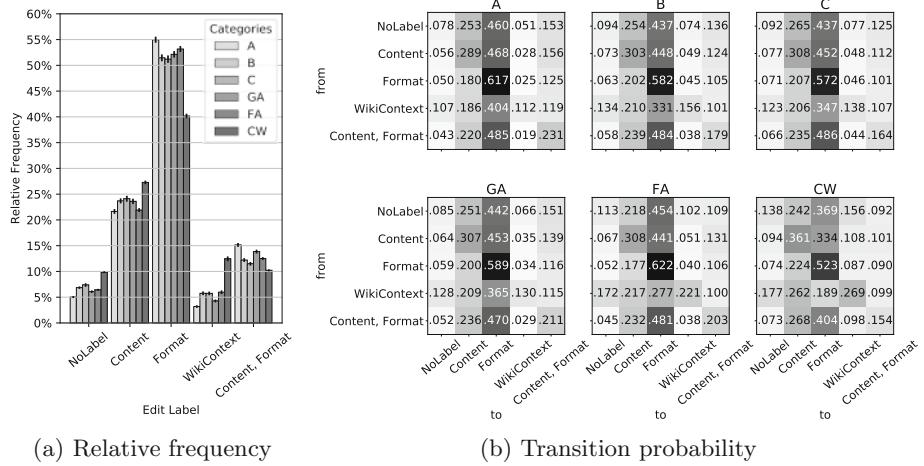


Fig. 1. Relative frequency and transition probability for edit labels. In Fig. 1a, we visualize relative label frequencies by article category (bootstrapped 95% confidence intervals), while we show transition probabilities between edit labels per article quality in Fig. 1b. Overall, CW results are considerably different from other categories.

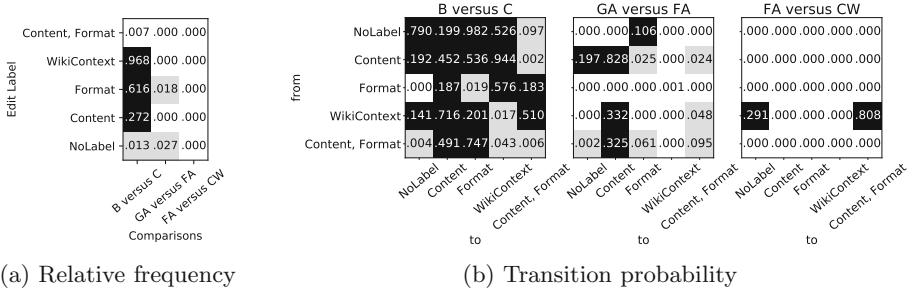
(<0.0004). Furthermore, revisions exist where our classifier could either not assign a definite label, revisions were deleted, or other inconsistencies occurred. We name this label class NoLabel.

We find several substantial disparities between various article quality categories. Relative edit label frequency for CW deviates considerably from all other categories. In particular, revisions for articles in the CW category, on average, contain 40.2% Format (which includes wikilink manipulations), 27.3% Content, 12.4% WikiContext, 10.2% Content and Format, and 9.9% NoLabel edits. On the contrary, in other categories Format actions constitute more than 51% of all edits. For formatting actions, A (54.9%) and FA (53.2%) are the overall highest-scoring categories. In addition, CW revisions are more than twice as likely to contain WikiContext edits than all other classes (12.4% versus 6% or less). On top of that, we detect a percentage increase of 18.6% for Content and 55.1% for NoLabel in comparison to other quality categories. Overall, we conclude that relative edit label frequency for CW differs considerably from all other categories. At the same time, we observe only minor differences between categories B and C as well as GA and FA.

Differences between CW and other categories are statistically significant, as our permutation tests for comparisons involving CW produce p-values <0.01. Additionally, tests confirm partial similarity of other categories (e.g., B and C). We show a selection of permutation test results for relative frequency in Fig. 2a.

4.2 Label Transition Probabilities

We present label transition probabilities in Fig. 1b, which unveils that probabilities for the CW category are considerably different from others. Particularly, CW shows a fairly low probability of consecutive formatting actions (0.523), especially in comparison to high-quality categories such as FA (0.622), A (0.617), and GA (0.589). Furthermore, CW articles more strongly lean towards sequences containing Content, NoLabel, and WikiContext, making their revision histories substantially dissimilar from other categories. In addition, results for B and C suggest similar transition probabilities for nearly all sequences. Likewise, GA and FA also share similar results, once again slightly differing from the rest.



(a) Relative frequency

(b) Transition probability

Fig. 2. Results of permutation tests comparing selected quality categories. In Fig. 2a, we visualize resulting p-values of permutation tests for relative edit label frequency of selected quality categories, while we show values for label transition probabilities in Fig. 2b. White boxes highlight statistically significant results for $p < 0.01$ in Fig. 2a and $p < 0.002$ in Fig. 2b (after Bonferroni correction). Results in black represent non-significant differences, which we operationalize as $p > 0.1$. Gray boxes signal p-values of $0.01 \leq p \leq 0.1$ in Fig. 2a and $0.002 \leq p \leq 0.1$ in Fig. 2b. Results for FA versus CW in Figs. 2a and 2b demonstrate the dissimilarity of CW to other categories.

Permutation test results principally confirm statistical significance of CW's dissimilarity. Alternatively, non-significant results for comparisons between B and C as well as GA and FA suggest similarities of these categories. In Fig. 2b, we show permutation test results for this category selection.

4.3 Network of Wikilinks

As we visualize in Fig. 3, CW exhibits fairly different CCDFs for multiple standard network metrics. Our results indicate that mean in-degree (deg^-), out-degree (deg^+), PageRank (PR), and k -core are substantially higher, while reciprocity (r) and clustering coefficient (C) are lower for CW than for other categories.

First, non-controversial categories rarely have a deg^- of more than 6,000, while over 20% of CW articles do (Fig. 3a). Secondly, we also observe an on average higher deg^+ for CW, although the effect is not as extreme as for deg^- (Fig. 3b).

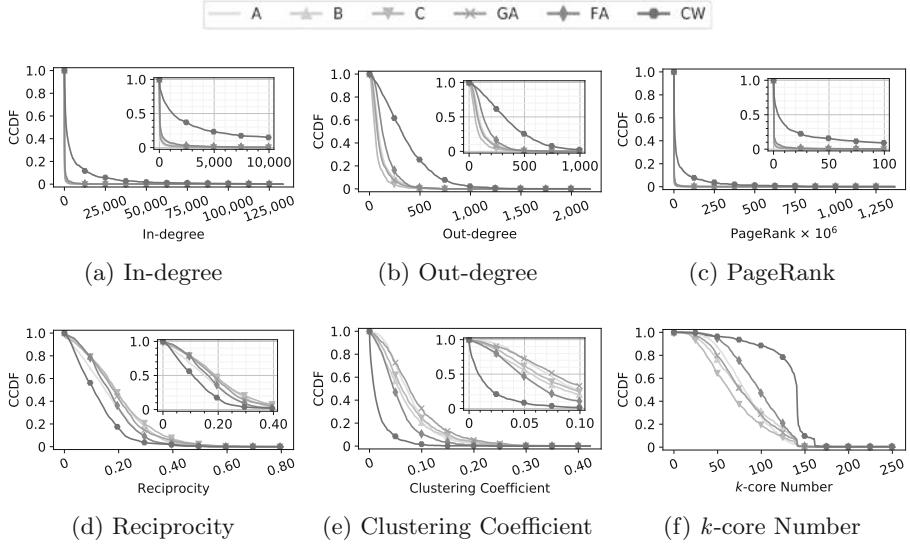


Fig. 3. CCDFs of network metrics across multiple quality categories, with insets highlighting relevant chart areas. In all CCDFs, we detect substantial differences for CW. We find that, on average, in-degree (Fig. 3a), out-degree (Fig. 3b), and PageRank (Fig. 3c) are considerably higher, while reciprocity (Fig. 3d) and clustering coefficient (Fig. 3e) are seemingly lower for CW than for other categories. Additionally, we attribute the higher k -core numbers (Fig. 3f) to their correlation with degree [28].

Thirdly, the majority of non-controversial articles holds extremely low PR , leading to hardly any articles reaching a value of 0.00006, while well over 10% of those in CW have a higher PR than that (Fig. 3c). Next, even though mean r is lower for CW than for other categories, it exhibits the smallest difference out of all metrics (Figs. 3d). Furthermore, over 97.5% of CW articles have a C lower than 0.1, whereas other categories seem to reach considerably higher values (Fig. 3e). Lastly, computed k -score numbers indicate that a large number of CW articles (205) is grouped in a 141-cluster, while other categories are more evenly scattered (Fig. 3f). Altogether, we conclude that CW articles show substantially different results for the considered network analysis metrics.

On a separate note, results for FA also tend to somewhat differ from C, B, GA, and A for specific metrics such as C or deg^+ . However, the observable contrast is by no means as definite as for CW.

5 Discussion

Edit Label Results. Edit actions in CW articles follow significantly different patterns than other quality categories. More precisely, the increased relative frequencies and transition probabilities for Content, WikiContext, or NoLabel indicate that editors consistently rewrite each others content, suggesting editor

disputes in CW. NoLabel edits generally signal that revisions contain untypical content, such as ASCII art¹⁴, or that they were removed by administrators. Therefore, we argue that the CW category's high frequency of NoLabel edits is a by-product of increased vandalism. Altogether, our results suggest that content is significantly more disputed in CW articles, triggering other destructive behavior. On the contrary, we detect a substantially lower amount of formatting actions in CW, potentially indicating less link-editing operations.

On a separate note, the considerably higher transition probabilities for consecutive Format labels in GA, A, and FA may stem from these categories depicting some of the highest-quality Wikipedia content, thus needing more polishing. Editors commonly improve format of articles when they are considered for promotion to a higher quality category, which further supports this assumption. Such behavioral information could be leveraged for editor role identification or similar applications benefiting from semantic edit labels.

Network Metrics. We observe a lower clustering coefficient for CW articles, which could be the consequence of such content being relevant to multiple diverse, closer connected article subgroups. For example, controversial articles such as “United States”, “Vladimir Putin”, and “World War II” connect different topical categories on Wikipedia, acting as a “waypoint” between communities. Considering previous findings about edit wars, this might be a peculiarity of English Wikipedia, where broader topics are more contested [35]. Apparently, the English version merges contributions of people with differing origin and background, thus generating conflict and edit wars—an effect which is not as prevalent for non-English Wikipedia. However, this effect may also be explained by the general negative correlation between clustering coefficient and degree [25].

Furthermore, conflicted articles exhibit, on average, higher in-degree and PageRank, signaling frequent referral from other (or more prominent) articles. Consequently, such articles “attract more attention” than others. Increased public exposure might lead to a boost in popularity and, as a result, controversy. A rapid increase in in-links might therefore signal an article gaining traction due to a recent event or current news. This could possibly be interpreted as a “warning signal” for administrators. It might be feasible to start monitoring or even semi-protect¹⁵ such content to potentially prevent edit wars.

Limitations. Firstly, instead of our simplistic approach, a different multi-label classification method might be more suitable for our crude taxonomy. Some label combinations are, although theoretically possible, highly unlikely in practice. Future studies should consider using approaches such as converting the task to a multi-class problem or implementing hierarchical multi-label classification.

Moreover, one could argue that 800 articles per category (4,800 overall) are a non-representative sample for Wikipedia. However, we sampled our article subset randomly from a more comprehensive dataset and only accounted for articles

¹⁴ https://en.wiktionary.org/wiki/ASCII_art.

¹⁵ https://en.wikipedia.org/wiki/Wikipedia:Protection_policy#semi.

with more than 50 revisions. Above all, our classification still quantitatively outmatches similar studies which apply semantic Wikipedia edit labeling [7,33].

On top of that, non-scaling network metrics with article length could skew results, especially for out-degree. However, we desist from such a scaling, since we mostly focused our elaborations on in-degree, PageRank, and clustering coefficient, which are not as dependent on article length.

6 Conclusions

In this work, we showed that edit wars and controversial articles significantly differ from articles in other quality categories, both in regard to editing as well as linking behavior. Firstly, our findings for relative label frequencies and label transition probabilities indicated substantial editor disputes over content in controversial articles, making editing behavior vastly different from regular articles. Overall, we demonstrated feasibility of deriving semantic editor behavior from revision histories, which may prove helpful to existing issues such as editor role identification on Wikipedia. Furthermore, we proposed application of network metrics for conflict detection, an alternative to existing algorithms such as those utilizing mutual reverts [14,34]. Besides, we elaborated the effect that controversial issues are not only more frequently referred to from other articles, but may also act as a connector between topical subgroups. As opposed to the English Wikipedia, the exact nature of controversy in general subjects in other language editions is still an open research question, thus making a similar analysis of non-English Wikipedia an interesting avenue for future research.

Acknowledgments. Tiago Santos is a recipient of a DOC Fellowship of the Austrian Academy of Sciences at the Institute of Interactive Systems and Data Science of the Graz University of Technology.

References

1. Adler, B.T., De Alfaro, L., Mola-Velasco, S.M., Rosso, P., West, A.G.: Wikipedia vandalism detection: combining natural language, metadata, and reputation features. In: CICLing, pp. 277–288. Springer (2011)
2. Borrà, E., Weltvrede, E., Ciuccarelli, P., Kaltenbrunner, A., Laniado, D., Magni, G., Mauri, M., Rogers, R., Venturini, T.: Societal Controversies in Wikipedia articles. In: SIGCHI, pp. 193–196 (2015)
3. Brandes, U., Kenis, P., Lerner, J., Van Raaij, D.: Network analysis of collaboration structure in Wikipedia. In: WWW, pp. 731–740. ACM (2009)
4. Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., Gilbert, E.: You Can't stay here: the efficacy of Reddit's 2015 ban examined through hate speech. HCI 1(CSCW), 31:1–31:22 (2017)
5. Consonni, C., Laniado, D., Montresor, A.: WikiLinkGraphs: a complete, longitudinal and multi-language dataset of the Wikipedia link networks. In: ICWSM, vol. 13, pp. 598–607 (2019)
6. Coursey, K., Mihalcea, R.: Topic identification using Wikipedia graph centrality. In: NAACL HLT, pp. 117–120 (2009)

7. Daxenberger, J., Gurevych, I.: A corpus-based study of edit categories in featured and non-featured wikipedia articles. In: COLING, pp. 711–726 (2012)
8. De La Robertie, B., Pitarch, Y., Teste, O.: Measuring article quality in Wikipedia using the collaboration network. In: ASONAM, pp. 464–471 (2015)
9. Dimitrov, D., Lemmerich, F., Singer, P., Strohmaier, M.: What Makes a Link Successful on Wikipedia? In: WWW, pp. 917–926 (2017)
10. Dimitrov, D., Singer, P., Helic, D., Strohmaier, M.: The role of structural information for designing navigational user interfaces. In: HT, pp. 59–68. ACM (2015)
11. Editorial: Britannica attacks. *Nature* **440**(582) (2006)
12. Faigley, L., Witte, S.: Analyzing revision. *College Compos. Commun.* **32**(4), 400–414 (1981)
13. Flöck, F., Erdogan, K., Acosta, M.: TokTrack: a complete token provenance and change tracking dataset for the English Wikipedia. In: ICWSM, pp. 408–417 (2017)
14. Gandica, Y., dos Aidos, F.S., Carvalho, J.: The dynamic nature of conflict in Wikipedia. *EPL* **108**(1), 18003 (2014)
15. Halfaker, A., Geiger, R.S., Morgan, J.T., Sarabadani, A., Wight, A.: ORES: facilitating remediation of Wikipedia's socio-technical problems (2018)
16. Hanada, R., Cristo, M., Pimentel, M.D.G.C.: How do metrics of link analysis correlate to quality, relevance and popularity in Wikipedia? In: WebMedia, pp. 105–112 (2013)
17. Ingawale, M., Dutta, A., Roy, R., Seetharaman, P.: Network analysis of user generated content quality in Wikipedia. *Online Inf. Rev.* **37**(4), 602–619 (2013)
18. Kamps, J., Koolen, M.: Is Wikipedia link structure different? In: WSDM, pp. 232–241 (2009)
19. Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, she says: conflict and coordination in Wikipedia. In: SIGCHI, pp. 453–462 (2007)
20. Kumar, S., Spezzano, F., Subrahmanian, V.: VEWS: a Wikipedia vandal early warning system. In: SIGKDD, pp. 607–616 (2015)
21. Li, X., Tang, J., Wang, T., Luo, Z., De Rijke, M.: Automatically assessing wikipedia article quality by exploiting article-editor networks. In: European Conference on Information Retrieval, pp. 574–580. Springer (2015)
22. Liu, J., Ram, S.: Using big data and network analysis to understand Wikipedia article quality. *Data Knowl. Eng.* **115**, 80–93 (2018)
23. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In: AAAI (2008)
24. Platt, E.L., Romero, D.M.: Network structure, efficiency, and performance in WikiProjects. In: ICWSM, pp. 251–260 (2018)
25. Ravasz, E., Barabási, A.L.: Hierarchical organization in complex networks. *Phys. Rev. E* **67**(2), 026112 (2003)
26. Sage Ross: Editing Wikipedia, a print guide for new contributors (2014). <https://w.wiki/86W>. Accessed 09 Apr 2019
27. Samoilenco, A., Lemmerich, F., Zens, M., Jadidi, M., Génois, M., Strohmaier, M.: (Don't) mention the war: a comparison of Wikipedia and britannica articles on national histories. In: WWW, pp. 843–852 (2018)
28. Shin, K., Eliassi-Rad, T., Faloutsos, C.: CoreScope: graph mining using k-core analysis - patterns, anomalies and algorithms. In: ICDM, pp. 469–478 (2016)
29. Sucheki, K., Salah, A.A.A., Gao, C., Scharnhorst, A.: Evolution of Wikipedia's category structure. *Adv. Complex Syst.* **15**, 1250068 (2012)
30. Sumi, R., Yasseri, T., et al.: Edit wars in Wikipedia. In: PASSAT/SocialCom, pp. 724–727 (2011)

31. Vautard, R., Mo, K.C., Ghil, M.: Statistical significance test for transition matrices of atmospheric Markov chains. *J. Atmos. Sci.* **47**(15), 1926–1931 (1990)
32. Yang, D., Halfaker, A., Kraut, R., Hovy, E.: Edit categories and editor role identification in Wikipedia. In: LREC, pp. 1295–1299 (2016)
33. Yang, D., Halfaker, A., Kraut, R., Hovy, E.: Identifying semantic edit intentions from revisions in Wikipedia. In: EMNLP, pp. 2000–2010 (2017)
34. Yasseri, T., Kertész, J.: Value production in a collaborative environment. *J. Stat. Phys.* **151**(3), 414–439 (2013)
35. Yasseri, T., Spoerri, A., Graham, M., Kertész, J.: The most controversial topics in Wikipedia. *Global Wikipedia* **25** (2014)
36. Yasseri, T., Sumi, R., Rung, A., Kornai, A., Kertész, J.: Dynamics of conflicts in Wikipedia. *PLoS ONE* **7**(6), 1–12 (2012)



Establish the Expected Number of Injective Motifs on Unlabeled Graphs Through Analytical Models

Emanuele Martorana¹, Giovanni Micale², Alfredo Ferro²,
and Alfredo Pulvirenti^{2(✉)}

¹ Department of Physics and Astronomy, University of Catania, Catania, Italy
emanuele@martorana.email

² Department of Clinical and Experimental Medicine,
University of Catania, Catania, Italy
{gmicale,ferro,apulvirenti}@dmi.unict.it

Abstract. Network motifs have a central role in explaining the functionality of complex systems. Establishing motif significance requires the computation of the expected number of their occurrences according to a random graph model. Few models have been proposed to analytically derive the expected number of non-induced occurrences of a motif. In this paper we present an analytical model to compute the expected number of occurrences of induced motifs in unlabeled graphs. We will illustrate two different algorithms for computing the occurrence probability of induced motifs. We evaluate the performance of our algorithms for calculating the expected number of induced motifs with up to 10 nodes.

Keywords: Injective motifs · Networks · Random graphs · Analytical models

1 Introduction

Network motifs are small subgraphs of a larger graph which occur more than expected according to a properly chosen random null model. Motif discovery has several applications ranging from biology to social networks and finance [1, 10, 17].

The expectation of a motif is evaluated with a random graph model, which is commonly defined as an ensemble of random graphs that preserve some features of the input network (e.g. the degree distribution). A subgraph is claimed to be a motif when its mean frequency in the null model is statistically significantly lower than in the input graph. Indeed, motif computation requires both fast techniques to establish the number of its occurrences and tools to avoid the actual generation of the random graph ensemble. However, counting the number of occurrences of a subgraph in a graph is related to subgraph isomorphism which is a NP-complete problem.

Given a subgraph m of an input graph G , the common approach to determining whether m is a motif consists of the following steps: (i) generate a large set of random networks sharing some of the characteristics of G according to a reference model; (ii) find the number of occurrences of m in each of these networks; (iii) estimate the p-value by comparing the number of occurrences in the input network with those in the random networks. If a motif m occurs very rarely in the random networks compared to in G , then m is over-represented in G . Conversely, m is under-represented in G . Examples of reference models include: the Erdős-Renyi model (ER model) [4], the Fixed degree distribution model (FDD model) [11], the Expected degree distribution model (EDD model) [2, 13], the Erdős-Renyi mixture for graphs model (ERMG model) [3, 12].

The above simulation-based method yields a measure of the significance of each candidate through the computation of a p-value using a resampling approach [9, 10, 15, 16]. Unfortunately, this method, while yielding sound results, requires a large number of random graphs whose analysis turns out to be computationally unfeasible. Over the last few years, some research has shown how to replace simulation methods with analytical ones. Approximation methods, based on the Erdős-Renyi (ED) model, have been proposed to compute the asymptotic normality of the distribution of topology counts [18]. In 2009, Picard et al. [14] proposed a model to exactly compute the mean and variance of the count of a given pattern under any exchangeable random graph model. The authors make use of the Pólya-Aeppli distribution [5]. The occurrence probability of a motif with k nodes does not depend on its position in the observed network; thanks to the exchangeability property every subset of k nodes in the graph could contain it in [14] authors provide equations to compute motif occurrence probability according to different reference random models (FDD, ER, EDD and ERMG). More recently, analytical models for computing expectations of the number of non-induced occurrences in node-labeled graphs [7] and multi-relational graphs [8] under the EDD model have been proposed.

The analytical model of [14] calculates expectations only for non-induced subgraphs. In order to extend this model to induced subgraphs, authors have proposed the application of Kocay Lemma [6]. This Lemma allows the computation of the number of occurrences of induced subgraphs as a linear combination of non induced ones and vice versa. However, calculating the induced mean using Kocay requires the computation of coefficients of the linear combination, which is a hard computational task, even for motifs with 7 or more nodes.

In this paper we present an efficient analytical model to compute the expected number of induced occurrences of a motif according to the EDD model avoiding the computation of Kocay coefficients. Our experimental analysis clearly shows that our method is capable to calculate the expected number of induced motifs with up to 10 nodes in reasonable running time, whereas the computation based on Kocay coefficients becomes unfeasible starting from motifs with 7 nodes.

2 Definitions

We first give some preliminary definitions concerning graphs and motifs. A *graph* (through the paper also called network) is a pair $G = (V, E)$, where V is the set of vertices and $E = \{(a, b) : a, b \in V\}$ is a set of pairs of vertices. The size of G is $|V|$, i.e. the number of vertices in G . A graph is undirected if $\forall(a, b) \in E, (b, a) \in E$, i.e. each edge can be traversed in both ways, otherwise the graph is directed, i.e. there is a one-way relationship between vertices. Two graphs $G = (V, E)$ and $G' = (V', E')$ are isomorphic iff there exists a bijective function $M : V \rightarrow V'$, called isomorphism, such that $\forall a, b \in V : (a, b) \in E \Leftrightarrow (M(a), M(b)) \in V'$. A subgraph $S = (V', E')$ of a graph $G = (V, E)$ is a graph where $V' \subseteq V$ and $E' \subseteq E$. S is called *induced* if all possible edges in E between nodes of V are also present in S (i.e., $\forall a, b \in V' : (a, b) \in E \Leftrightarrow (a, b) \in E'$). S is called *non-induced* if one or more edges in E between nodes of V are not present in S . So, the definition of induced subgraph is more restrictive than the one of non-induced subgraph. A graph $G' = (V', E')$ is said to be subgraph isomorphic to a graph $G = (V, E)$ if G' is isomorphic to a subgraph of G . Subgraph isomorphism problem is known to be NP-Complete and can have more than one solution. The number of occurrences of G' in G is the number of subgraph isomorphisms of G' in G . A *motif* is a graph which occur more than expected with respect to a null reference model, which is an ensemble of random graphs that preserve some features of the input network (e.g. the degree distribution). In other words, a motif is a graph whose mean frequency in the null model is statistically significantly lower than in the input graph.

3 Analytical Model to Assess Significance of Non-induced Motifs

In [14] authors propose an analytical model to assess the significance of non-induced motifs with respect to several random graph models in directed and undirected graphs. In this paper we focus on undirected graphs and use the Expected Degree Distribution (EDD) model as random graph. However, our model can be easily extended to directed graphs and applied to any exchangeable random model, i.e. any model in which the occurrence probability of a given motif in a graph does not depend on the occurrence position.

EDD model generates random graphs in which node degrees follow the degree distribution of the input graph. Let $Deg(G)$ the degree distribution of graph G with n nodes and X_{ij} an indicator random variable that equals 1 iff there is an edge between nodes i and j . According to the EDD model, the probability of observing an edge between two nodes i and j in a graph is given by

$$P(X_{ij} = 1 | D_i, D_j) = \min(1, \gamma D_i D_j) \quad (1)$$

where $\gamma = \frac{1}{(n-1)\mathbb{E}[Deg]}$, and D_i is the degree of node i sampled according to the Deg distribution. Since EDD is an exchangeable random model, the conditional

occurrence probability of a motif m with k nodes in G , given an assignment of expected degrees D_i to the nodes of the motif, can be expressed as the product of the edge probabilities. To compute the occurrence probability of m , we have to sum all probabilities obtained assigning all possible degrees D_i present in the input network. Such probability can be expressed as products of some moments of the degree distribution Deg of G , as follows:

$$\mu(m) = \gamma^{m_{++}/2} \prod_{u=1}^k \mathbb{E}[\text{Deg}]^{m_{u+}}$$

where m_{++} is the number of edges in m , m_{u+} is the degree of node u in m and $\mathbb{E}[\text{Deg}]^{m_{u+}}$ is the moment of order m_{u+} of Deg distribution. Starting from $\mu(m)$ it is possible to compute both the mean and the variance of the number of non-induced occurrences of m under the EDD model. Thanks to the exchangeability property of the EDD model, $\mu(m)$ does not depend on the specific location of m . All possible k -tuples of vertices in the graph are potential locations of a motif occurrence. If the graph has n nodes, the number of all k -tuples is $\binom{n}{k}$. Moreover, in a given k -tuple of vertices α , a motif can occur in different configurations considering all possible permutations of positions of vertices in m . The number of such permutations is $k!$. However, some of them actually produce redundant occurrences with the same adjacency matrix. We denote with $R(m)$ the set of this Non-Redundant Permutations (NRPs) and with $\varrho(m) = |R(m)|$. Finally, the mean number of non-induced occurrences is given by:

$$\mathbb{E}[N(m)] = \binom{n}{k} \varrho(m) \mu(m) \quad (2)$$

The calculation of variance is based on the expectation of the squared count of non-induced occurrences of m , i.e. $\mathbb{E}[N^2(m)]$. To compute $\mathbb{E}[N^2(m)]$ we need to consider the possible overlap in nodes and edges of two non-redundant occurrences of m . Given two NRPs m' and m'' of m , [14] define the overlapping operation with s nodes, $m' \cap_s m''$, whose result is a super-motif with $2k - s$ nodes. To define the adjacency matrix of the super-motif, the adjacency matrices of m' and m'' are splitted into four blocks of variable sizes as follows:

$$m' = \left(\begin{array}{c|c} m'_{11} & m'_{12} \\ \hline [k-s, k-s] & [k-s, s] \\ \hline m'_{21} & m'_{22} \\ \hline [s, k-s] & [s, s] \end{array} \right) \quad m'' = \left(\begin{array}{c|c} m''_{11} & m''_{12} \\ \hline [s, s] & [s, k-s] \\ \hline m''_{21} & m''_{22} \\ \hline [k-s, s] & [k-s, k-s] \end{array} \right)$$

The adjacency matrix of super-motif is given by:

$$m' \cap_s m'' = \left(\begin{array}{c|c|c} m'_{11} & m'_{12} & 0 \\ \hline m'_{21} & \max(m'_{22}, m''_{11}) & m''_{12} \\ \hline 0 & m''_{21} & m''_{22} \end{array} \right)$$

where the max function in the central term indicates that for the s common vertices of m' and m'' all edges of m'_{22} and m''_{11} must be present. The max function is equivalent to the logical OR. Considering that two occurrences can overlap in one or more nodes (up to k nodes) and taking into account all possible ways two occurrences can overlap with s nodes, the expected square count $\mathbb{E}[N^2(m)]$ is given by:

$$\mathbb{E}[N^2(m)] = \binom{n}{n-2k, k, k} \left[\sum_{m' \in R(m)} \mu(m') \right] + \sum_{s=1}^k \left[\binom{n}{k-s, s, k-s, n-2k+s} \times \right. \\ \left. \times \sum_{m', m'' \in R(m)} \mu(m' \cap_s m'') \right]$$

Finally, the variance is given by $\mathbb{V}[N(m)] = \mathbb{E}[N^2(m)] - \mathbb{E}[N(m)]^2$

4 Analytical Model to Assess Significance of Induced Network Motifs

In [14] authors also describe the computation of the expected number of occurrences of an induced motif under any exchangeable random model. They show that such a number can be expressed as a linear combination of the expected number of non-induced occurrences of all motifs of the same size, using the Kocay Lemma [6]. For example, the expected number of induced paths with 3 nodes is equal to the expected number of non-induced paths with 3 nodes minus the number of non-induced cliques with 3 nodes. The coefficients of the linear combination are given by the Kocay Lemma and they are the entries of the inverse of a matrix K , called Kocay matrix, where $K[i, j] = x$ if there are x non-induced occurrences of motif i within motif j . The same coefficients can be used for calculating the variance of the induced count. However, the computation of variance also requires the covariance between all possible NRPs of a motif. Therefore, in order to calculate induced mean and variance we need to compute the inverse of the Kocay matrix and the means and the variances of all motifs of a given size, which makes the overall computation very hard even for motifs of small size.

In what follows we propose an alternative computation of the expected number of induced occurrences of a motif that avoids the calculation of Kocay matrix. Based on this, we describe an analytical model to derive occurrence probability, mean and variance of induced motifs.

4.1 Occurrence Probability of Induced Motifs

Let $G = (V, E)$ a graph with n nodes and Deg its degree distribution. Let D_i the degree of i -th node. As described in the previous section, the probability of observing an edge between two nodes i and j is obtained with Eq. 1. Then, the probability that no edge is observed between i and j is $P(X_{ij} = 0 | D_i, D_j) = 1 - \gamma D_i D_j$. To evaluate the occurrence probability of an induced motif under

the EDD model, we need to take into account both the existence and the non-existence of an edge within the motif. The occurrence probability can then be expressed as a product of edge and non-edge probabilities. For instance, the induced occurrence probability of the 4-square motif (Fig. 1) can be expressed as:

$$P \{ \exists(AB, BC, CD, DA), \nexists(AC, BD) | D_A, D_B, D_C, D_D \} = \gamma^4 D_A^2 D_B^2 D_C^2 D_D^2 - \\ - \gamma^5 D_A^3 D_B^2 D_C^3 D_D^2 - \gamma^5 D_A^2 D_B^3 D_C^2 D_D^3 + \gamma^6 D_A^3 D_B^3 D_C^3 D_D^3$$

Each term in this summation actually corresponds to the non-induced occurrence probability of a certain motif of the same size (Fig. 1) and the sign in the summation depends on the number of edges of the corresponding motif. Note also that the sign of the term changes every time we go from one motif with h edges to a motif with $h + 1$ edges.

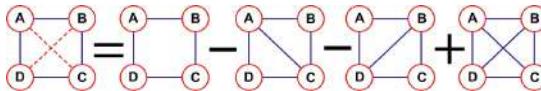


Fig. 1. Induced occurrence probability of 4-node square motif. The probability is obtained as a linear combination of non-induced occurrence probabilities of different motifs of the same size. *Dashed red lines* are missing edges in the 4-square motif, *blue lines* are present edges.

4.2 Additive Set: An Effective Data Structure for Induced Probability Computation

To correctly identify the set of motifs in the linear combination that expresses the induced occurrence probability and their sign in the combination we need to introduce a data structure that we call Additive Set (AS). The AS is a Directed Acyclic Graph (DAG) which represents all motifs that can be obtained by adding one edge at the time from a set of given motifs with k nodes. The AS is characterized by the following properties:

- Root nodes represent starting motifs;
- Each level contains motifs with the same number of edges;
- DAG levels encode the coefficients of the linear combination of the induced probability equation;
- Transition from level L to level $L + 1$ is characterized by the addition of exactly one edge in one of the motifs at level L ;
- Levels go from 0 to r , where r indicates the maximum number of edges that can be added.

Figure. 2 shows an example of AS starting from the 4-node path. Note that the AS may contain many isomorphic motifs. So, even for moderate motif size the AS cannot be represented in main memory. By removing redundant occurrences

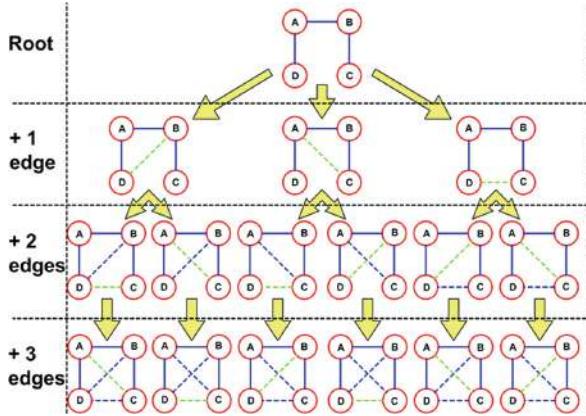


Fig. 2. Uncompressed Additive Set (AS) of the 4-node path. The root of the AS is the 4-node path. Internal nodes of the AS contain motifs built from the 4-node path by adding 1, 2 or 3 edges. *Dashed green lines* are newly added edges, *dashed blue lines* are previously added edges.

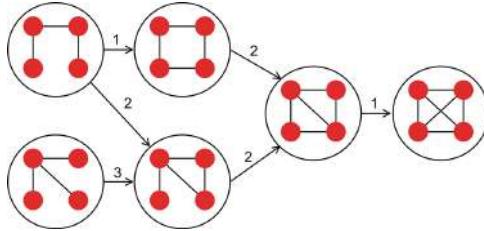


Fig. 3. Compressed AS for all 4-node motifs. Edge weights indicate the number of motifs that can be obtained from another motif by adding one edge in all possible ways.

we can obtain a compressed representation of the AS where edges are weighted according to the number of non-redundant motifs that can be obtained from another motif. Figure 3 shows the compressed AS with all 6 motifs of size 4. Given an AS \mathcal{T} , we can define the Topology Induced Additive Set (TIAS) of a motif m , which is a sub-DAG of \mathcal{T} formed by m and all motifs that can be obtained from m by adding one or more edges (Fig. 4b). Given two motifs m and m' , we define the weight of a path P from m to m' in the TIAS of m , $w(P)$, as the product of the weights of all the edges in the path. If \mathcal{P} is the set of all paths between m and m' in the TIAS of m and p is the level of m' in the TIAS, the quantity $(\sum_{P \in \mathcal{P}} w(P))/p!$ corresponds to the number of occurrences of m' that can be generated from m by adding edges in all possible ways (Fig. 4b). Using the AS we can calculate induced occurrences probabilities of a motif m of size k as described in Algorithm 1. Formally, the induced occurrence probability of a

motif m is given by the following equation:

$$\mu_I(m) = \sum_{i=0}^{L_{max}} \frac{(-1)^i}{i!} \left[\sum_{m \in \mathcal{M}(i)} \beta(m, i) \mu(m) \right]$$

where L_{max} is the maximum level reached in the TIAS of m , $\mathcal{M}(i)$ is the set of motifs stored in the nodes of the TIAS at level i and $\beta(m, i)$ is a recursive function defined as:

$$\beta(m, i) = \begin{cases} 1 & \text{if } i = 0 \\ \sum_{v \in \mathcal{M}(i-1)} w(v, m) \beta(v, i-1) & \text{otherwise} \end{cases}$$

where $w(v, m)$ is the weight of the edge (v, m) in the TIAS.

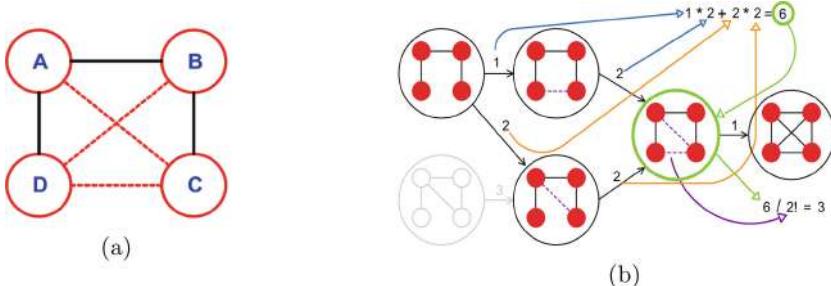


Fig. 4. (a) 4-node path, where black lines represent existing edges, dashed red lines are missing edges. (b) Example of calculation of the number of occurrences of the 4-node square with diagonal, m' , (surrounded by a green circle) generated from the 4-node path. There are two paths between the two motifs, with weights 2 and 4, respectively, and m' is at level 2 in the TIAS, so the number of such occurrences is $6 / 2! = 3$.

Algorithm 1: Induced Probability

1 InducedProbability(m, k);

Input: An adjacency matrix m of a motif and its size k ;

Output: The induced probability of m ;

2 Build the complete AS for topologies with k nodes;

3 Extract the TIAS of motif m from the AS;

4 Starting from root m in the TIAS perform a BFS and for each visited motif m' calculate the number of occurrences of m' that can be generated from m and multiply it by the non-induced probability of m' ;

5 Sum all the terms computed at step 4, where the sign of the term related to motif m' in the sum depends on the level of m' in the TIAS: if the level is even the sign is positive otherwise it is negative;

4.3 Mean and Variance of Induced Motifs

Equations for the mean and the variance of the number of occurrences of induced motifs are similar to those presented in [14] for non-induced motifs. The expectation $\mathbb{E}[N_I(m)]$ can be computed using Eq. 2 replacing $\mu(m)$ by $\mu_I(m)$. Concerning computation of the variance we first need to change the definition of the overlapping operation between NRPs of a motif m . More specifically, the adjacency matrix of the super motif resulting from the overlapping operation (named here \cap_s^I) between two NRPs of m , m' and m'' , with s nodes can be written as:

$$m' \cap_s^I m'' = \left(\begin{array}{c|cc|c} m'_{11} & m'_{12} & 0 \\ \hline m'_{21} & \Delta(m'_{22}, m''_{11}) & m''_{12} \\ \hline 0 & m''_{21} & m''_{22} \end{array} \right) \quad (3)$$

All the blocks except the central one are defined as in the case of non-induced motifs. The Δ function in the central block of super-motif is defined as:

$$\Delta(m'_{22}, m''_{11}) = \begin{cases} m'_{22} \vee m''_{11}, & \text{if } \sum_{i,j} m'_{22_{i,j}} = \sum_{i,j} m''_{11_{i,j}} \\ \text{discard,} & \text{otherwise} \end{cases}$$

where *discard* indicates that the super-motif is not considered in the computation of variance. So, the overlap is defined iff all the edges in the overlapping region come from both m' and m'' . In order to correctly compute the variance, for each super-motif S generated with Eq. 3, we also need to consider all motifs that can be obtained by S by adding one or more edges in all possible ways between two nodes, outside the overlapping region. So, the expected square count $\mathbb{E}[N_I^2(m)]$ is computed considering the induced occurrence probability of all super-motifs generated using the overlapping function \cap_s^I and the induced occurrence probability of all extensions of each super-motif generated as previously described. Finally, the variance $\mathbb{V}[N_I(m)]$ is given by $\mathbb{E}[N_I^2(m)] - \mathbb{E}[N_I(m)]^2$.

4.4 RaME - RApid Matrix Elaboration

The proposed analytical model has been implemented in Java in a package called RaME (RApid Matrix Elaboration). RaME uses efficient matrix operations to compute induced motif probabilities. Indeed, we can observe that in order to calculate induced probabilities we do not need the explicit generation of motifs in the AS but we just need moments of the degree distributions of these motifs. Let m be a motif with k nodes and let $E_A(m)$ be the set of its absent edges. We denote with \mathcal{C} the set of all $2^{|E_A|} - 1$ possible combinations of edges in $E_A(m)$. We can compute a matrix $M_{\mathcal{C}A}$ with $2^{|E_A|} - 1$ rows and k columns, where $M_{\mathcal{C}A}[c, x]$ is the number of edges of combination c to which node x is incident. Table 1 shows an example of computation of $M_{\mathcal{C}A}$ for the 4-node path of Fig. 4a. If we add to each cell (i, j) of $M_{\mathcal{C}A}$ the degree of node j in m , we obtain a new matrix M_D (see Table 2 as an example) where each row contains the degrees of nodes of a motif m' in the TIAS \mathcal{T} of m (Fig. 4b). Starting from node degrees of each motif in the TIAS, we can finally calculate a matrix of moments M_E where $M_E[m', u]$ contains the moment of order m'_{u+} of the degree distribution of motif m' and

$\prod_{u=1}^k M_{\mathbb{E}}[m', u] = \mu(m')$. Then, the induced occurrence probability of m can be written as:

$$\mu_I(m) = \sum_{m' \in \mathcal{T}} \text{Sign}(m') * \gamma^{\frac{m'_{++}}{2}} * \mu(m')$$

where $\text{Sign}(m') = (-1)^{(\frac{m'_{++}}{2} \bmod \frac{m_{++}}{2})}$. Concerning the complexity, given a motif with k nodes, RaMe requires $\mathcal{O}(2^{k^2} k)$ and $\mathcal{O}(k^2(k!)^3(2^{k^2})^2)$ to compute expected counts and variance of the motif within the graph. Whereas, Kocay requires $\mathcal{O}(2^{k^2} k!k^2)$ and $\mathcal{O}(k^2(k!)^2)$ for the expectation of the counts and variance of the motif within the graph.

Table 1. Combination of absent edges

Edges comb	A	B	C	D
AC	1	0	1	0
BD	0	1	0	1
CD	0	0	1	1
(AC,BD)	1	1	1	1
(AC,CD)	1	0	2	1
(BD,CD)	0	1	1	2
(AC,BD,CD)	1	1	2	2

Table 2. Matrix of resulting degrees

A	B	C	D
3	2	2	1
2	3	1	2
2	2	2	2
3	3	2	2
3	2	3	2
2	3	2	3
3	3	3	3

Table 1 is the matrix $M_{\mathcal{CA}}$ of combinations of absent edges of m , where $M_{\mathcal{CA}}[c, x]$ is the number of edges in combination c to which node x is incident. Table 2 is the matrix M_D of node degrees of motifs in the TIAS of the 4-node path, obtained by adding node degrees of the 4-node path $(2, 2, 1, 1)$ to each row of $M_{\mathcal{CA}}$.

5 Experimental Results

In this section we evaluate the performance of RaME for the computation of the mean count of induced motifs in a dataset of real undirected networks of different size. We compared RaME to a Java implementation of the Kocay-based analytical model described in Sect. 4.

5.1 Dataset

The dataset used for our experiments includes 4 networks downloaded from KONECT¹: **Human Protein (Vidal)** is a proteome-scale map of physical interactions between human proteins; **CAIDA** is an infrastructural network of

¹ <http://konect.cc>.

communications between autonomous systems of Internet; **DBLP** is a coauthorship network where nodes are authors and an edge connects two authors if they published at least one paper together; **LiveJournal** is the map of russian social network LiveJournal where users are nodes and edges are their relationship status. In Table 3 we report the size of these networks.

5.2 Results

Tests were performed using an Intel core i5-7400 processor with 8 GB of RAM. Table 3 shows the running time (in seconds) of both algorithms for the calculation of the expected number of induced motifs in each network. For RaME, we report results for all motifs of a given size and for a single topology of that size (the star topology) Table 4, which represent the worst-case for our algorithm. In fact, star topologies have the smallest number of edges and the highest number of motifs in their TIAS. So, the running time for any single motif with k nodes in RaME will be less than or equal to the one of the k -node star topology. For Kocay algorithm we only report results for all k -node motifs, since both for all k -node motifs and for a single motif the algorithm must compute the entire Kocay Matrix, the inverse of the matrix and the non-induced probabilities of all k -node motifs. RaME is faster than Kocay algorithm for larger motifs. Moreover, the running time of RaME does not depend on network size, because RaME is only based on the degree distribution of the input network. Due to the combinatorial explosion of the number of possible motifs of size k , for $k = 8, 9, 10$ we just focused on the induced occurrence probability of the star topology. Note that in these cases RaME can still compute expected counts in reasonable running time (few hours, see Table 5), while Kocay algorithm cannot finish the task or runs out of memory. In our experiments we did not compute the variance of the induced count, because the calculation of variance using our model is computational intensive and therefore unfeasible in a single machine .

Table 3. Information of network dataset

Network	Category	Nodes	Edges
Human Protein	Metabolic	3'133	6'726
CAIDA	Computer	26'475	53'381
DBLP	Coauthorship	317'080	1'049'866
LiveJournal	Social	5'204'176	49'174'464

Table 4. Execution time (seconds) comparison for induced mean calculation for all motifs of size ranging from 3 to 7 nodes

Motif size	Human Protein		CAIDA		DBLP		LiveJournal	
	Kocay	RaME	Kocay	RaME	Kocay	RaME	Kocay	RaME
3	0.12	0.02	0.22	0.10	1.28	1.01	86.21	88.23
4	0.13	0.02	0.25	0.10	1.30	1.03	89.35	90.65
5	0.20	0.02	0.32	0.12	1.37	1.05	92.72	93.43
6	1.75	0.21	1.65	0.25	2.70	1.17	97.28	94.67
7	117.44	48.14	118.92	59.73	113.25	52.08	245.07	160.56

Table 5. RaME execution time (s) for induced mean calculation for star topology

Star size	Human protein	CAIDA	DBLP	LiveJournal
3	0.01	0.04	0.87	73.46
4	0.01	0.05	0.87	75.81
5	0.02	0.05	0.88	78.62
6	0.03	0.05	0.93	79.31
7	0.06	0.07	1.15	80.36
8	0.50	0.43	1.62	81.54
9	33.87	33.52	36.38	134.60
10	9'517.49	8'861,13	9'009.11	13'152.37

6 Conclusions

In this paper we introduced a novel method to compute the expected number of induced motifs on undirected large networks. Our model, compared with the Kocay-based model, is faster when the size of the motif increases. We plan to implement such algorithm and its approximated variants on distributed environment on top of SPARK framework to deal with very large networks.

References

- Chen, J., Yuan, B.: Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics* **22**(18), 2283–2290 (2006)
- Chung, F., Lu, L.: The average distances in random graphs with given expected degrees. *Proc. Natl. Acad. Sci.* **99**(25), 15879–15882 (2002)
- Daudin, J.J., Picard, F., Robin, S.: A mixture model for random graphs. *Stat. Comput.* **18**(2), 173–183 (2008)
- Erdos, P., Renyi, A.: On random graphs. *Publ. Math.* **6**, 290–297 (1959)
- Johnson, N.L., Kotz, S., Kemp, A.W.: *Univariate Discrete Distributions*. Wiley (1992)

6. Kocay, W.: An extension of Kelly's lemma to spanning subgraphs. *Congr. Numer.* **31**, 109–120 (1981)
7. Micale, G., Giugno, R., Ferro, A., Mongiovì, M., Shasha, D., Pulvirenti, A.: Fast analytical methods for finding significant labeled graph motifs. *Data Min. Knowl. Discov.* **32**(2), 1–28 (2018)
8. Micale, G., Pulvirenti, A., Ferro, A., Giugno, R., Shasha, D.: Fast methods for finding significant motifs on labelled multi-relational networks. *J. Complex Netw.* cnz008 (2019)
9. Milo, R., Kashtan, N., Itzkovitz, S., et al.: On the uniform generation of random graphs with prescribed degree sequences. *Cond. Mat.* **0312028**, 1–4 (2004)
10. Milo, R., Shen-Orr, S., Itzkovitz, S., et al.: Network motifs: simple building blocks of complex networks. *Science* **298**(5594), 824–827 (2002)
11. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001)
12. Nowicki, K., Snijders, T.: Estimation and prediction for stochastic block structures. *J. Am. Stat. Assoc.* **96**, 1077–1087 (2001)
13. Park, J., Newman, M.: The origin of degree correlations in the Internet and other networks. *Phys. Rev. E* **68**, 026112 (2003)
14. Picard, F., Daudin, J.J., Koskas, M., et al.: Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15**(1), 1–20 (2008)
15. Prill, R., Iglesias, P.A., Levchenko, A.: Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol.* **3**(11), e343 (2005)
16. Shen-Orr, S.S., Milo, R., Mangan, S., et al.: Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* **31**, 64–68 (2002)
17. Squartini, T., Garlaschelli, D.: Analytical maximum-likelihood method to detect patterns in real networks. *New J. Phys.* **13**(8), 083001 (2011)
18. Wernicke, S.: Efficient detection of network motifs. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **3**(4), 347–359 (2006)



Network Rewiring Dynamics to Form Clustered Strategic Networks

Faisal Ghaffar^{1,2(✉)} and Neil Hurley²

¹ Innovation Exchange, IBM Ireland, Dublin, Ireland
faisalgh@ie.ibm.com

² University College Dublin, Dublin, Ireland
neil.hurley@ucd.ie

Abstract. Burt’s *Structural Hole Theory* provides a theoretical foundation for individuals in a network to strategically seek a position in the network that gives them advantages by connecting them with a diverse range of others in different social cliques. Kleinberg et al. in [10] proposed an algorithm for the best response dynamics for the individuals in a network when they act strategically to maximize the number of structural holes in their neighbourhood during the formation of links. In this paper, we demonstrate through a set of experiments that networks that emerge at equilibria of strategic games such as the one proposed in [10], do not have characteristics of real-world networks. This leads us to follow an approach of studying the capacity of a network to hold maximum number of structural holes while maintaining its properties such as degree distribution and clustering coefficient. We also propose a new payoff utility function and a stochastic dynamic rewiring process with modified pairwise stability. Carrying out a set of experiments on real-world and synthetically generated networks, we empirically examine the number of structural holes that can be maintained in a network with realistic characteristics. We demonstrate that our payoff utility is able to maintain the clustering coefficient in a degree preserving rewiring scheme.

Keywords: Strategic network formation · Network rewiring · Payoff maximization

1 Introduction

Complex networks are flexible and generic enough that they can virtually represent any natural domain, in which interacting entities can be identified, and an analysis of the resulting network can yield useful insights into that domain. Investigations of complex networks typically start with an analysis of the topological features of the obtained representation. It is remarkable that similar network characteristics have been observed in networks arising from many diverse contexts, and the question arises as to what generative process accounts for this observed structure. Several network models have been proposed for the purpose

of studying topological features of complex systems. A model, which has been well-studied theoretically, is the Random Graph model (developed by Erdős and Rény [4]) in which the underlying structure is formed through a uniformly random process. However, real world networks have characteristics which are not explained by random processes or models and this fact has been the main driver of research that has focused on other models of complex networks generation. The network characteristics that such models seek to explain include community structure, the existence of hub nodes with very high connectivity and a power law degree distribution, among other structural features. Models proposed to explain these characteristics include: Watts and Strogatz's small-world networks [12], Barabási and Albert's scale-free networks [1], and Girvan and Newman's identification of the community structure [5] present in many networks. However, none of these models have the ability to explain the strategic reasons behind formation of complex networks. For this purpose, researchers have looked at network formation from an economic perspective and have proposed strategic network formation models.

1.1 Strategic Network Models

Social networks provide a breeding ground to spread information and ideas. Individuals that disseminate information in social networks receive benefits and incur costs in terms of time and effort as a consequence of making connections with others. With costs associated with links, individuals in the network act strategically while selecting their connections. It is important to understand the effect of strategic behaviour of individuals in the network on the formation of links. Recently many researchers have proposed several models of social network formation using game-theoretic approaches to explain the costs and benefits of links [6, 9]. The crux of these strategic network formation models is a strategy in the form of a game in which players are individuals in the network, the strategy of each individual is the choice of neighbours and the utility of each individual depends on its neighbourhood and the structure of the network.

From the perspective of the diffusion of information or ideas, one important concept in sociology is that individuals benefit from being intermediaries or bridges between others who are not directly connected. Such individuals are said to occupy *Structural holes* in the network [2]. In particular, Burt's *structural hole theory* at its core argues that the gap or "hole" between two unconnected neighbours of an individual gives structural advantage to that individual in leveraging neighbours for its own benefit. A scenario with such structural advantages can be observed in Fig. 1, which shows a node A acting as a bridge for nodes B and C and K. The benefits that accrue to node A are due to its being a brokerage for the flow of information between nodes B, C and K. Node A is in a position to receive novel information by synthesizing ideas arising from different parts of the network, i.e., from the communities of B, C and K.

With structural hole benefits known, it might be expected that individuals seek out opportunities to realize them by strategizing their decisions about link formation. A number of models of such strategic network formation have been

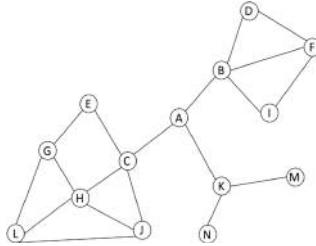


Fig. 1. A toy network with node A spanning structural hole between B and C

proposed. In [7], Jackson gives a good review of such models. Also, strategic formation models with structural holes have been studied by [6, 10]. In particular, Kleinberg in [10] proposes a general model in which an individual's payoff is modelled as being dependent on the number of structural holes that she bridges in the network and individuals form links in order to maximise their payoff. The equilibrium networks that result from playing this strategy link formation 'game' are studied.

However, the networks resulting from strategic models for structural holes like the one in [10] lack the characteristics of real-world networks. In particular, emerged networks do not exhibit the following characteristics of real-world socially generated networks:

- high clustering co-efficient (in comparison to networks generated through an independent random process);
- power-law degree distribution;
- small average path length and network diameter (on the order of the log of the number of nodes).

1.2 Our Contributions

First through a set of experiments on real-world networks, we demonstrate the payoff function of [10] does not produce a network that has the properties of a real-world network. We then propose a new payoff and demonstrate that it addresses the challenge of maintaining the clustering coefficient in a set of experiments on real-world and synthetically generated networks. Specifically our contributions are as follows:

- We propose a new payoff function based on the intuition that structural holes are only useful if they bridge clustered nodes. The proposed payoff is compared against Kleinberg's payoff through a strategic game. We show that networks that maximize our proposed utility preserve the clustering coefficient of the original network.
- We extend the pairwise stability concept for two edges. The pairwise stability notion in its original form has a weak requirement in that it checks whether utilities of end-points are worse off if a single node unilaterally severs an edge or both nodes add a new link.

- A network degree preserving rewiring mechanism is implemented using the proposed utility and the pairwise stability to demonstrate the network's ability to maximize its overall payoff (in terms of structural holes) without changing its structure, i.e., clustering coefficient and degree distribution.

2 Preliminaries

2.1 Networks

Let $G = (V, E)$ be a graph with a set of vertices (also referred to as nodes) V , $|V| = n$ and a set of edges (also referred to as links or connections) $E \subseteq (i, j)|i, j \in V$. Edges are often represented by e_{ij} where i and j denote the terminal vertices of edge e . The graph G is regarded as undirected if each edge e_{ij} in G is defined by unordered pairs of vertices i and j , which also means that the relationship between vertices is symmetric, i.e. $e_{ij} = e_{ji}$. For a vertex $i \in V$, we define the *neighborhood* of i in G to be $N(i) = j \in V|e_{i,j} \in E$ and the degree of i as $d_i = |N(i)|$.

The distance $d(i, j)$ between two nodes i and j is the minimum path length between i and j , where a *path* in a network $G = (V, E)$ between two nodes i and j is a sequence of nodes i_1, \dots, i_K such that $e_{k,k+1} \in E$ for each $k \in \{1, \dots, K-1\}$, with $i_1 = i$ and $i_K = j$. The length of such path is $K - 1$. The *diameter* of G is then defined as $\max_{i,j \in V} d(i, j)$, the maximum distance between any two connected nodes.

2.2 Clustering

The *local clustering coefficient* of vertex i indicates the cohesiveness of a single node and is defined as a ratio between the number of triangles that vertex i is involved in, denoted here as Δ_i , given in Eq. 1, to the number of possible triangles i could be involved in, represented as τ_i , given in Eq. 2.

$$\Delta_i = |e_{ik} \in E| e_{ij} \in E, e_{jk} \in E| \quad (1)$$

$$\tau_i = \frac{d_i(d_i - 1)}{2} \quad (2)$$

The local clustering coefficient of vertex i is:

$$c_i = \frac{\Delta_i}{\tau_i} = \frac{2\Delta_i}{d_i(d_i - 1)} \quad (3)$$

With the node level clustering coefficient known, we can define the *average clustering coefficient* for the network as:

$$c(G) = \frac{1}{n} \sum_{i \in V} c_i \quad (4)$$

2.3 Utility and Efficiency

The utility $u_i(G)$ of a node i is the net benefit that a node receives in a network G . A network is considered *efficient* if the overall utility of all nodes in the network is maximum.

2.4 Pairwise Stability

A network g is pairwise stable (see [9]) if no single node in the network would like to sever an edge and no two nodes both want to add an edge, as, if they did so, the utility of the endpoints in both cases (deletion or addition of an edge) would decrease. The pairwise stability for an undirected network g can be defined as

1. $\forall e_{ij} \in E, u_i(g) \geq u_i(g - e_{ij})$ and $u_j(g) \geq u_j(g - e_{ij})$
2. $\forall e_{ij} \notin E, u_i(g) < u_i(g + e_{ij})$ then $u_j(g) > u_j(g + e_{ij})$

where we write $g \pm e_{i,j}$ to represent the removal or addition of the edge.

2.5 Improving Paths

In [8] the concept of *improving paths* to analyse the process of network evolution is proposed. An improving path is a sequence of networks that emerges when nodes add or sever edges and each subsequent network g_k emerging in the evolution process has improved overall utility over its predecessor. The notion of improving path is myopic in that each network in the sequence differs from the previous network by the addition or deletion of a single edge which benefits the nodes involved. Formally, an improving path from a network g to a network g' is a finite sequence of adjacent networks g_1, g_2, \dots, g_K with $g_1 = g$ and $g_K = g'$ such that for any $k \in \{1, \dots, K\}$ either:

1. $g_{k+1} = g_k - e_{ij}$ for some e_{ij} such that $u_i(g_k - e_{ij}) > u_i(g_k)$; or
2. $g_{k+1} = g_k + e_{ij}$ for some e_{ij} such that $u_i(g_k + e_{ij}) > u_i(g_k)$, and $u_j(g_k + e_{ij}) \geq u_j(g_k)$.

3 Payoff Model and Rewiring Mechanism

In this section, we define our network formation model. First, we define our payoff utility function and then we define our model of network generation and the process of initializing our model.

3.1 Clustered Nodes Based Payoff

The model proposed in [10] captures direct link benefits along with bridging benefits for nodes that lie in the middle of length-two paths between unconnected pairs of nodes. The model also proposes an algorithm in polynomial time for a node in the network to determine its best response as nodes can chose to connect

with any subset of other nodes and the model also guarantees an equilibrium with ‘Nash equilibrium’ as the notion of stability. In strategic terms, a network is *efficient* if the sum of utilities of the nodes in the network is maximal [3, 9]. From the topological point of view, the only possible topologies for efficient networks are the topologies of a complete graph or a start graph [11]. The authors in [10] also characterized the structure of stable networks that arise as a result of equilibria of the strategic game as complete, multipartite graphs. In other words, the resulting networks have maximum clustering coefficient which means there are no open triangles and no structural holes. From a strategic point of view, this means if everyone in the network strives to maximize their structural holes, the purpose is lost and there is no unique information flowing among nodes in the network. Therefore, our first goal is to allow nodes to maintain some closed triangles or clusters as well as open triangles as structural holes. We define our payoff utility based on this concept.

Our payoff definition is similar to the payoff model of [10] with an exception in the manner that the intermediary benefits are measured. We define intermediary benefits of a node as the number of two-length paths that involve the node and have terminal nodes whose clustering coefficient is higher than a certain threshold. The intuition behind this definition is the core concept of structural hole theory—the nodes that act as brokers between different parts or groups of the network are better positioned for access to unique information. The clustering coefficient metric is an indication of the embeddedness of nodes within their neighbourhood. This payoff utility function captures benefits not only from direct contacts but also from indirect contacts by lying in the paths connecting different communities or cliques. Formally, the intermediary benefit, $I_i(G)$ of a node i with edges e_{ij} and e_{ik} in network G is given as:

$$I_i(G) = \sum_{j,k \in N(i)} \delta r_{jk} \quad (5)$$

where r_{jk} is the number of paths of length two or more between nodes $j, k \in N(i)$, such that the clustering coefficients of these nodes satisfy $c_j > t$ and $c_k > t$, where t is a threshold value. δ is the direct link benefits. With intermediary benefits I_i known, δ and c as link benefits and costs(resp.), we define a node i ’s payoff in G as:

$$u_i(G) = d_i(\delta - c) + I_i \quad (6)$$

Under a uniform metric, in which $\delta = c$, nodes payoff are based only on their intermediary benefits.

3.2 Dynamic Simulation Algorithm with Pairwise Stability

In this section, we describe the process of stochastic network reformation which is based on the notion of improving path described in Sect. 2. Under this notion, the process starts from a given network g and at discrete set of times, $\{1, 2, 3, \dots\}$ a decision of a two edge (let’s say e_{xy} and e_{uv}) swap is made. The nodes involved

Algorithm 1. DPR Algorithm

```

Input:  $G, c, \delta$ 
Output:  $G'$ 
1 Function CalculatePayoffs( $G, c, \delta$ ):
2    $payoffs \leftarrow \{\}$ 
3   for all  $i \in V(G)$  do
4      $u_i = d_i(\delta - c) + I_i$      $payoffs\{i\} \leftarrow u_i$ 
5   return payoffs
6 initialPayoffs  $\leftarrow$  CalculatePayoffs ( $G, c, \delta$ )
7 iterations  $\leftarrow N$ 
8 swapSuccessful  $\leftarrow$  False
9 for  $i = 0; i < N; i = i++$  do
10    $H \leftarrow G.\text{copy}()$ 
11   Find randomly a pair of edges  $(e_{uv}, e_{xy}) \in H$ 
12   if  $!H.\text{hasEdges}(e_{uy}, e_{vx})$  and  $!swapSuccessful$  then
13      $H.\text{addEdges}([e_{uy}, e_{vx}])$ 
14      $H.\text{removeEdges}([e_{uv}, e_{xy}])$ 
15     swapSuccessful  $\leftarrow$  True
16   end
17   updatedPayoffs  $\leftarrow$  CalculatePayoffs ( $H, c, \delta$ )
18   if  $updatedPayoff > initialPayoff$  then
19      $G \leftarrow H.\text{copy}()$ 
20     initialPayoff  $\leftarrow$  updatedPayoff
21   end
22 end

```

act myopically, swapping the edges if such a swap makes each at least as well off and two strictly better off in the context of our payoff definition. In this scenario, the adjacent network g' to the given network g in the improving path is $g' = g - e_{xy} - e_{uv} + e_{uy} + e_{yv}$ and the process strictly follows the condition $u_i(g') > u_i(g)$ **or** $(u_i(g') > u_i(g)$ **and** $u_j(g') > u_j(g))$ which, in simple terms, means that the edge swap should strictly benefit at least two nodes. Under the above terminology, the network is pairwise stable if it is not defeated by an adjacent network. According to [8], our process is in a *stable state* if there is some time t after which no edges are swapped.

Algorithm 1 - Degree Preserving Rewiring (DPR). Considering the improving path and pairwise stability for a pair of edges, an algorithm for a degree preserving rewiring of the network is designed (given in Algorithm 1). We assume a setting in which a set of vertices V are connected with each other via a fixed number of edges in the form of a graph G as described in the previous section. We model the network rewiring as a strategic game where the rewiring mechanism maximizes each node's payoff (measured in this article as the number of structural holes) while keeping the network characteristics close to the original network. Our model enables us to identify the maximum number of structural holes, a network can sustain without losing its structural characteristics. The

fundamental idea of the model and algorithms we present here is as follows: we wish to understand the network that emerges when all nodes in the network are trying to increase the structural hole opportunities for themselves.

4 Experiments with Real-World Networks

In this section, we present results from experiments on a number of real-world networks. In particular, we test our modified payoff function and rewiring algorithm on the following networks: (i) Karate Club (KC), (ii) Football Network (FN) [5], (iii) Social graph of IBM's one business unit technical leadership team, (iv) Erdős and Rény random network (ER) [4], and (v) Co-Inventors patent collaboration network¹. The properties of these networks are given in Table 1.

Table 1. Properties of real-world and random networks

Network	#nodes	#edges	$\langle k \rangle$	cco
Karate club	34	78	4.5	0.57
ER random	33	77	4.66	0.106
Football network	115	613	10.66	0.403
Co-inventors network	232	494	4.25	0.638
IBM social graph	206	971	9.4	0.4232

4.1 Results from Best Response Dynamics Using Kleinberg vs Modified Payoff

In the first set of experiments we play Kleinberg's game in its simple form - add or delete edges for each node using the node's best response which is determined by the utility function given in [10]. Figure 2a shows the simulation results of Kleinberg's game for the KC network and the resulting degree histogram. As seen in the plots, the clustering coefficient (CC) gradually increases whereas the overall network's payoff decreases after first 5 iterations. If the simulation is run for more iterations, a stable complete network emerges as observed by the degree histogram in Fig. 2b. With the modified payoff function, we observe an increase in the payoff instead of a decrease (Fig. 2c) which gets stable for longer iterations of the simulation. There are total 1,852 structural holes with the modified payoff compared to 396 using Kleinberg's payoff function in stable networks. Additionally, with the modified payoff, the network diameter is reduced from 5 to 2 as there are more clustered nodes. However, the modified payoff did not help in preserving the structure of the network. This leads us to think of the number of structural holes that can be sustained in a given network while keeping the network characteristics such as degree distribution and clustering coefficient unchanged.

¹ <http://www.patentsview.org/download/>.

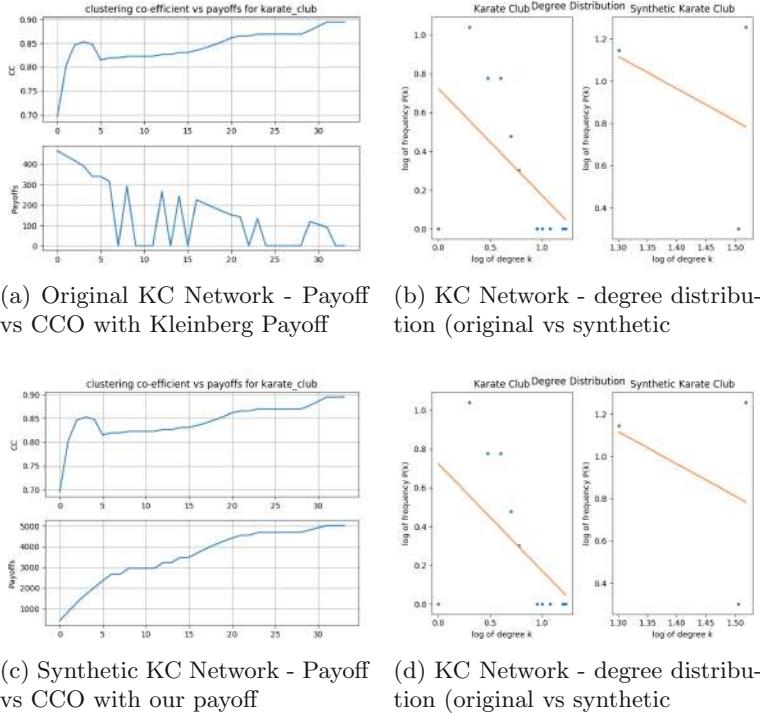


Fig. 2. Simulation results and degree histogram for Kleinberg vs modified payoff

4.2 Results from Structure Preserving Algorithms

In a second series of experiments, we tested our rewiring algorithm with both payoff utility functions. Rewiring simulations randomly swap edges between two pairs of connected nodes if rewiring strictly satisfies the pairwise stability condition that we have adapted for our mechanism and is defined in Sect. 3.2. Our goal here is twofold; (i) determine how many structural holes a network can sustain while preserving the degree distribution and clustering coefficient, (ii) observe properties such as the payoff max/min and distribution of the emerged stable networks and compare them with the original ones.

Figure 3a and c shows KC network's simulation results for clustering coefficient vs network payoff for the degree preserving rewiring using Kleinberg's payoff utility function and our modified payoff utility respectively. In the plots, we can see that with Kleinberg's payoff utility model, the network is not able to maintain its average clustering coefficient whereas with the modified payoff, the network's clustering coefficient stays very close to its original value while gradually increasing the payoff. However, the degree distribution in both cases is same and follows a power-law distribution as shown in Fig. 3b and d. The results on rest of the networks are shown in Fig. 4. The networks are not able to maintain their average clustering coefficient when rewired using Kleinberg's

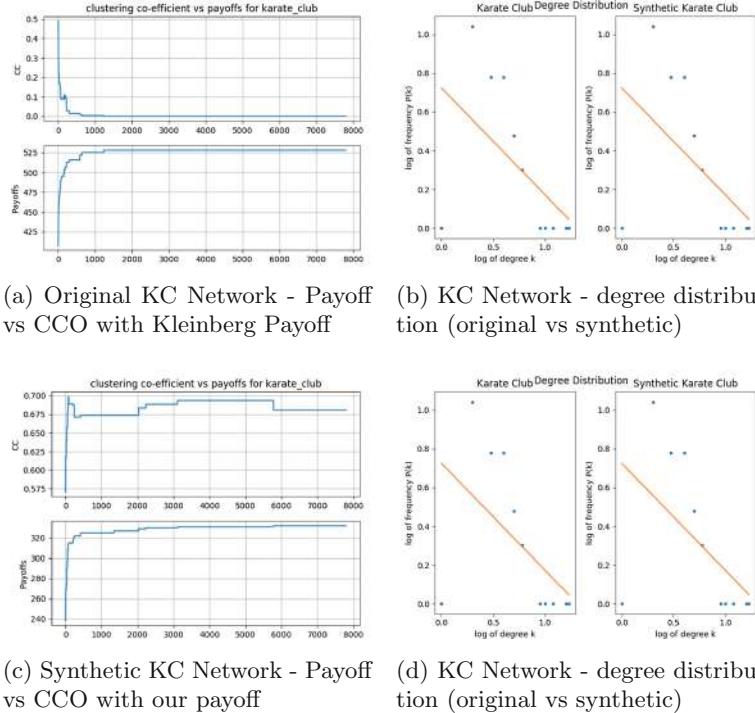


Fig. 3. Simulation results and degree histogram for structure preserving rewiring

utility function (as can be seen in Fig. 4a, c, e, and g). However, with our modified payoff function, all real-world networks are able to maintain clustering coefficient except random network as shown in Fig. 4, b, d, f, and h.

In addition to the clustering coefficient vs the payoff simulation plots, we also compared the emerged stable networks after rewiring with both payoff functions and summarised this comparison in terms of the network properties given in Table 2. The values for each property are given in the format “real network/generated network”. All the bold values highlight where each payoff utility function performed better. The clustering coefficient values are important to note in this table. As we can see that the networks that result from kleinberg’s payoff

Table 2. Kleinberg payoff vs modified payoff - Rewiring results comparison - values are in the format of “Real network/Generated network”

Networks	#structural holes	Clustering Coefficient	Assortivity	Diameter	Transitivity
	Kleinberg Ours	Kleinberg Ours	Kleinberg Ours	Kleinberg Ours	Kleinberg Ours
Karate Club	393/528	238/366	0.570/0.0 0.570/ 0.606 0.475/-0.612 0.475/- 0.3985	5/4	5/5 0.255/0.0 0.255/ 0.306
ER Random	210/219	31/123	0.027/0.0 0.208/ 0.574 0.185/-0.248 0.033/- 0.168	5/4	6/8 0.041/0.0 0.118/ 0.419
Football Network	3537/5967	2884/5579	0.403/0.0 0.403/ 0.065 0.162/0.084 0.162/0.081	4/3	4/3 0.407/0.0 0.407/ 0.064
Patent Collaboration	1296/2709	592/840	0.638/0.0 0.638/ 0.501 0.436/0.0299 0.436/ 0.382	∞/∞	∞/∞ 0.521/0.0 0.521/ 0.423
IBM Social Graph	12590/20129	11347/12519	0.423/0.125 0.399/ 0.388 0.063/-0.403	0.063/ 0.16	∞/∞ ∞/∞ 0.503/0.205 0.503/ 0.456

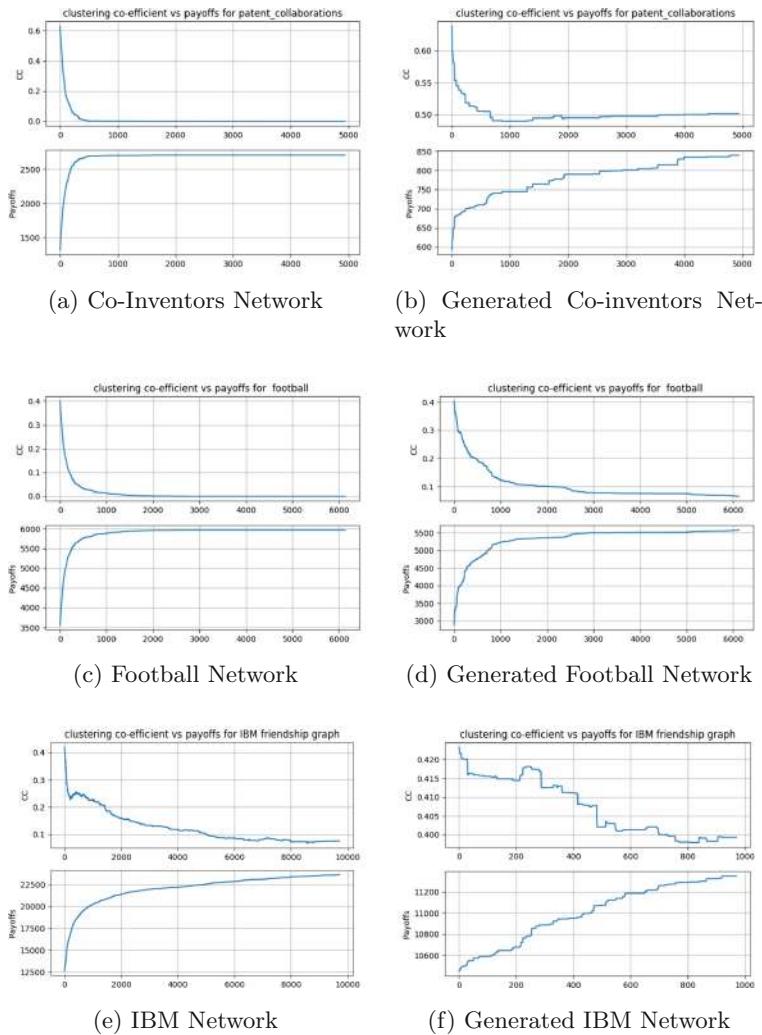


Fig. 4. DPR algorithm results on a number of real-world networks

utility have clustering coefficient either close to or equal to zero. However, the clustering coefficient values using our modified payoff function are close to the original values or at least above zero. The results on *#structural holes* are interesting. For all networks, Kleinberg's payoff model performed better than our modified version but that performance comes with the expense of the network losing its clustering coefficient. We observe that our modified payoff is capable of maintaining average clustering coefficient for emerged stable networks in almost all the cases. This indicates the capacity of each network to have maximum number of structural holes before loosing it's properties.

5 Conclusion

In this paper, we played the Kleinberg's [10] strategic game to maximize the number of structural holes for individuals in a number of real-world networks. We modified the payoff definition with the idea that structural holes are only beneficial if they are among clustered nodes. We observed that the network converges into a stable network more quickly compared to the original payoff definition. However, the resulting network is still not able to maintain its characteristics. This gives rise to question, how many structural holes a real-world network can sustain before it loses its structural properties such as degree distribution and clustering coefficient? For this purpose, we first extended the pairwise stability notion for two edges and then designed a stochastic rewiring mechanism with extended pairwise stability as the notion of stability. Our experiments demonstrate that unlike Kleinberg's payoff utility function, our modified version is able to maintain the clustering coefficient property of the network while maximizing the overall network payoff. However, our experiments involve maximizing the payoff for the overall network and we have left the evaluation of our modified utility function for individual's myopic response when used in a strategic game such as the one in [10], as future work.

References

1. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999)
2. Burt, R.S.: *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge (1992)
3. Buskens, V., Van de Rijt, A.: Dynamics of networks if everyone strives for structural holes. *Am. J. Sociol.* **114**(2), 371–407 (2008)
4. Erdős, P., Rényi, A.: On random graphs i. *Publicationes Mathematicae Debrecen*, vol. 6, p. 290 (1959)
5. Girvan, M., Newman, M.E.: Community structure in social and biological networks. *Proc. Nat. Acad. Sci.* **99**(12), 7821–7826 (2002)
6. Goyal, S., Vega-Redondo, F.: Structural holes in social networks. *J. Econ. Theory* **137**(1), 460–492 (2007)
7. Jackson, M.O.: The stability and efficiency of economic and social networks. In: *Advances in Economic Design*, pp. 319–361. Springer, Heidelberg (2003)
8. Jackson, M.O., Watts, A.: The evolution of social and economic networks. *J. Econ. Theory* **106**(2), 265–295 (2002)
9. Jackson, M.O., Wolinsky, A.: A strategic model of social and economic networks. *J. Econ. Theory* **71**(1), 44–74 (1996)
10. Kleinberg, J., Suri, S., Tardos, É., Wexler, T.: Strategic network formation with structural holes. In: *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 284–293. ACM (2008)
11. Narayananam, R., Narahari, Y.: Topologies of strategically formed social networks based on a generic value function-allocation rule model. *Soc. Netw.* **33**(1), 56–69 (2011)
12. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. *Nature* **393**(6684), 440 (1998)



Measuring Local Assortativity in the Presence of Missing Values

Jan van der Laan^(✉) and Edwin de Jonge

Statistics Netherlands (CBS), The Hague, The Netherlands
{dj.vanderlaan,e.dejonge}@cbs.nl

Abstract. Assortativity or homophily, the tendency of network nodes to connect with similar nodes, often is an useful property to monitor in social networks. For example, when applied to a demographic social network, assortativity on variables such as ethnicity, education level and income is an indication of social segregation. As, for larger networks, the assortativity can vary over the network and between nodes, it is of interest to calculate the assortativity locally. For each person node a local assortativity is calculated while addressing two practical problems. First we apply a normalisation to the local assortativity score to cope with imbalance in group sizes. Secondly, we address missing values in the group identity, which often is a practical problem. We demonstrate the procedures with a real dataset of the Dutch population.

Keywords: Assortativity · Segregation · Networks · Missing values

1 Introduction

In networks there is a tendency for nodes to connect to nodes with similar attributes. For example, sharing hobbies and interest increases the amount of contact with friends [5] and in many countries people tend to have more contact with persons of the same ethnic background [2]. Assortativity in a demographic social network can be seen as a measure for the segregation of its members. In social science segregation in ethnicity, education level or income often is of interest because it is an important indicator for social inequality.

For larger networks assortativity will not be homogeneous over the complete network. For some subgroups the assortativity will be higher or lower and also between individual nodes the assortativity can vary. Therefore, it is of interest to measure assortativity per node. Assortativity at higher levels can be obtained by aggregating the assortativity of nodes.

We will be building on the paper by Ballester *et al.* [1] and combine it with results from Peel *et al.* [7]. Both papers introduce local assortativity measures based on a local random walk. First, we will harmonize the concepts of the two papers. The first paper focuses on ethnic segregation; the second focuses on the traditional degree assortativity. The method can be used to define a locally weighted ego network from which all kinds of local summaries can be

computed. Second, we will discuss how to normalise the assortativity measures, which is necessary to make comparisons between groups. Third, we will discuss how to handle missing values in the node attributes. In our case, we would like to estimate segregation on education level. However, for approximately forty percent of our population we do not know the education level.

We will apply and test the methods on a social network of the complete Dutch population. As social network structures are very important for many social and economic processes, we have been working at our institute on deriving networks for the complete Dutch population. The nodes of our network are all persons that were registered in the official population register on October 1st 2014. The additional sources are used to define the edges: family relations, household membership, neighbours, co-workers and children going to same school (and being of the same age). The resulting network contains 16.9 million nodes and 764.2 million edges [3].

2 Methods

2.1 Local Assortativity

Our graph consists of a finite set $V = \{1, 2, \dots, n\}$ of individuals. We assume that there exist connections between individuals. We further assume that each relation has a weight that measures the strength of the relation. The weight of the relation between v and u is noted by e_{vu} ($e_{vu} \geq 0$). When e_{vu} equals zero there is no relation between v and u . Connections do not have to be symmetric: in general $e_{vu} \neq e_{uv}$ and $e_{vu} > 0$ does not imply that $e_{uv} > 0$. We will further assume that the outgoing weights of an individual add up to one:

$$\sum_u e_{vu} = 1. \quad (1)$$

The intuitive idea behind this is that each individual has a finite amount of resources (e.g. time) that it can spend on its connections. Then e_{vu} indicates the fraction of resources v spends on u .

A simple method of defining a local assortativity measure for node v is to look only at the direct neighbours of v . However, this has some disadvantages. First, this introduces issues with precision when the number of neighbours of v is small. Second, when looking for example at segregation, it is not only the direct neighbours of v that determine how segregated v is, but also how segregated those neighbours are i.e. the neighbours of those neighbours etc. Both Ballester *et al.* [1] and Peel *et al.* [7] introduced local assortativity measures based on a local random walk. The first is based on node-homogeneity of the egonetwork, and the second on edge-homogeneity. This paper is restricted to measures in node-homogeneity.

Starting from an individual v one starts to follow random connections. A connection from v to u is followed with a probability e_{vu} . After every step there is a probability $1 - \alpha$ that one stops at that point and a probability α that

one follows a random connection to the next node. The parameter α determines the size of the ego network. The average number of steps of random walk is $1/(1 - \alpha)$. Therefore, a smaller value of α means that a larger area around the starting node is taken into account. When α is zero only the direct neighbours are taken into account.

It was already assumed that the weights of the out-going connections sum to one: $\sum_u e_{vu} = 1$. Therefore, the $n \times n$ matrix \mathbf{E} with elements e_{vu} is a transition matrix. Each row gives the probabilities of transitioning in the next step to each of the other nodes. It can be shown [1] that the probabilities of individual v stopping at individual u , p_{vu} , are given by the matrix $\mathbf{P} = (p_{vu})_{v,u \in V}$, which is given by

$$\mathbf{P} = (1 - \alpha)(\mathbf{I} - \alpha\mathbf{E})^{-1}\mathbf{E}. \quad (2)$$

The rows of \mathbf{P} sum to one and for a row v the elements p_{vu} will be decrease for nodes u further away from v . These elements p_{vu} can therefore be used to calculate a locally weighted average of a property x_v of the nodes:

$$\tilde{x}_v^\alpha = \sum_u p_{vu}x_u, \quad (3)$$

using $\mathbf{x} = (x_1, x_2, \dots, x_n)'$, this can also be written as

$$\tilde{\mathbf{x}}^\alpha = \mathbf{P}\mathbf{x}. \quad (4)$$

In case x_v is a categorical variable ($x_v \in G$ with $G = \{1, 2, \dots, m\}$) indicating group membership of node v , we can define binary variables $x_v^{(g)} = \delta(x_v, g)$ with $x_v^{(g)}$ being one when v is member of g and zero otherwise. A group can be, for example, a set of individuals that share a given ethnicity or education level. The exposure of v to group g is then defined as:

$$e_v(g) = \sum_u p_{vu}x_u^{(g)}. \quad (5)$$

The exposure $i_v = e_v(x_v)$ of v to its own group is a measure for local assortativity. In the context of segregation this measure is also called isolation, and can be used as a measure of segregation for v .

There are a number of ways to calculate this index (Eq. (2) is impractical for large problems). One option is to make use of the fact that this score is related to the PageRank [6]. This is discussed further in Ballester *et al.* [1]. In practice it appears to be faster and more memory efficient to make use of the fact that \mathbf{P} can be written as an infinite series. Therefore, Eq. (3) can be written as

$$\tilde{\mathbf{x}}^\alpha = \mathbf{r}_g(\alpha) = \mathbf{P}\mathbf{x} = (1 - \alpha)\mathbf{Ex} + \alpha(1 - \alpha)\mathbf{E}^2\mathbf{x} + \alpha^2(1 - \alpha)\mathbf{E}^3\mathbf{x} + \dots \quad (6)$$

For $\alpha < 1$ these terms approach zero and the series can be limited to a finite number of terms.

The value of α determines the size of the ego-network. For small values of α the direct neighbours receive the highest weight, while for large values of α

the neighbourhood will consist of a large portion of the network. Peel *et al.* [7] suggest varying α from 0 to 1 and averaging the scores. Ballester *et al.* [1] use a value of 0.85 (which happens to be the same value often used in PageRank). Although they do not have a strong argument for this value, we follow Ballester *et al.* and use a fixed value of 0.85. Determining an optimal value of α is a subject for future research. For the conclusions of the paper, the exact value is not relevant.

2.2 Normalisation

A practical consideration in comparing local segregation scores is their normalisation. The local assortativity calculated is an individual score measuring the exposure of each individual to (members of) each group. If group sizes are unbalanced it is difficult to compare scores between groups. For example in case a randomly connected network consists for 95% out of nodes of group A and 5% of group B, the method would assign a high segregation score to members of group A and low to B, which is dubious since there is no assortativity mechanism in which similar nodes are more connected. There are simply more nodes of type A. The standard isolation or exposure makes it difficult to compare scores of the different groups. It is therefore sensible to normalize scores: is a score higher than one should expect? We will compare the raw local assortativity scores with three normalisation schemes.

First we adopt the normalisation used in [1] by dividing the isolation score by the relative group size:

$$i_v^d = \frac{i_v}{g_v} \quad (7)$$

with i_v the unnormalised isolation score and g_v the relative group size of the group to which v belongs. This normalisation reduces isolation scores for nodes of large groups, and enlarges isolation scores of small groups. Since $i_v^d \in [0, \inf]$, a down side of this normalisation is that the isolation score is not upper bounded and may exaggerate isolation for very small groups.

Alternatively we normalise the local isolation score of a node by subtracting by its group size and rescaling:

$$i_v^s = \frac{i_v - g_v}{1 - g_v} \quad (8)$$

with i_v the unnormalised isolation score and g_v the relative group size of the group to which v belongs. This normalisation reduces isolation scores for nodes belonging to a large group and has an upper bound of 1, but has no lower bound.

Third, Peel *et al.* [7] addresses the issue by calculating the local assortativity as the deviation of a global assortativity score. Instead of comparing the group fractions, the global assortativity score calculates the relative number of edges between groups and how it deviates from a randomly connected network with the same group size and degree distribution. The local assortativity score is then the deviation from the global assortativity score: how do the edges of the ego

network deviate from what you would expect from the global group sizes? In our experimental setup we made a slight alteration in which we compare the locally expected with the globally expected.

$$i_v^a = \frac{1}{\sum_g 1 - a_g^2} \sum_g (e_v(g)^2 - a_g^2) \quad (9)$$

with $e_v(g)$ the exposure of node v to group g within the ego-network, $a_g = \frac{k_g}{2m}$ the number of nodes of type g divided by the number of edges. The normalization constant in front of the sum assures that the assortativity is bounded: $-1 < a_v < 1$.

2.3 Missing Values

In this section we will consider the case where some of the values of \mathbf{x} are missing. We assume that the complete graph is still known. In our case, we don't know the education level of approximately forty percent of our population. Other examples would be online social network data where the network is known, but the ethnicity, education level or other property of interest is only known for a subset of the nodes in the network. Often the missingness of the values is not random, but is often directly or indirectly related to the variable of interest. In our case, for example, the number of missing values in education level increases for lower education levels and higher age.

One solution that is often used in survey statistics is weighting [8]. Survey weights are inversely proportional to the probability that an element is included in the survey (is observed). Observations on elements can be missing because it is a design choice to not include all elements (e.g. because of cost), or because observation was not possible (e.g. because element refused to participate). Often it is a combination of the two. In the first case, the inclusion probabilities are generally known (we know the selection process). In the second case, inclusion probabilities are generally estimated using other variables that are available (at the aggregated level) for the complete population and for the observed elements. For example, when the population consists of fifty percent men and fifty percent women, and the observed percentages in the data set are forty and sixty percent, we can derive that women have a 1.5 times higher probability of being observed than men. Therefore, using variables known for all nodes in the network, weights $\mathbf{w} = (w_1, w_2, \dots, w_n)'$ are calculated for the nodes with missing variables on the target variable \mathbf{x} .

As our local weighted average is already a weighted average of our target variable, we can simply add the survey or non-response weight directly into to the equation. Therefore Eqs. (3) and (4) become

$$\tilde{x}_v^\alpha = \sum_u p_{vu} w_u x_u, \quad (10)$$

and

$$\tilde{\mathbf{x}}^\alpha = \mathbf{P} \mathbf{W} \mathbf{x}, \quad (11)$$

respectively, with \mathbf{W} a diagonal matrix with the elements of \mathbf{w} on the diagonal.

3 Application

3.1 Data

Using data available at our institute a network was derived for the complete Dutch population. The nodes in our network consist of persons registered in the population register at October 1st 2014. This results in 16.9 million nodes.

The population register also registers the parents of each person. From this we were able to reconstruct other family relations: sibling, grand-parent, grand-child, uncle/aunt, niece/nephew, cousin. From the household register it is possible to determine household members and partners (married or not). Neighbours are the 10 closest households within 50 m. Data from the tax office enabled us to derive colleague-colleague relations. In order to limit the size of the colleagues network and because it is unlikely that someone knows all of its colleagues in very large companies, the number of colleagues was limited to a maximum of 100 (colleagues were randomly sampled otherwise). Finally, data on education enabled us to derive (a proxy for) class-mates: persons going to the same school and being of the same birth year. Table 1 shows the number of relations of each type in the resulting network.

Table 1. Number of edges in the social network of the Netherlands.

Relation	Count ($\times 1\ 000$)	Relation	Count ($\times 1\ 000$)
Grand parent	13 080	Cousin	83 213
Child	18 703	Household member	12 383
Grand child	13 068	Partner	7 786
Niece/nephew	39 392	Neighbour	195 131
Aunt/uncle	39 393	Colleague	118 105
Sibling	28 894	Schoolmate	176 318
Total			764 209

One of the goals of our research is to measure local assortativity on a number of variables related to social segregation: ethnicity, education level and income.

3.2 Normalisation

An important and often used segregation variable is ethnicity, which is available in the network we derived. We compare three different normalization schemes for the local assortativity score for the different ethnic groups.

Figure 1 shows the average local isolation scores per ethnic group. The right chart is identical to the left, but is truncated at isolation score 1, because i_v^d runs up to 10 and has extreme average scores as shown in the left. i_v^a is the most balanced, which is due to the calculation of the score: it weights in the

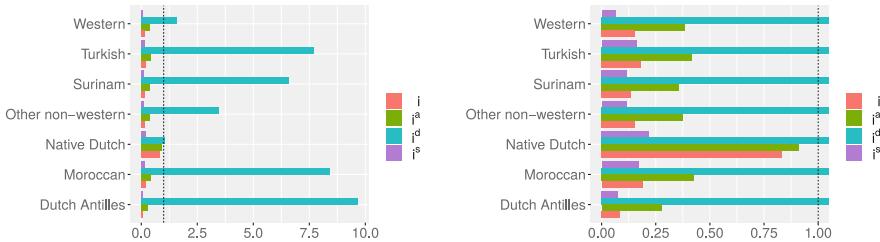


Fig. 1. Average local isolation scores per ethnicity group, i_v , i_v^d , i_v^s and i_v^a , rhs, same graph, but clipped to scores [0,1]

exposure to all groups instead of only the exposure to the own group. i_v^s reduces the average scores to below 0.3.

If a spatial average per region is taken as shown in Fig. 2, i_v and i_v^s are similar, with the last more spatially smooth. This spatial distribution seems plausible since the areas with a high isolation score are rural areas and the areas with low isolation score are urban. The spatial i_v^d however assigns high isolation scores to the major urban areas and seems contradictory. The explanation is that these areas have neighborhoods with high isolation score i_v^d and since these scores are more extreme as seen in Fig. 1, the average isolation score is affected. Therefore i_v^d is less suitable for aggregating individual scores.

We note that segregation often is analysed spatially and a network approach of calculating segregation scores adds value, by including relationships that are not spatially local. An interesting point for further investigation is that the spatial neighbourhood can also be used as a normalisation context.

3.3 Missing Values

One of the goals of our research is to measure local assortativity on a number of variables related to social segregation: ethnicity, education level and income. Unfortunately education level is not observed for approximately 40% of the population. Furthermore, the missingness is not random. For example, there are more missing values for older persons, lower educated persons and immigrants. The other two variables are observed for the complete population. Fortunately, pre-determined weights for education level are available based on a large number of background variables available for the complete population. In order to test the weighting method introduced in Sect. 2.3, missing values were introduced in the ethnicity variable for records that have a missing value for education level. Three types of estimates were then calculated. First, estimates were obtained using the data without missing values. This is the benchmark to which the other estimates are compared. Second, the weights available for education level were then used in Eq. (11) to obtain estimates for the exposure of individuals to the different ethnic groups (see (5)). Third, an unweighted estimate was obtained by using the same weight for all observations. This was done to see if weighting indeed improves the estimate compared to not weighting. Table 2 shows the sizes

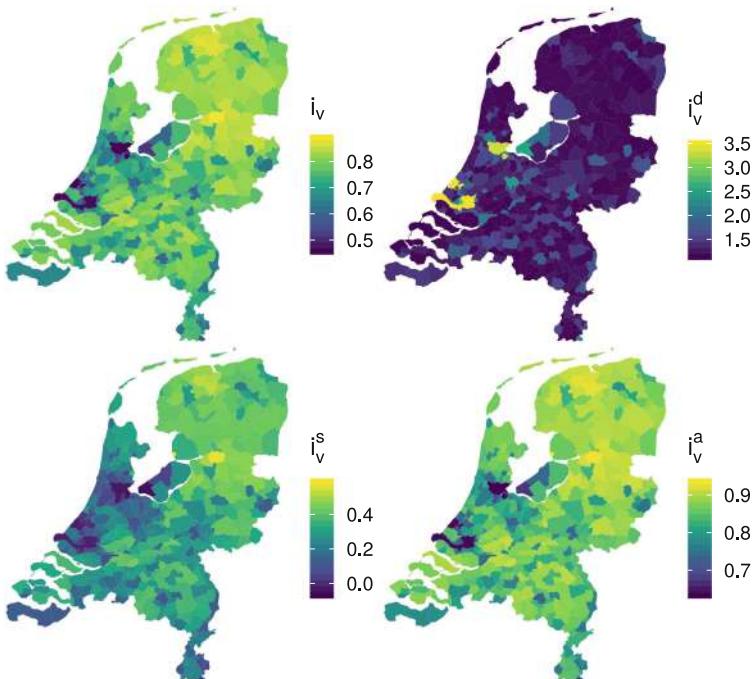


Fig. 2. Spatial comparison of i_v , i_v^d , i_v^s , i_v^a .

of each of the ethnic groups and the fraction of missing values introduced in each of the groups. As can be seen in the table, this fraction differs for each of the groups. This is among other things, related to the age distributions, as education is more often missing for older persons.

Local exposure values were calculated for each individual using the three methods: original, weighted and unweighted. The difference between the estimates and the original values is the error at the individual level. The individual values were also aggregated to the municipal level. The differences between the aggregates and the aggregate of the original values are the errors municipality level. Table 3 shows the mean of these errors and the spread in these errors. Looking at the mean error for the weighted estimate, we see that the bias is very small (relative error is less than 2%). The unweighted estimates have a substantial bias, indicating that weighting does help in removing the bias. As expected, the variances of the estimates increase when missing values are introduced. The variances of the weighted estimates are higher than the unweighted estimates. This is a common aspect of weighting [8].

Table 2. Size of each of the ethnic groups and fraction of missing values introduced in each of the groups.

Ethnicity	Size ($\times 1\ 000$)	Fraction missing
Native Dutch	13 238	40.2%
Moroccan	379	24.0%
Turkish	396	28.0%
Surinam	349	28.4%
Dutch Antilles	149	19.2%
Other non-western	754	28.2%
Western	1 622	45.4%
Total	16 887	39.1%

Table 3. Comparison of estimated exposure for ethnicity with missing values introduced to values obtained without missing values.

Ethnicity	Relative errors (%)							
	Weighted				Unweighted			
	Municip.		Indiv.		Municip.		Indiv.	
	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
Native Dutch	-0.2	1.2	-0.1	5.0	-1.0	1.6	-1.3	3.1
Moroccan	1.8	10.9	0.6	50.9	21.7	25.3	19.4	33.7
Turkish	1.7	8.4	0.7	54.9	16	18.1	14.1	28.7
Surinam	1.3	7.3	0.5	52	14.5	16.8	12.5	27.6
Dutch Antilles	1.2	18.1	0.9	67.9	26.1	32.9	24.7	50.7
Other non-western	1.2	17	0.8	51.8	15.9	20.4	14.2	32.2
Western	0.3	9.8	0.2	38.3	-9.3	12.7	-9.3	21.9

3.4 Local Assortativity of Education Level

Figure 3 shows the distribution of the exposure of member of each education level to members of the same education level that is isolation. As was mentioned in Sect. 2.1 this is a measure of segregation. The normalised local isolation is generally quite high (close to one). This indicates that members of each group mainly have members of their own group in their ego network. However, there are substantial differences between different persons. For each group the isolation can vary from approximately 0 to 1.

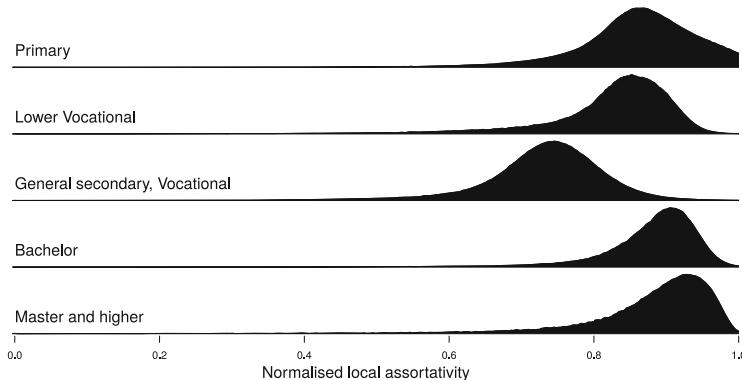


Fig. 3. Weighted density distribution of normalised local isolation for each education level.

4 Conclusion

Measuring assortativity at the individual level has a number of advantages to measuring it at the global level. First and most important, is the fact that for most variables of interest there will be substantial differences in the assortativity between individuals and sub networks. This can be seen in the results for education level where the isolation can vary between 0 and 1, and is even more pronounced when looking at ethnicity where the spread is even higher [3,4]. Second, a local measure makes it possible to look at assortativity at different levels by aggregating the individual values. Third, related to the previous, this makes it possible to study the relationship between individual assortativity and various other processes in the network. For example, we see substantial differences in the isolation values for education level. It would be interesting to investigate what processes explain these differences and to see what effect these differences have on other processes (e.g. social status).

We tried to unify existing measures and presented methods for handling missing values in the variables of interest and normalising the scores to make them comparable between groups. It is not always possible to measure all variables of interest for the complete population. However, we demonstrated with education levels that using weighting it is possible to obtain unbiased estimates, enabling local assortativity measures for a much larger set of variables. Other examples of missing variables are online social networks, which often have partially missing characteristics, e.g. gender, nationality or language. As the exposure depends on the sizes of the groups, it is necessary to control for the group sizes. The i_v^s measure we presented should make it possible to make comparisons between groups of different sizes.

As was mentioned above, being able to calculate local assortativity values for each member of the population, makes it possible to study the relationship between the ego network and various other social processes.

References

1. Ballester, C., Vorsatz, M.: Random walk-based segregation measures. *Rev. Econ. Stat.* **96**, 383–401 (2014)
2. Curry, O., Dunbar, R.I.M.: Do birds of a feather flock together? *Hum. Nat.* **24**(3), 336–347 (2013)
3. van der Laan, D.J., de Jonge, E.: Producing official statistics from network data. Paper presented at The 6th International Conference on Complex Networks and Their Applications (2017)
4. van der Laan, D.J., de Jonge, E.: Measuring segregation using a network of the Dutch population. Paper presented at The 5th International Conference on Computational Social Science (2019)
5. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* **27**(1), 415–444 (2001)
6. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. In: Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, pp. 161–172 (1998)
7. Peel, L., Delvenne, J.C., Lambiotte, R.: Multiscale mixing patterns in networks. *Proc. Nat. Acad. Sci.* **115**(16), 4057–4062 (2018)
8. Särndal, C.E., Swensson, B., Wretman, J.: Model Assisted Survey Sampling. Springer, Heidelberg (2003)



The Case for Kendall’s Assortativity

Paolo Boldi and Sebastiano Vigna^(✉)

Dipartimento di Informatica, Università degli Studi di Milano, Milan, Italy
{paolo.boldi,sebastiano.vigna}@unimi.it

Abstract. Since the seminal work of Litvak and van der Hofstad [12], it has been known that Newman’s assortativity [14, 15], being based on Pearson’s correlation, is subject to a pernicious size effect which makes large networks with heavy-tailed degree distributions always unassortative. Usage of Spearman’s ρ , or even Kendall’s τ was suggested as a replacement [6], but the treatment of ties was problematic for both measures. In this paper we first argue analytically that the tie-aware version of τ solves the problems observed in [6], and we show that Newman’s assortativity is heavily influenced by tightly knit communities. Then, we perform for the first time a set of large-scale computational experiments on a variety of networks, comparing assortativity based on Kendall’s τ and assortativity based on Pearson’s correlation, showing that the pernicious effect of size is indeed very strong on real-world large networks, whereas the tie-aware Kendall’s τ can be a practical, principled alternative.

1 Introduction

Assortativity (or *assortative mixing*) is a property of networks in which similar nodes are connected. More in detail, here we consider the *degree assortativity* of (directed) networks, that looks at whether the indegree/outdegree of x is correlated to the indegree/outdegree of y over all the arcs $x \rightarrow y$ of the network. One has thus four types of assortativity (denoted by $+/-$, $-/+$, $-/-$ and $+/-$), and for each type one has to choose which measure of correlation should be used between the lists of degrees at the start/end of each arc. The classical definition of assortativity given by Newman [14, 15] employed *Pearson’s correlation*.

In a seminal paper, Litvak and van der Hofstad [12] have shown analytically that on pathological examples and on heavy-tailed undirected networks Pearson’s degree-degree assortativity tends to zero as the network gets large, because of a size effect induced by the denominator of the formula. They go on to suggest to use of Spearman’s ρ (which is Pearson’s correlation on the rank of the values) to correct this defect.

A subsequent paper [6] has shown that the same problem plagues Pearson’s degree-degree correlation in the directed case: the authors explore also briefly, besides the ρ coefficient, the possibility of using Kendall’s τ , but they do not completely resolve the problem of ties. In their work, they suggest to use averages or randomization to correct for the presence of ties in the case of Spearman’s ρ ,

and simply neglect them in the case of τ , noting that this choice constraints (in a negative way) the possible values that τ can assume.

In this paper, first we extend analytically some of the results above, showing that the version of Kendall's τ that takes ties into consideration (sometimes called τ_b) has the same (better) behavior as Spearman's ρ on the pathological examples, and thus does not suffer from the limitations highlighted in [6]. Moreover, we show that tightly knit communities can influence in pernicious ways assortativity. Then, we perform several computational experiments on various web graphs and other types of complex networks, computing both Pearson's and Kendall's degree-degree correlations: by looking inside the data, we confirm the bias in the former when large networks are involved, and also give empirical evidence of the effect of tightly knit communities.

All data used in this paper are publicly available from the LAW, and the code used for the experiments is available as free software as part of the LAW Library.¹

2 Definitions and Conventions

In this paper, we consider directed graphs defined by a set N of n nodes and a set $A \subseteq N \times N$ of arcs; we write $x \rightarrow y$ when $a = \langle x, y \rangle \in A$ and call x (respectively, y) the source (respectively, target) of the arc a , denoted by $s(a)$ (respectively, $t(a)$). Our graphs can contain *loops* (arcs a such that $s(a) = t(a)$). A *successor* of x is a node y such that $x \rightarrow y$, and a *predecessor* of x is a node y such that $y \rightarrow x$. The *outdegree* $d^+(x)$ of a node x is the number of its successors, and the *indegree* $d^-(x)$ is the number of its predecessors.

A *symmetric graph* is a graph such that $x \rightarrow y$ whenever $y \rightarrow x$; such a graph can be identified with an undirected graph, that is, a graph whose arcs (usually called *edges*) are subsets of one or two nodes. In fact, in the following, all definitions are given for directed graphs, and apply to undirected graphs through their loopless symmetric representation.

3 Assortativity

Degree-degree assortativity measures the propensity of nodes to create links to nodes with similar degrees. Since we consider directed graphs, there are four types of assortativity: outdegree/outdegree, indegree/outdegree, indegree/indegree, and outdegree/indegree, denoted by $+/-$, $-/+$, $-/-$, and $+/+$, respectively. As noted in [6], the only case in which an arc contributes to the degrees on both of its sides is $+/-$, which makes the $+/-$ case the natural generalization of the undirected case [12]. The assortativity is defined as a measure of correlation between the appropriate list of degrees at the two sides of the list

¹ <http://law.di.unimi.it/>.

of all arcs of the graph. More precisely, for any given correlation index c , and every choice of $\alpha, \beta \in \{+, -\}$, we define the c -assortativity of type (α, β) as

$$c_\alpha^\beta(G) = c \left([d^\alpha(s(a))]_{a \in A}, [d^\beta(t(a))]_{a \in A} \right)$$

where $[-]_{a \in A}$ are used to denote a list ranging over all arcs of the graph (the order is immaterial, provided that it is coherent, i.e., the same for both lists).

Newman's definition of assortativity [14, 15] uses Pearson's correlation coefficient for c . However, there are many other possibilities to measure the correlation between degrees. One can use, as an alternative, Spearman's ρ [16], which is Pearson's correlation between the *ranks* of the values in the lists: this choice has some advantages, but it does not provide a solution for *ties*, that is, duplicate degrees.

Correct handling of ties in degree lists is of utter importance because in a real-world graph a large percentage of the arcs is involved in a tie (e.g., the outdegree of the target of all arcs pointing at the same large-indegree node are the same). In [6], the authors resort to typical solutions such as averaging or randomization, which however have been shown to be detrimental and, in fact, decrease the amount of correlation [18].

An alternative that handles ties in a principled way is Kendall's τ correlation index, when formulated properly (see next section). However, the authors of [6] used a formulation that had been designed for lists without ties, obtaining pathological results. We are going to show that this problem can be fixed using a proper version, thus providing more proper handling of degree ties.

4 Kendall's τ , 1945

The original and most commonly known definition of Kendall's τ is given in terms of concordances and discordances. We consider two real-valued vectors \mathbf{r} and \mathbf{s} (to be thought of as scores) of n elements, and assume that no score appears twice in a vector (i.e., there are no *ties*). We say that a pair of indices $\langle i, j \rangle$, $0 \leq i < j < n$, is *discordant* if $s_i < s_j$ and $t_i > t_j$, or $s_i > s_j$ and $t_i < t_j$, *concordant* otherwise. Then, the τ between the two vectors is given by the number of concordances, minus the number of discordances, divided for the number of pairs. Note that all scores must be distinct.

In his 1945 paper about ranking with ties [7], Kendall, starting from an observation of Daniels [4], reformulates his correlation index using a definition similar in spirit to that of an inner product. Let us define

$$\langle \mathbf{r}, \mathbf{s} \rangle := \sum_{i < j} \text{sgn}(r_i - r_j) \text{sgn}(s_i - s_j),$$

where

$$\text{sgn}(x) := \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

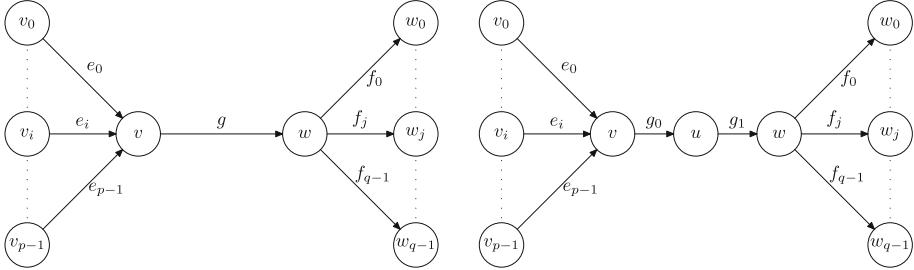


Fig. 1. The graphs $G(p, q)$ and $\hat{G}(p, q)$.

Note that the expression above is actually an inner product in a larger space of dimension $n(n-1)/2$: each score vector \mathbf{r} is mapped to the vector with coordinate $\langle i, j \rangle$, $i < j$, given by $\text{sgn}(r_i - r_j)$. Thus, we can define

$$\tau(\mathbf{r}, \mathbf{s}) := \frac{\langle \mathbf{r}, \mathbf{s} \rangle}{\sqrt{\langle \mathbf{r}, \mathbf{r} \rangle} \cdot \sqrt{\langle \mathbf{s}, \mathbf{s} \rangle}}. \quad (1)$$

Essentially, we are defining a cosine similarity, which we can compute easily as follows: given a pair of distinct indices $0 \leq i, j < n$, we say that the pair is

- *concordant* iff $r_i - r_j$ and $s_i - s_j$ are both nonzero and have the same sign;
- *discordant* iff $r_i - r_j$ and $s_i - s_j$ are both nonzero and have opposite signs;
- a *left tie* iff $r_i - r_j = 0$;
- a *right tie* iff $s_i - s_j = 0$.

Let C, D, T_r, T_s be the number of concordant pairs, discordant pairs, left ties, right ties, and joint ties, respectively. We have

$$\tau(\mathbf{r}, \mathbf{s}) = \frac{C - D}{\sqrt{\binom{n}{2} - T_r} \sqrt{\binom{n}{2} - T_s}}.$$

Note that a pair that is at the same time a left tie and a right tie (a so-called *joint tie*) will be counted both in T_r and in T_s .

5 The Case for Ties in Kendall's Assortativity

As an example of the importance of ties in the computation of Kendall's assortativity, we consider the graphs $G(p, q)$ and $\hat{G}(p, q)$ defined in [6, Sect. 5.1] and shown in Fig. 1. The graphs are made by a directed 1-path or 2-path (whose arcs we will call g , or g_0 and g_1 , respectively), with p further nodes v_0, v_1, \dots, v_{p-1} pointing to the source of the path (the p corresponding arcs are named e_0, e_1, \dots, e_{p-1}) and q nodes w_0, w_1, \dots, w_{q-1} pointed by the target of the path (the q corresponding arcs are named f_0, f_1, \dots, f_{q-1}). Overall, the graph has $G(p, q)$ has $p + q + 1$ arcs, whereas $\hat{G}(p, q)$ has $p + q + 2$ arcs.

On these graphs, Pearson's assortativity of type $-/+$ behaves in a completely pathological way: for $G(n, an)$, it tends to 1 as $n \rightarrow \infty$, because the mass associated with the arc g becomes very large, due to its very large indegree, whereas the assortativity of $\hat{G}(n, an)$ tends to 0 [6]. This is extremely counterintuitive, because the two networks are almost identical. In particular, in both cases one expects to measure a significant disassortativity², because a large fraction of arcs are disassortative (both the e_i 's and the f_j 's are such).

Using Spearman's ρ makes assortativity correctly tend to -1 in both cases if one solves ties by giving equal rank to equal elements; one has, however, widely different results with different methods for ranking ties.

The limit of Kendall's τ as $n \rightarrow \infty$ without taking ties into consideration is $-2a/(a+1)^2$, which tends to zero as a grows. The authors comment that this is due to the influence of ties, and indeed we are going to show that using the correct tie-aware version Kendall's τ solves the problem: the network becomes disassortative, as the fraction of concordant pairs goes to zero whereas the fraction of discordant pairs does not. This happens naturally, without having to choose a policy for ties.

Looking again at the $-/+$ assortativity, we see that in both graphs there are p arcs (the arcs e_i) with degree pairs $\langle 0, 1 \rangle$ and q arcs (the arcs f_j) with degree pairs $\langle 1, 0 \rangle$. Finally, in $G(p, q)$ we have one arc with degree pair $\langle p, q \rangle$, whereas in $\hat{G}(p, q)$ we have one arc with degree pair $\langle p, 1 \rangle$ and one arc with degree pair $\langle 1, q \rangle$. We will assume $2 < p < q$ in the following.

In $G(p, q)$ we have:

- $p + q$ concordances, given by the pairs $\langle e_i, g \rangle$ and $\langle f_j, g \rangle$.
- pq discordances, given by the pairs $\langle e_i, f_j \rangle$.
- $\binom{p}{2} + \binom{q}{2}$ left ties, given by distinct pairs of e_i 's, and distinct pairs of f_j 's.
- $\binom{p}{2} + \binom{q}{2}$ right ties, given by the same pairs.

All in all, we thus have

$$\tau_{G(p,q)} = \frac{p + q - pq}{\sqrt{\left(\binom{p+q+1}{2} - \binom{p}{2} - \binom{q}{2}\right) \cdot \left(\binom{p+q+1}{2} - \binom{p}{2} - \binom{q}{2}\right)}}.$$

In $\hat{G}(p, q)$, by an analogous analysis we have

$$\tau_{\hat{G}(p,q)} = \frac{p + q - pq - 1}{\sqrt{\left(\binom{p+q+2}{2} - \binom{p}{2} - \binom{q}{2} - q\right) \cdot \left(\binom{p+q+2}{2} - \binom{p}{2} - \binom{q}{2} - p\right)}}.$$

² We use “unassortative” for networks with a correlation close to 0, and “disassortative” for networks with a correlation close to -1 .

If we consider the case $p = n$, $q = an$ with a constant and $n \rightarrow \infty$, as in [6], we have

$$\tau_{G(n,an)} = \frac{n + an - an^2}{\sqrt{\left(\binom{n+an+1}{2} - \binom{n}{2} - \binom{an}{2}\right) \cdot \left(\binom{n+an+1}{2} - \binom{n}{2} - \binom{an}{2}\right)}} \rightarrow -1$$

$$\tau_{\hat{G}(n,an)} = \frac{n + an - an^2 - 1}{\sqrt{\left(\binom{n+an+2}{2} - \binom{n}{2} - \binom{an}{2} - an\right) \cdot \left(\binom{n+an+2}{2} - \binom{n}{2} - \binom{an}{2} - n\right)}} \rightarrow -1$$

Thus, the proper definition aligns on this example Kendall's τ with the results from Spearman's ρ , using constant ranks for ties.³

6 The Tightly Knit Community Effect, Again

We are now going to discuss another, and possibly more pernicious, effect of size on Pearson's assortativity. This phenomenon is akin to the well-known *tightly knit community* (TKC) effect on certain ranking algorithms such as HITS [8]: a small group of tightly connected users ends up being ranked unfairly high. For this section, we consider undirected graphs, as it is much simpler to compute Pearson's assortativity using the formulae from [17], but the same considerations apply to directed graphs.

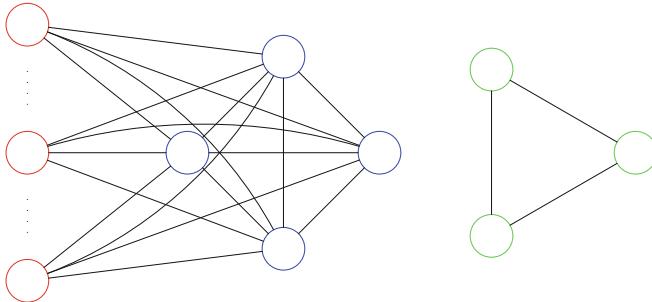


Fig. 2. The graph $H(p, q, k)$. There are p red nodes (left), q blue nodes (center) and k green nodes (right).

Let us start from the graph $H(p, q)$ defined as a (p, q) complete bipartite graph, in which the q nodes are further completely connected (i.e., they form a q -clique). This graph (the red and blue nodes in Fig. 2) is highly unassortative,

³ We mention that also Goodman–Kruskal's γ [5], defined as the difference between concordances and discordances divided by their sum, provides a principled treatment of ties. However, Kendall's τ has some advantages, and in particular the possibility of defining tie-aware weighted versions [18].

in the sense that its Pearson's assortativity, which is $p/(1-p-q)$, goes to zero if $q, p \rightarrow \infty$, as long as $p = o(q)$.

Let us now consider a graph $\hat{H}(p, q, k)$ formed by $H(p, q)$ plus a (disjoint) clique of k vertices (see Fig. 2). Our interest is in measuring how much the clique (the most assortative graph) will influence the unassortative graph $H(p, q)$. In particular, we will look at $\hat{H}(p, p^2, p^{1/2+\varepsilon})$: this is the very unassortative graph $H(p, p^2)$, with $p+p^2$ nodes, to which we are adding a clique of size $p^{1/2+\varepsilon}$ (note that the number of added nodes for small ε is negligible in the size of $H(p, p^2)$).

The formula for the Pearson's assortativity of $H(p, q, k)$ is

$$1 - \frac{pq(p-1)^2}{(k(k-1)^3 + pq^3 + qs^3 - \frac{1}{2pq+q(q-1)+k(k-1)}(pq^2 + qs^2 + k(k-1)^2)^2)},$$

where $s = q + p - 1$, and dominant term of $H(p, p^2, p^{1/2+\varepsilon})$ as $p \rightarrow \infty$ is

$$\frac{p^{2\varepsilon}}{1 + p^{2\varepsilon}}. \quad (2)$$

Thus, we have a threshold effect as $p \rightarrow \infty$: for $\varepsilon < 0$, the network becomes unassortative; for $\varepsilon = 0$, assortativity tends to $1/2$; but for $\varepsilon > 0$, the network will become completely assortative. In other words, *a tightly knight community of order $\Omega(n^{1/4+\varepsilon})$ can drive a large unassortative network to assortativity*. This impressive effect is evidently pathological.

Is there a similar phenomenon for Kendall's assortativity? The value of Kendall's assortativity on $H(p, q)$ is (maybe surprisingly) the same of Pearson's assortativity, whereas the Kendall's assortativity of $\hat{H}(p, q, k)$ is

$$\frac{k(k-1)(2p+q-1) - qp^2}{k(k-1)(2p+q-1) + pq(q+p-1)}.$$

When we examine $H(p, p^2, p^{1/2+\varepsilon})$ and let $p \rightarrow \infty$ we find a completely different situation, as assortativity tends to zero as $-1/p$. However, the leading term of $H(p, p^2, p^{3/2+\varepsilon})$ is again (2), and a similar transition effect appears.

In other words, *also Kendall's assortativity is subject to the TKC effect, but one needs a community asymptotically much larger to obtain the same effect*, that is, $\Omega(n^{3/4+\varepsilon})$ vs. $\Omega(n^{1/4+\varepsilon})$. As an example, the graph $H(100, 10000)$ has a (Pearson's and Kendall's) assortativity of -0.010 , but if look at $\hat{H}(100, 10000, 200)$ (i.e., we add a clique of 200 nodes, increasing the size of the graph by less than 2%) we get an impressive increase in Pearson's assortativity (it goes up to 0.997), whereas the Kendall's assortativity is still very small (0.029).

7 Experiments

We consider a set of networks available from the repository of the Laboratory for Web Algorithmics.⁴ The graphs are listed and briefly described in Table 1,

⁴ <http://law.di.unimi.it/datasets.php>.

Table 1. The graphs used in the experiments.

Name	Nodes	Arcs	
arabic-2005	22 744 080	639 999 458	A crawl of Arabic countries [1]
cnr-2000	325 557	3 216 152	A crawl of the CNR [1]
dblp-2010	326 186	1 615 400	The co-authorship graph from DBLP
dblp-2011	986 324	6 707 236	The co-authorship graph from DBLP
dewiki-2013	1 532 354	36 722 696	German Wikipedia
enwiki-2013	4 206 785	101 355 853	English Wikipedia
eswiki-2013	972 933	23 041 488	Spanish Wikipedia
frwiki-2013	1 352 053	34 378 431	French Wikipedia
eu-2005	862 664	19 235 140	A crawl of .eu [2]
gsh-2015-host	68 660 142	1 802 747 600	Host graph of a general crawl [2]
gsh-2015-tpd	30 809 122	602 119 716	Top domains of a general crawl [2]
hollywood-2009	1 139 905	113 891 327	The co-starship graph from the IMDB
hollywood-2011	2 180 759	228 985 632	The co-starship graph from the IMDB
hu-tel-2006	2 317 492	46 126 952	Call graph from Hungarian Telekom [10]
in-2004	1 382 908	16 917 053	A crawl of .in [1]
indochina-2004	7 414 866	194 109 311	A crawl of Indochina [1]
it-2004	41 291 594	1 150 725 436	A crawl of .it [1]
itwiki-2013	1 016 867	25 619 926	Italian Wikipedia
ljournal-2008	5 363 260	79 023 142	LiveJournal [3]
orkut-2007	3 072 626	234 370 166	The Orkut social network [13]
sk-2005	50 636 154	1 949 412 601	A crawl of .sk [1]
twitter-2010	41 652 230	1 468 365 182	Twitter [11]
uk-2002	18 520 486	298 113 762	A crawl of .uk [1]
uk-2005	39 459 925	936 364 282	A crawl of .uk [1]
uk-2014-host	4 769 354	50 829 923	Host graph of .uk [2]
uk-2014-tpd	1 766 010	18 244 650	Top domains of .uk [2]
webbase-2001	118 142 155	1 019 903 190	A crawl from the Stanford WebBase

which shows some of their basic properties: more information can be found on the repository website. The list includes a variety of types of graphs, both directed and undirected (i.e., symmetric), including web crawls, host graphs, Wikipedia graphs, social networks (e.g., Twitter), telephone-call graphs, and co-authorship/co-starship graphs; their sizes range from a few hundred thousands to billion of edges.

We computed assortativity values based on Pearson’s correlation and on Kendall’s τ using the implementations available in the LAW library, and we report the values in Table 2. Large differences (≥ 0.20) are shown in boldface.

First of all, we remark that we can confirm the results about Wikipedia graphs discussed in [6]: they are unassortative for all types. However, when we consider social networks such as LiveJournal, Twitter and Orkut, where we do

Table 2. Pearson's and Kendall's assortativity values for the graphs of Table 1. Bold-faced entries have a difference larger than 0.20. Note that all values for the undirected graphs (e.g., `orkut-2007`) are all identical.

Name	+/-		-/+		-/-		+/-	
	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
<code>hu-tel-2006</code>	-0.0063	0.0660	0.0037	0.0768	-0.0107	0.0349	-0.0418	0.0266
<code>ljournal-2008</code>	0.2665	0.2835	0.1919	0.2656	0.0508	0.2689	0.0674	0.2989
<code>twitter-2010</code>	-0.0301	0.0410	-0.0089	0.3232	-0.0121	0.1886	-0.0506	-0.1395
<code>orkut-2007</code>	0.0158	0.2528	0.0158	0.2528	0.0158	0.2528	0.0158	0.2528
<code>dblp-2010</code>	0.3300	0.2827	0.3300	0.2827	0.3300	0.2827	0.3300	0.2827
<code>dblp-2011</code>	0.1296	0.1757	0.1296	0.1757	0.1296	0.1757	0.1296	0.1757
<code>hollywood-2009</code>	0.3555	0.3278	0.3555	0.3278	0.3555	0.3278	0.3555	0.3278
<code>hollywood-2011</code>	0.2073	0.2585	0.2073	0.2585	0.2073	0.2585	0.2073	0.2585
<code>enwiki-2013</code>	-0.0715	-0.0542	-0.0007	0.0126	-0.0077	-0.0385	-0.0553	-0.1287
<code>frwiki-2013</code>	-0.0469	-0.0136	0.0028	0.0037	-0.0110	-0.0458	-0.0564	-0.0778
<code>dewiki-2013</code>	-0.0398	-0.0395	0.0052	0.0295	-0.0109	-0.0058	-0.0518	-0.0934
<code>eswiki-2013</code>	-0.0301	0.0078	-0.0054	-0.0244	-0.0261	-0.1774	-0.1049	-0.1183
<code>webbase-2001</code>	0.4005	0.2817	0.2635	0.1483	-0.0048	-0.0486	-0.0107	0.1234
<code>arabic-2005</code>	0.9350	0.5456	0.5378	0.2766	-0.0288	-0.0916	-0.0539	0.0406
<code>indochina-2004</code>	0.9965	0.6195	0.9837	0.5068	0.0333	0.1704	0.0332	0.2397
<code>eu-2005</code>	0.0832	0.2942	0.0394	0.1388	-0.0239	-0.1541	-0.0815	-0.0947
<code>in-2004</code>	0.3219	0.4761	0.2883	0.2865	-0.0458	-0.0722	-0.0925	0.0908
<code>it-2004</code>	0.9003	0.4083	0.3582	0.1790	-0.0113	-0.1033	-0.0191	0.0304
<code>sk-2005</code>	0.9534	0.3375	0.2177	0.1353	-0.0078	-0.1351	-0.0343	-0.0633
<code>uk-2002</code>	0.5083	0.4219	0.2755	0.1338	-0.0050	-0.1383	-0.0209	0.0679
<code>uk-2005</code>	0.8334	0.6246	0.0337	0.2010	-0.0031	-0.1364	-0.0818	0.1781
<code>cnr-2000</code>	-0.0439	0.1237	-0.0027	0.1079	-0.0317	-0.1850	-0.0986	0.0099
<code>itwiki-2013</code>	-0.0678	-0.0209	0.0034	0.0429	-0.0114	-0.0284	-0.0666	-0.0758
<code>uk-2014-host</code>	-0.0221	0.2792	-0.0093	0.2743	-0.0123	0.2005	-0.0296	0.2282
<code>gsh-2015-host</code>	0.0962	0.3674	0.5688	0.3547	-0.0297	0.2159	-0.0205	0.2558
<code>uk-2014-tpd</code>	-0.0406	0.0755	0.0128	0.1679	-0.0100	0.1695	-0.0376	0.0988
<code>gsh-2015-tpd</code>	-0.0092	0.2595	0.0544	0.3749	-0.0252	0.3471	-0.0243	0.3263

expect some kind of assortativity, Kendall's assortativity provides significant larger values, proving for the first time that on real-world networks the size effect is substantial, as it makes such networks appear unassortative according to Pearson's correlation.

In Fig. 3 we show a scatter plot of the values obtained by Pearson's correlation versus those obtained by Kendall's τ , sizing the dots depending on the number of arcs of the graph. In the $-/-$ and $+/-$ case one can see immediately the strip of small graphs on the diagonal showing correlation, and the pile of large graphs, all stationing around the value zero of Pearson's correlation.

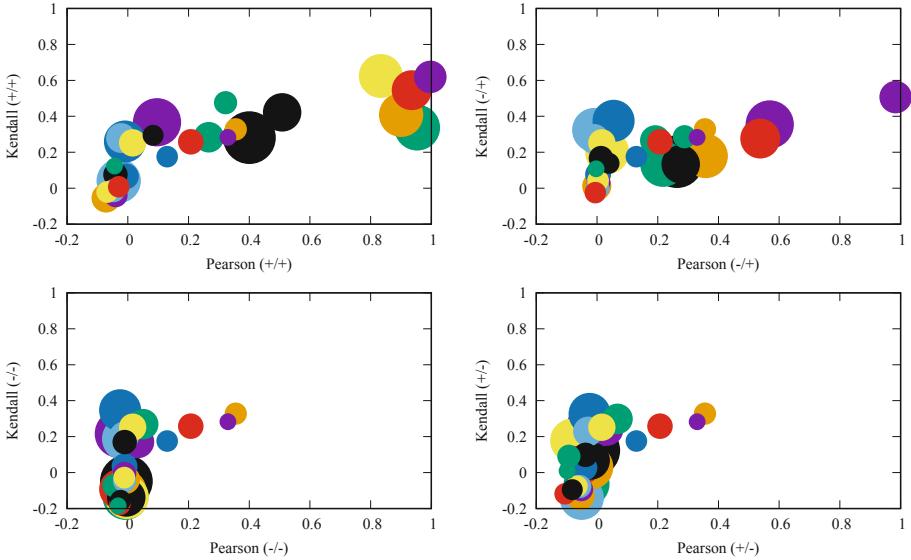


Fig. 3. Plots displaying the correlation between Pearson’s and Kendall’s assortativity.

There is of course another feature that is evident: several web graphs have incredibly large assortativity of type $+/+$ (e.g., `indochina-2004` has Pearson’s assortativity 0.9965 and Kendall’s assortativity 0.6195), which shows up as the block of large circles in the upper right part of the top left graph of Fig. 3. Is it really possible that web pages tend to link mostly to pages with the same number of links?

The answer is no: the preposterously high score we are observing are simply due to the TKC effect. Extremely connected sites (e.g., machine-generated tables or calendars) can have a very strong impact on assortativity. For example, the densest website of non-negligible size (probably a link farm) in `indochina-2004` contains 7 611 nodes and 48 231 874 arcs (a density of 83%!). To study the role of such dense websites, we can try to remove them from the graph (or, in the opposite direction, to add a fictitious large clique) and see how this operation impacts on assortativity. Table 3 shows the results for `indochina-2004`:

- The TKC effect is evident for $+/+$ and $-/+$, but the increase is more dramatic in Pearson’s than in Kendall’s assortativity (in line with the discussion of Sect. 6).
- The same observation holds for $-/-$ and $+/-$ but in this case the phenomenon in Pearson’s assortativity is diluted by the size effect: this web graph contains pages extremely large indegree and zero outdegree (simply because the crawl was stopped before the outlinks of those pages could be fetched); the arcs toward these pages contribute to a very large second component in $-/-$ and $+/-$ (whereas they do not show up in $+/+$ and $-/+$).

- Once we remove the noise, the Kendall $-/-$ values tend to *disassortativity*, which is actually what we expect from a web graph (highly pointed nodes of influential websites are, by and large, linked by “normal” nodes); in other words, the noise from the TKC and the size effects completely hide the actual disassortative nature of the network.

Other web crawls behave similarly: the take-home message here is that one should be very cautious when using assortativity values measured on noisy large-scale data such as web crawls, and that, in any case, Kendall's τ is more robust and less sensitive to the TKC and size effects. In fact, computing *both* measures is an excellent way to spot anomalous substructures in a network.

Table 3. Pearson's and Kendall's assortativity values for variants of the graph G of *indochina-2004*. Here H_1 and H_2 are the densest and second-densest non-negligible websites of this web graph, whereas K_t is a clique of size t .

	$G - H_{1,2}$		$G - H_1$		G		$G + K_{1000}$		$G + K_{20\,000}$	
	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
$+/-$	0.5659	0.4167	0.7784	0.4342	0.9965	0.6195	0.9965	0.6233	0.9998	0.8026
$-/+$	0.2070	0.2005	0.3622	0.2252	0.9837	0.5068	0.9837	0.5118	0.9993	0.7917
$-/-$	-0.0200	-0.1344	-0.0269	-0.1262	0.0333	0.1704	0.0337	0.1717	0.5596	0.7075
$+/-$	-0.0672	0.0027	-0.0665	0.0063	0.0332	0.2397	0.0335	0.2403	0.5597	0.7112

8 Conclusions

We have discussed important, practical shortcomings of measures of degree-degree correlation, in particular Newman's assortativity, when applied to large networks. We believe that using Kendall's τ in place of Pearson's correlation might mitigate parts of the problems. More theoretical analysis and experiments are however necessary to understand in detail the sensitivity of these measures to small locally dense graphs. Kendall's τ requires some more computational effort, that is, $O(m \log m)$ (where m is the number of arcs) rather than the $O(m)$ time of Spearman's and Pearson's correlation. However, there are $O(m \log m)$ algorithms based on sorting [9] that are easily parallelized or distributed among multiple computational units, which should help to mitigate the problem.

References

- Boldi, P., Codenotti, B., Santini, M., Vigna, S.: UbiCrawler: a scalable fully distributed web crawler. *Softw. Pract. Exp.* **34**(8), 711–726 (2004)
- Boldi, P., Marino, A., Santini, M., Vigna, S.: BUbiNG: Massive crawling for the masses. *ACM Trans. Web* **12**(2), 12:1–12:26 (2019)

3. Chierichetti, F., Kumar, R., Lattanzi, S., Mitzenmacher, M., Panconesi, A., Raghavan, P.: On compressing social networks. In: KDD 2009: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 219–228. ACM, New York (2009)
4. Daniels, H.E.: The relation between measures of correlation in the universe of sample permutations. *Biometrika* **33**(2), 129–135 (1943)
5. Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. *J. Am. Stat. Assoc.* **49**(268), 732–764 (1954)
6. van der Hoorn, P., Litvak, N.: Degree-degree dependencies in directed networks with heavy-tailed degrees. *Internet Math.* **11**(2), 155–179 (2015)
7. Kendall, M.G.: The treatment of ties in ranking problems. *Biometrika* **33**(3), 239–251 (1945)
8. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *J. ACM* **46**(5), 604–632 (1999)
9. Knight, W.R.: A computer method for calculating Kendall's tau with ungrouped data. *J. Am. Stat. Assoc.* **61**(314), 436–439 (1966)
10. Kurucz, M., Benczur, A., Csalogany, K., Lukacs, L.: Spectral clustering in telephone call graphs. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, WebKDD/SNA-KDD 2007, pp. 82–91. ACM (2007)
11. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 591–600. ACM (2010)
12. Litvak, N., van der Hofstad, R.: Uncovering disassortativity in large scale-free networks. *Phys. Rev. E* **87**, 022801 (2013)
13. Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, pp. 29–42. ACM (2007)
14. Newman, M.E.J.: Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002)
15. Newman, M.E.J.: Mixing patterns in networks. *Phys. Rev. E* **67**, 026126 (2003)
16. Spearman, C.: The proof and measurement of association between two things. *Am. J. Psychol.* **15**(1), 72–101 (1904)
17. Van Mieghem, P., Wang, H., Ge, X., Tang, S., Kuipers, F.A.: Influence of assortativity and degree-preserving rewiring on the spectra of networks. *Eur. Phys. J. B* **76**(4), 643–652 (2010)
18. Vigna, S.: A weighted correlation index for rankings with ties. In: Srinivasan, S., Ramamritham, K., Kumar, A., Ravindra, M.P., Bertino, E., Kumar, R. (eds.) Proceedings of the 24th International Conference on World Wide Web, pp. 1166–1176. ACM (2015)

Modeling Human Behavior



Modelling Opinion Dynamics and Language Change: Two Faces of the Same Coin

Jérôme Michaud^(✉)

Department of Sociology, Uppsala University,
Thunbergsvägen 3H, 752 38 Uppsala, Sweden
jerome.michaud@soc.uu.se

Abstract. While opinion dynamics models and language change models seem to be very different at first sight, we demonstrate in this paper that they are more similar than they look like. Here, we analyse the similarities and differences between the Social Influence with Recurrent Mobility model (Phys. Rev. Lett. **112**, 158701) and the Utterance Selection Model of language change (Phys. Rev. E **73**, 046118), two models that simulate a stochastic dynamics on a complex network, and show that their mesoscopic dynamic is strikingly similar. By drawing an analogy between the two models, we discuss possibilities of cross-fertilization of research between the two problems.

Keywords: Social Influence with Recurrent Mobility · Utterance Selection Model · Stochastic dynamics on network · Mesoscopic approximation · Stochastic differential equation · Multiplicative noise

1 Introduction

At a first sight, opinion dynamics and language change seem to be very different topics and similarities between them are not evident. This paper aims at exemplifying such similarities between these two problems by looking at stochastic models of opinion dynamics and language change on complex networks. The analysis justifies to make an analogy between the two problems and enables a reinterpretation of the dynamics of both problems. As a result, the models for opinion dynamics and language change can be seen at two faces of the same modelling coin and this duality can be utilized to improve both models.

The chosen model for opinion dynamics is the Social Influence with Recurrent Mobility model [5, 10], which is a generalization of the well-studied Voter Model [7]¹ and the chosen model for language evolution is the Utterance Selection Model (USM) for language change [1], which is a model grounded in evolutionary

¹ But note that the SIRM model relaxes some of the problematic assumptions of the Voter Model. Errors can occur in the copying processes and unilateral change of opinion is possible.

theory and inspired by the Wright-Fischer model for population genetics [6, 12]. These two models will be shown to have a very similar dynamics, at least at some mesoscopic scale.

This paper is organized as follows. Section 2 introduces the two studied models and their respective mesoscopic approximation that takes the form of a stochastic differential equation (SDE) with multiplicative noise. Section 3 discusses the similarities and differences between the two models. Section 4 provides some concluding remarks and suggests directions in which the models studied could be improved.

2 Two Stochastic Models of Opinion Dynamics and Language Change on Networks

The two models we are analyzing in this paper are stochastic models on networks whose nodes can be partitioned into a finite number of groups. The level at which the models will be compared is at this group level, where the mesoscopic dynamics take a very similar form.

The SIRM and the USM share the property of having two sources of stochasticity: the standard stochasticity originating from random ordering of pairwise interactions and another level of stochasticity originating in imperfect interactions. One specificity of this second source of stochasticity is that it survives a (heterogeneous-) mean-field treatment, leading to a stochastic dynamics of the mesoscale quantities.

2.1 The Social Influence with Recurrent Mobility (SIRM) Model

The SIRM model [5] is a generalization of the well-studied Voter Model [7] that considers the evolution of the opinions of a population partitioned into geographic/administrative regions. The coupling between the regions is encoded using the commuting pattern of the individuals, i.e., using the number of people living in one region and working in another. It has been applied to US presidential elections [5] and to Swedish parliamentary elections [10]. The formulation we are using is that of [10], where some issues in the original mathematical formulation of the SIRM have been fixed.

Network Structure. Consider a population of N individuals divided into M regions usually representing electoral regions, such as counties, municipalities or states. Let N_{ij} be the number of commuters between regions i and j , that is the number of individuals living in region i and working in region j . The commuting network is constructed by taking the M municipalities as nodes and linking them by weighted edges of weight N_{ij} . The commuting network is a directed weighted network with self-loops, since there are people who live and work in the same place. In the rest of the paper, we will refer to the sub-population N_{ij} as the *commuting cell ij* .

Additionally, taking the commuting cells as nodes, one can define the *recurrent mobility network* in which each node has an attribute N_{ij} and is connected to all others commuting cells sharing either the first or the second index with them. That means that all commuting cells $i \cdot$ or $\cdot j$ are connected to the commuting cell ij . The recurrent mobility network can be partitioned into regions by grouping the commuting cells according to their first index.

From the N_{ij} quantities, one can construct the number N_i of people living in region i by summing over the second index and the number N'_j of people working in region j by summing over the first index. Hence, we define

$$N_i := \sum_j N_{ij} \quad \text{and} \quad N'_j := \sum_i N_{ij}. \quad (1)$$

The total population $N = \sum_i N_i = \sum_j N'_j$. The primed quantities refer to the working population.

Opinion Structure and Vote Shares. Let us assume that individuals can choose between K different opinions. We denote by V_{ij}^k the number of people in commuting cell ij that have opinion k . For consistency, we must have

$$\sum_{k=1}^K V_{ij}^k = N_{ij}. \quad (2)$$

The state of the model is fully defined by the quantities V_{ij}^k . For convenience, we also introduce the vote shares

$$v_{ij}^k := \frac{V_{ij}^k}{N_{ij}}. \quad (3)$$

With this definition one can easily show that

$$v_i^k = \sum_j \frac{N_{ij}}{N_i} v_{ij}^k \quad \text{and} \quad v'^k_j = \sum_l \frac{N_{lj}}{N'_j} v_{lj}^k \quad (4)$$

are the vote shares of opinion k for the population living in region l and for the population working in region j .

Dynamics. Let us now define the transition operators $R_{ij}^{kk'}$

$$R_{ij}^{kk'} := P[(V_{ij}^k, V_{ij}^{k'}) \rightarrow (V_{ij}^k - 1, V_{ij}^{k'} + 1)] = \frac{N_{ij}}{N} v_{ij}^k p_{ij}^{k \rightarrow k'}, \quad (5)$$

that defines the probability that an individual changes from opinion k to opinion k' in the commuting cell ij . For the left most term, the first factor is the probability to choose an individual in the commuting cell ij ; the second factor is the probability that this individual holds opinion k and the third factor is the probability that she changes opinion to k' .

A procedure to get a stochastic version of a probability distribution using a Dirichlet distribution has been developed in [10]. This procedure can be utilized to add stochasticity to the rates of the SIRM model. In the rest of this paper, we consider the following model:

$$p_{ij}^{k \rightarrow k'} = \lambda \left(\alpha v_i^{k'} + (1 - \alpha) v_j^{k'} \right) + \beta \tilde{v}_{ij,D}^{k'} + \gamma / K, \quad (6)$$

where $\lambda, \beta, \gamma \geq 0$ and satisfy $\lambda + \beta + \gamma = 1$ and where

$$\tilde{\mathbf{v}}_{ij,D} = \text{Dir}(\mathbf{v}_{ij}/D), \quad (7)$$

where $\text{Dir}(\mathbf{v}_{ij}/D)$ denotes a sample from the Dirichlet distribution of parameter \mathbf{v}_{ij}/D and D controls the level of noise associated with the intra-commuting-cell influence. The parameter β represents the strength of the intra-commuting-cell influence and mainly act as a source of stochasticity, the parameter γ controls the intensity of free will encoding unilateral change of opinion, where all opinions are equally likely to be chosen. One parameter can be eliminated by rescaling the time. The state of the model is updated by resampling the population of each commuting cells according to the probabilities (6) as described in [10].

Mesoscale Approximation. A continuous time limit can be derived for the evolution of the commuting cell vote shares v_{ij} . As obtained in [10], it takes the form

$$dv_{ij} = \hat{d}(v_{ij})dt + \sqrt{\hat{D}(v_{ij})}dW_{ij}^*(t), \quad (8)$$

where $dW_{ij}(t)$ is a white noise, $\hat{d}(v_{ij})$ is a drift function and $\hat{D}(v_{ij})$ are diffusion coefficient.

In the case of two variants with the transition probability defined in (6), the drift and diffusion coefficients, $\hat{d}(v_{ij})$ and $\hat{D}(v_{ij})$, take the form

$$\begin{aligned} \hat{d}(v_{ij}) &= \lambda[\alpha v_i + (1 - \alpha)v'_j - v_{ij}] + \beta(\tilde{v}_{ij,D} - v_{ij}) + \gamma \left(\frac{1}{2} - v_{ij} \right) \\ \hat{D}(v_{ij}) &= \frac{1}{N_{ij}} \left[(1 - 2v_{ij}) [\lambda[\alpha v_i + (1 - \alpha)v'_j] + \beta \tilde{v}_{ij,D}] + (1 - \gamma)v_{ij} + \frac{\gamma}{2} \right]. \end{aligned} \quad (9)$$

These two functions are stochastic functions because of the presence of $\tilde{v}_{ij,D}$ in the formulation, which is a stochastic variable coming from the Dirichlet sampling process (7). As in previous analysis of the SIRM model, we neglect the diffusion function $\hat{D}(v_{ij})$ and use the normal approximation of Dirichlet distributed variable to obtained the SDE for the evolution of v_{ij}

$$\begin{aligned} dv_{ij} &= \left\{ \lambda [\alpha v_i + (1 - \alpha)v'_j - v_{ij}] + \gamma \left(\frac{1}{2} - v_{ij} \right) \right\} dt \\ &\quad + \beta \sqrt{\frac{\tilde{D}}{\tilde{D} + 1}} \sqrt{v_{ij}(1 - v_{ij})} dW_{ij}. \end{aligned} \quad (10)$$

The mesoscopic approximation of the dynamics is then given by the evolution of v_i , which can be obtained by combining (4) with (10). Reorganizing the terms, we obtain the following SDE

$$\begin{aligned} dv_i = & \left[C_1^{\text{SIRM}} \left(\frac{1}{2} - v_i \right) + C_2^{\text{SIRM}} \left(\sum_j \frac{N_{ij}}{N_i} v'_j - v_i \right) \right] dt \\ & + C_3^{\text{SIRM}} \sum_j \frac{N_{ij}}{N_i} \sqrt{v_{ij}(1-v_{ij})} dW_{ij}, \end{aligned} \quad (11)$$

where the coefficients are given by

$$C_1^{\text{SIRM}} = \gamma, \quad (12)$$

$$C_2^{\text{SIRM}} = \lambda(1-\alpha), \quad (13)$$

$$C_3^{\text{SIRM}} = \beta \sqrt{\frac{D}{D+1}}. \quad (14)$$

Equation (11) will be used as a basis for comparing the SIRM model with the USM.

2.2 The Utterance Selection Model (USM) for Language Change

The USM for language change [1] is an evolutionary model of language change inspired by population genetics. It has been shown in [4] that it is formally equivalent to the Wright-Fischer model of population genetics, where the population is divided into islands and where migration is possible between the islands. In this analogy, every speaker is an island and variants compete for being used. Variants are transmitted from one speaker to another through conversation and the production of utterances, which corresponds to the migration process between islands. This model and its extensions have been used to test hypothesis on new language formation [2], to explain sociolinguistic patterns [3] or to provide a theory for the self-actuation of changes in the language dynamics [8,11].

Network Structure. The USM describes the evolution of N speakers represented as a node of a network. The network of speakers is static and a parameter h_{ij} is associated with the edge connecting speaker i to speaker j . This parameter controls the attention that speaker i pays to speaker j , which is not necessarily symmetrical. For simplicity, this parameter will be assumed not to depend on the identity of the speakers $h_{ij} = h$ for all i, j . This parameter plays a similar role than the α parameter of the SIRM model as it controls the importance of neighbors on the network.

We will further assume that speakers are partitioned into M communities. These communities can be defined as any group of speakers, but the heterogeneous mean-field approximation developed in [9] and used in this paper works better if the nodes associated with the agents are well-connected. For each community c , the number of speakers in that community is denoted by N_c and the averaged node degree of that community is denoted by k_c .

State Vector. Every agent i is characterized by a state vector \mathbf{x}_i of length V representing the discrete probability distribution of using one of the available V variants, i.e., x_i^v is the probability that agent i uses variant v .

For each community c , the average state vector \mathbf{x}_c is defined by

$$\mathbf{x}_c := \frac{1}{N_c} \sum_{j \in c} \mathbf{x}_j, \quad (15)$$

where N_c is the number of agents in community c and the sum runs over the member of the community.

Dynamics. The dynamics of the USM is driven by pairwise interactions along the edges of the network. At each step a speaker i and a speaker j exchange utterances \mathbf{u} defined as a biased sample of their state vector \mathbf{x} , i.e.,

$$\mathbf{u} := \frac{1}{L} M \text{Multi}(L, \mathbf{x}), \quad (16)$$

where L is the length of the utterance, M is an innovation matrix (row stochastic) and $\text{Multi}(L, \mathbf{x})$ denotes a multinomial sample of length L and parameter \mathbf{x} , and update their state \mathbf{x} .

The change in state vectors of the USM [1, 9] is given by

$$\delta \mathbf{x}_i = \lambda [(1 - h)(\mathbf{u}_i - \mathbf{x}_i) + h(\mathbf{u}_j - \mathbf{x}_i)], \quad (17)$$

where λ is a learning parameter and h is the attention parameter controlling the weight of the incoming utterance from speaker j with respect to speaker i own utterance. Equation (17) fully determines the dynamics of the USM.

Mesoscopic Approximation. A *stochastic heterogeneous mean-field* (SHMF) approximation of the USM has been developed and analysed in [9]. In order to compare the dynamics of the SIRM model with that of the USM, we provide the SHMF of the USM for two variants. In this case, state vectors take the form

$$\mathbf{x} = \begin{pmatrix} x \\ 1 - x \end{pmatrix}, \quad (18)$$

and the equation for the first component is sufficient to characterize the dynamics.

In the utterance production rule (16), an innovation matrix M should be specified. For this analysis, we use the symmetrical case and set

$$M := \begin{bmatrix} 1 - q & q \\ q & 1 - q \end{bmatrix}, \quad (19)$$

where q is an innovation parameter.

With these specifications, the SHMF approximation for the quantities x_c describing the communities dynamics is given by

$$\begin{aligned} dx_c = & \left[C_1^{\text{USM}} \left(\frac{1}{2} - x_c \right) + C_2^{\text{USM}} \left(\sum_{c'} p(c'|c) x'_{c'} - x_c \right) \right] dt \\ & + C_3^{\text{USM}} \left[(1-h) \sqrt{\frac{x_c(1-x_c)}{k_c N_c}} dW_c + h \sum_{c'} p(c'|c) \sqrt{\frac{x_{c'}(1-x_{c'})}{k_{c'} N_{c'}}} dW_{c'} \right], \end{aligned} \quad (20)$$

where dW_c and $dW_{c'}$ are white noises, and $p(c'|c)$ is the probability that a speaker in community c interacts with a speaker in community c' , which can be computed by computing the fraction of edges originating in community c that terminate in community c' . For consistency, we have

$$\sum_{c'} p(c'|c) = 1. \quad (21)$$

The variable $x'_{c'} := (M\mathbf{x}_{c'})_1$ is the first component of the biased state vector $M\mathbf{x}_{c'}$ and is given by

$$x'_{c'} = x_{c'} - 2q \left(\frac{1}{2} - x_{c'} \right). \quad (22)$$

The coefficients of (20) are given by

$$C_1^{\text{USM}} = \lambda k_c 2q(1-h), \quad (23)$$

$$C_2^{\text{USM}} = \lambda k_c h, \quad (24)$$

$$C_3^{\text{USM}} = \lambda k_c (1-2q) \frac{1}{\sqrt{L}}, \quad (25)$$

where λ is a learning parameter, k_c is the averaged node degree of community c , h the attention parameter, q the innovation parameter and L is the length of utterances. Equations (11) and (20) will be compared and discussed in the next section.

3 Analogy Between the SIRM and the USM

In the previous section, we have derived mesoscopic approximations of both the SIRM model and the USM. These two Eqs. (11) and (20) are strikingly similar, but also have some significant differences that we will now explore. We start by comparing the deterministic parts (first line of (11) and (20)) and then discuss the stochastic parts.

3.1 Deterministic Terms

The deterministic parts of (11) and (20) contain two terms: a term due to free will/innovation and a term encoding the effect of interactions. We now discuss these two components.

The free will/innovation term takes the form $C_1(\frac{1}{2} - x)$ in both models and only differs by the form of the C_1 constant. In the SIRM model, this constant is a model parameter, whereas in the USM it takes a complex form given by (23). Assuming that the γ parameter of the SIRM model is usually small, and that the innovation parameter q of the USM is also a small quantity, we can argue that these two parameters play the same role; that of controlling the strength of unilateral change of opinion or variant either by free will or by innovation. As a result, we say that γ corresponds to q and encodes *unilateral change*.

The interaction term of (11) and (20) encodes the average influence of the other groups. Both terms are of the form $C_2 \left(\sum_j w_j x'_j - x_i \right)$, where the sum is the weighted average of the importance of neighboring groups.

The weights w_j encodes in both case topological features of the underlying networks. In the SIRM, $w_j := \frac{N_{ij}}{N_i}$ weight the importance of a neighboring region j by the fraction of people living in i commuting there. In the USM, $w_{c'} := p(c'|c)$ computes the fraction of people in community c talking to people in community c' . Therefore, we say that $\frac{N_{ij}}{N_i}$ corresponds to $p(c'|c)$ and encodes the *relative weight of neighbors*.

In the SIRM, the incoming signal from neighbors is encoded by $x'_j := v'_j$ the vote share at work in region j , whereas in the USM it is encoded by $x'_{c'}$, which represents the perceived utterance from community c' . The forms of these two terms are different and depend on the definition of the models. In the SIRM, the vote share at work in region j is given by (4) and introduce a coupling with all regions l for which $N_{lj} \neq 0$. Such a coupling between multiple regions is absent from the USM in which the incoming signal is simply a biased version of the state vector of the community and is given by (22). Such a difference can be interpreted by the fact that commuting cells encode specific behavior between regions i and j . In the language change context of the USM, this would correspond to adapting the utterances to the interlocutor. This feature is absent from the USM and could be added to it using “commuting cell” corresponding to state vectors tailored to the interlocutor. In addition, the biased aspect of the incoming signal of the USM encoded in the matrix M is not encoded in the incoming signal, but as an external noise. This difference explains the absence of the γ parameter in the incoming signal of the SIRM model. For the analogy, we say that v'_j corresponds to $x'_{c'}$ and encodes the *neighbor signal*.

Looking at the C_2 constant (13) and (24), we see that the parameter $1 - \alpha$ in the SIRM plays a similar role to the parameter h in the USM. Both weight the importance of interactions with respect to other processes. Therefore, we say that $1 - \alpha$ in the SIRM corresponds to the attention parameter h in the USM representing the *importance of interactions*.

3.2 Stochastic Terms

The stochastic terms in (11) and (20) take a slightly different form, that originates in where incertainty in interactions is introduced. In the SIRM model,

uncertainty is added into the intra-commuting cell term $\beta\tilde{v}_{ij,D}^{k'}$ in (6). This formulation enables to recover the original formulation of the SIRM as a limit [5]. In the USM, however, the uncertainty in interactions comes from the definition of utterances (16) and affect both utterances in the update rule (17). This difference is at the origin of the different form of the stochastic terms. However, the two formulations share the property to have multiplicative noise terms of the form $\sqrt{x(1-x)}dW$ originating either from a multinomial or a Dirichlet (Beta for two opinions) sampling process. In the SIRM, these multiplicative noises are indexed to the commuting cell ij , whereas in the USM they are indexed to the communities and not to subpart of them. It is possible to modify the dynamics of the SIRM to ressemble even more that of the USM by changing (6) to

$$p_{ij}^{k \rightarrow k'} = \lambda \left(\alpha \tilde{v}_{i,D}^{k'} + (1 - \alpha) \tilde{v}_{j,D}^{k'} \right) + \gamma/K, \quad (26)$$

where uncertainty is added to the vote shares at home and at work. Furthermore, since the uncertainty is included at the group level in the USM, the prefactors of noise are also dependent on the connectivity and size of the groups, whereas in the SIRM, the noise is added at the level of commuting cells, which are taken as nodes of the recurrent mobility network. This explains why in (11) there is no dependence of topological quantities in the noise term.

Despite these differences in structure, there is a similarity in the constant C_3 that enables to further extend the analogy between the two models. For instance, the parameter D in the SIRM model and the parameter L in the USM both controls the strength of the uncertainty in the incoming signal through a sampling process. It turns out that the covariance matrix of multinomial and Dirichlet processes only differ by a constant, since the Dirichlet distribution is the continuous analog to the multinomial distribution. Therefore, there is a correspondance between $\frac{D}{D+1}$ in the SIRM model and $\frac{1}{L}$ in the USM and encodes a *noise parameter*.

4 Discussion and Outlook

As we have shown, the SIRM model and the USM for language change are deeply related. The correspondance between the SIRM model and USM is summarized in Table 1.

This correspondance stresses the similarities between the two models and illustrates the fact that these two approaches are two faces of the same modelling coin, that of modelling stochastic dynamics on complex networks. Since their formulations comes from different lines of research they also have some significant differences. For example, the working population of a region i is made of contributions from many different regions whose vote shares feed back to region i . This long range coupling is absent from the USM, in which every speaker contributes either to the region where they live or where they work.

The differences between the two models suggest ways to develop and improve both models. For example, the USM for language change has the property that

Table 1. Correspondance between the parameters of the SIRM model and of the USM.

	SIRM	USM
Unilateral change	γ	q
Neighbor signal	v'_j	$x'_{c'}$
Relative weight of neighbors	$\frac{N_{ij}}{N_i}$	$p(c' c)$
Importance of interactions	$1 - \alpha$	h
Noise parameter	$\frac{\bar{D}}{1+\bar{D}}$	$\frac{1}{L}$

speakers use the same state vector \mathbf{x} in all their interactions, whereas the coupling between region i and region j is specific to this particular pair of regions and encoded through \mathbf{v}_{ij} . Such a substructure could be added to the USM to encode adaptation of speech to the identity of the neighbors and make this model more realistic. Furthermore, the SIRM model could be modified to incorporate more naturally the different sources of noise and, therefore, have a more elegant form. This could be done by adding noise on the vote shares at home and at work as in (26) and by assuming that innovation or errors can occur at this level by, for example, redefining the quantity $\tilde{\mathbf{v}}_{ij,D}$ to be

$$\tilde{\mathbf{v}}_{ij,D} = M\text{Dir}(\mathbf{v}_{ij}/D), \quad (27)$$

where M is a mutation matrix similar to that of the USM.

In conclusion, we have shown in this paper how the SIRM model and the USM are similar in many ways and how their differences can lead to improvements of both models. Since the USM has connections to evolutionary models, one can further the analogy presented here to interpret opinion dynamics models such as the Voter Model or the SIRM model in evolutionary terms. Furthermore, extensions of the USM have been developed [3, 8, 11] and the analogy developed here can be extended as well, opening new avenues of research.

References

1. Baxter, G.J., Blythe, R.A., Croft, W., McKane, A.J.: Utterance selection model of language change. *Phys. Rev. E* **73**(4), 046118 (2006). <https://doi.org/10.1103/PhysRevE.73.046118>
2. Baxter, G.J., Blythe, R.A., Croft, W., McKane, A.J.: Modeling language change: an evaluation of Trudgill's theory of the emergence of New Zealand English. *Lang. Var. Change* **21**(02), 257–296 (2009). <https://doi.org/10.1017/S095439450999010X>
3. Baxter, G.J., Croft, W.: Modeling language change across the lifespan: individual trajectories in community change. *Lang. Var. Change* **28**(02), 129–173 (2016). <https://doi.org/10.1017/S0954394516000077>
4. Blythe, R.A., McKane, A.J.: Stochastic models of evolution in genetics, ecology and linguistics. *J. Stat. Mech: Theory Exp.* **2007**(07), P07018 (2007). <https://doi.org/10.1088/1742-5468/2007/07/P07018>

5. Fernández-Gracia, J., Suseck, K., Ramasco, J.J., San Miguel, M., Eguíluz, V.M.: Is the voter model a model for voters? *Phys. Rev. Lett.* **112**, 158701 (2014). <https://doi.org/10.1103/PhysRevLett.112.158701>
6. Fisher, R.A.: *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford (1930)
7. Holley, R.A., Liggett, T.M.: Ergodic theorems for weakly interacting infinite systems and the voter model. *Ann. Probab.* **3**(4), 643–663 (1975). <https://doi.org/10.1214/aop/1176996306>
8. Michaud, J.: Dynamic preferences and self-actuation of changes in language dynamics. *Lang. Dyn. Change* **1**, 61–103 (2019). <https://doi.org/10.1163/22105832-00901003>
9. Michaud, J.: Continuous time limits of the utterance selection model. *Phys. Rev. E* **95**, 022308 (2017). <https://doi.org/10.1103/PhysRevE.95.022308>
10. Michaud, J., Szilva, A.: Social influence with recurrent mobility and multiple options. *Phys. Rev. E* **97**, 062313 (2018). <https://doi.org/10.1103/PhysRevE.97.062313>
11. Stadler, K., Blythe, R.A., Smith, K., Kirby, S.: Momentum in language change. *Lang. Dyn. Change* **6**(2), 171–198 (2016). <https://doi.org/10.1163/22105832-00602005>
12. Wright, S.: Evolution in mendelian populations. *Genetics* **16**(2), 97 (1931)



Networks of Intergenerational Mobility

Tanya Araújo^{1,2(✉)}, David Neves^{2,3}, and Francisco Louçã^{1,2}

¹ ISEG, University of Lisbon, Lisbon, Portugal
tanya@iseg.ulisboa.pt

² UECE - Research Unit on Complexity in Economics, Lisbon, Portugal

³ PSE - Paris School of Economics, Paris, France

Abstract. The work of Piketty has documented the long-run evolution of capital accumulation and the increasing wealth inequalities in western countries. This work re-opened the policy debate around the taxation of capital and inheritance. Standard models of capital taxation prescribe a null optimal tax rate for capital holdings on the basis that the supply of capital is infinitely elastic and that economic agents fully optimize over their life-cycle. These assumptions rarely hold. In turn, meritocracy and equality of opportunity are also presented as rationales for the capital taxation of inheritance. To which extent does inequality in inherited wealth magnify wealth inequality and by how much? What is the tax rate on capital flows and inheritance stocks that stabilizes wealth inequality, for a given growth rate of population and growth rate of income? To investigate these questions, we construct an agent-based Sugarscape model of wealth accumulation with overlapping generations, heterogeneous skill endowments, skill inheritability, assortative mating based on skills and wealth position. Simulations end up with networks of relatives, usually comprising four overlapping generations. Network structures show that assortative mating based on skills tend to have a more pronounced role in promoting equality in inherited wealth. Intergenerational mobility is strongly related with inherited skills being reinforced by marital sorting based either on skills or wealth levels.

Keywords: Skill inheritability · Assortative mating · Sugarscape model · Wealth distribution · Intergenerational mobility

1 Introduction

A curious historical investigation on search costs, market segmentation and the play of other social and cultural preferences in marriage was offered by Goñi [1]. Discussing evidence from 10364 marriages between 1531 and 1880 in Britain, he found that marriage between sons and daughters of aristocrats and commoners was negatively related to inequality of the distribution of landed wealth, as expected. But then he studied a curious social pattern defined between 1851 and 1875, when the landlords and their families met in London from Easter to August to perform a large series of social events, organized in order to define

marriage, what was called the Season. This established a marriage market and the norm of social reproduction through alliances of landowners. But, as the mother of Queen Victoria died in 1861 and then Prince Albert passed away the next year, these social events were interrupted until 1863. The interruption of the Season provoked an increase of marriage of peers with commoners by 80%, since age reduced the possibility of marriage, with the consequent loss of capital accumulation, political power and wealth concentration. Although our model does not capture the intricacies of this market, we simulate combinations of families and wealth through marriage as a technique of social reproduction.

The factors governing marriages deviate from randomness in all possible ways, yet the single deviation from randomness that does seem to hold is the popular saying that “opposites attract”. Opposites frequently do not attract and the most common deviation from randomness in marriage patterns is precisely matching based on similarity of either biological or cultural traits, embodied in the concept of assortative matching. Indeed, multi-dimensional similarity has been pointed out as the main driver of human mating, as cultural, biological and even genetic traits have been found to be highly correlated between spouses [2]. But if multi-dimensional similarity drives human mating, what are then the evolutionary, economic or social consequences of this process? In this study we investigate the consequences of assortative mating on wealth inequality and its interplay with wealth taxation. Today’s marriage markets are mostly informal, and it is virtually impossible to isolate a particular institution governing the matching process. That was not the case in the past, where there were explicit institutions designed to match individuals along specific traits. A prominent example is the case of the “Bottin Mondain”, an institution designed to promote homogamous marriages among French aristocrats Bisin and Verdier [3]. In a similar fashion, during the Victorian England, a series of royal parties were held in London from Easter to August to promote homogamous mating among the English aristocracy, as presented in the above mentioned work of Goñi [1]. On the other hand, mating along the wealth dimension might also be viewed as part of a more broad process of assortative matching driven by cultural traits and the desire of “vertical socialization”. Research by Bisin and Verdier [3] models human matching along cultural traits as an explicit mechanism to enforce cultural transmission across generations. Parents are assumed to be driven by “imperfect empathy”, seeing their children well-being through the lens of their own preferences, so that matching along cultural traits motivated by the desire to enforce vertical socialization of children. In turn, homogamous marriages are faced as the most efficient technology to enforce vertical socialization. In the absence of frictions in the marriage market (searching costs), this model boils down into perfect homogamy and increases the frequency of extreme cultural traits, decreasing the average ones.

More recently, Ghidi [4] linked the process of “vertical socialization” on the basis of cultural traits, with market outcomes, arguing that socialization efforts of offspring take into account whether cultural traits are associated with poor economic outcomes, thereby concentrating on traits that have market value.

Overall, the author links matching along cultural traits with transmission of skills, by establishing that parents do not jeopardize their offspring' economic opportunities for the sake of trait transmission alone, as poorly valued market traits are not transmitted.

Our work also contributes to the literature on policy implications of inter-generational transmission of wealth. Piketty [5] has pointed out the fact that attitudes towards taxation and redistribution exhibit a very high rate of dynastic reproduction. Parental income or dynastic experiences of income mobility seem to strongly influence current visions towards redistribution (which is at odds with the implications of standard choice models of redistributive politics, which only associate current income to agent' preferences towards redistribution). If past mobility experiences across generations are attributed to the role of effort, then individuals are socialized towards meritocratic beliefs and disregard the role of predetermined factors. Hence, in the long-run, different income trajectories generate dynasties that converge towards different beliefs regarding the role of effort and the role of predetermined factors, and therefore to different beliefs about the optimal tax rate and the corresponding level of redistribution. This model contradicts the standard intuitions from the models of redistribution politics, as it places the role of the vertical socialization towards meritocratic beliefs at the heart of political dispute on the levels of income taxation, rather than just the levels of current income that an individual attains.

Our study of the effect of assortative mating based on wealth and skills relies on the characterization of the resulting networks of agents. To this end, after each simulation, the surviving agents define the network nodes where the links are defined exclusively by the parental bonds between each surviving agent and its parents, provided that the parents are alive (at least one of them). Two nodes having a common offspring are also linked. Therefore, the resulting network has some clustering as besides the horizontal links child-mother and child-father there are also mother-father links provided they both survive.

2 The Sugarscape Model

The well-known book “Growing Artificial Societies” by Epstein and Axtell [6] introduced the original Sugarscape agent-based model, targeted at investigating a variety of social phenomena, among which the economics of wealth distribution. The Sugarscape model has two main elements, a landscape and a set of agents. The landscape is the place where events unfold, a 2D grid where each cell contains a certain amount of abstract wealth, called sugar in the original model. The landscape sets the spatial distribution of this generalized resource, which is the agent revenue level at that location. Each agent is characterized by fixed states (genetic characteristics): its metabolism (rate of which it consumes wealth) and length of vision (range of neighboring cells it can observe) which here is represented by skill, and variable states: its present location and amount of wealth, which is initially set with the amount of that resource in the location of the Sugarscape in which the agent was born. The agents move around the

landscape following simple rules of behavior. At each time step, the agent looks at its neighboring cells and moves to the cell with the most wealth. These rules can be extended to include features as reproduction, death and other interaction phenomena. The Sugarscape, even in its simplest version, is able to generate some real world characteristic outcomes. As observed in most real economies, in this case modeling the evolution of wealth generates a long-tailed distribution, showing that some agents are vastly richer than others. Sugarscape models also contain a growback rule specifying how wealth grows back at each time step. It may grow back to full capacity immediately or grow back at a rate of units per time interval up to the capacity at that position. Other rules concern wealth taxation, death, reproduction and inheritance. When taxation is present, it occurs at a rate r and at every t time steps, so that a fraction f of the wealth of each agent is collected and, for simplicity, equally redistributed. Reproduction/marriage is ruled under assortative matching, which can be driven by similarity in either skills or wealth.

At each time step, every pair of surviving agents is visited provided that their age is between 20 and 45 and, depending on the choice of the scenario, which may be skill-driven or wealth-driven, the similarity among their skills or wealth positions is computed. If they coincide in their skill levels or if the difference between their accumulated wealth is below a given threshold, they produce an offspring which is endowed with their skill level and the average value of the parents' wealth plus a small random value. The inherited values of both skill and wealth include a small random component in their computation, for the sake of representation of omitted variables. Agents die at the age of 80 or when their wealth level goes to zero. Each simulation run (time step) corresponds to one year.

2.1 Simulations

All simulations end after 150 runs ($runs = 150$), start with around 400 agents placed on a two-peak 50×50 landscape (the sugarscape) as Fig. 1 shows. Initial conditions are set such as each agent skill is an integer value between 0 and 10 ($maxSkill = 10$), randomly generated. The amount of wealth is also randomly defined over the interval $[0, 5]$ ($maxWealth = 5$).

The landscape has two symmetric concentrations of wealth, one in the southeast side of the grid, and the other in the northwest. In the 50×50 grid, the southeast peak is approximately on the $(0.75 * 50, 0.25 * 50)$ coordinate, while the other is on the $(0.25 * 50, 0.75 * 50)$ coordinate. From the peaks down, the level of wealth at each location will follow a decreasing path.

Two scenarios are investigated, each of which, with and without inheritance taxation: one is the case in which assortative mating is skill-driven, and the other in which it is wealth-driven. In each case, similarity between agent' skill or wealth positions is computed, driving (constraining) reproduction. In the former, a pair of agents produces an offspring if they coincide in their skill levels; in the latter, reproduction occurs if and only if the difference between the parents' accumulated wealth is below a given threshold (tw). In the two scenarios,

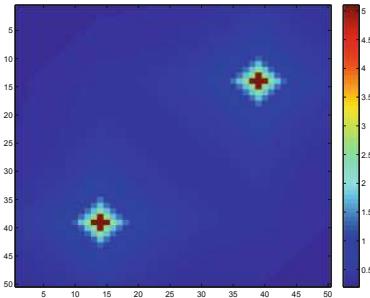


Fig. 1. A two-peak landscape. The color bar indicates the initial amount of wealth at each location.

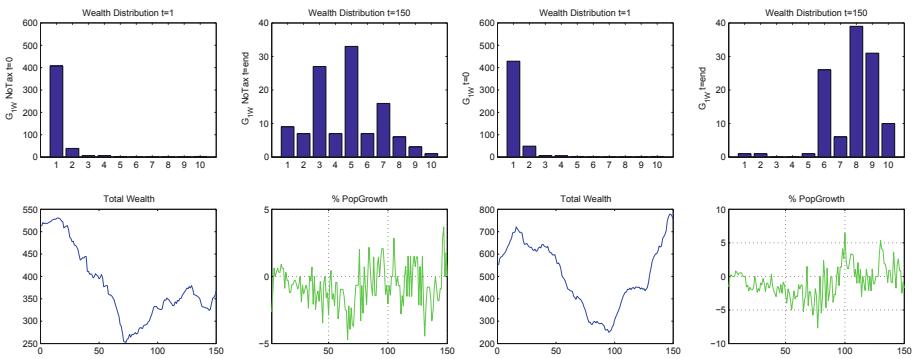


Fig. 2. Results from the wealth-driven scenario without (left) and with taxation (right).

initial conditions are as presented above ($runs = 150$; landscape $size = 50 \times 50$; $maxMetabolism = 1.3$; $maxSkill = 10$; $maxWealth = 5$; Growback at 10% per time step).

3 Results

Figure 2 shows the results from the wealth-driven scenario. The four subplots on the left side show the outcomes obtained without taxation of inheritance, while the four subplots on the right side show results coming from the introduction of a tax rate of 0.5% on the amount of inherited wealth by each newborn agent. In each set of four plots, the first and second subplots show the initial and final distribution of wealth ($t = 1$ and $t = 150$, respectively). The third subplots show the evolution of the total amount of wealth, while the fourth ones present the evolution of the rate of population growth. In the wealth-driven scenario, these results show that the introduction of inheritance taxation led to an increase in the amount of wealth and in the rate of population growth. Simultaneously, it also led to a slightly decrease of wealth inequality, as the histograms in the

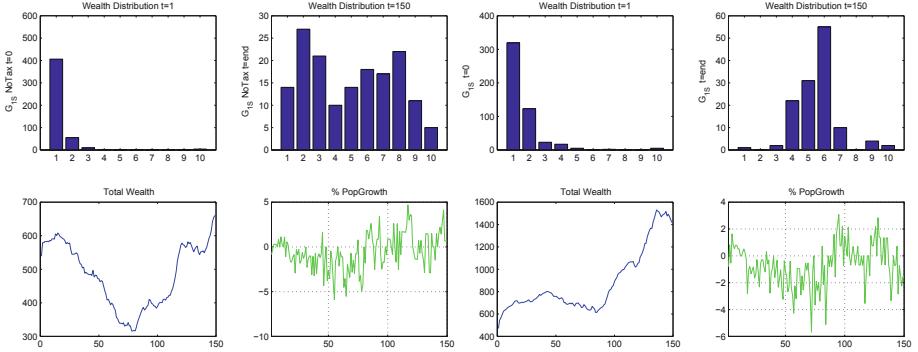


Fig. 3. Results from the skill-driven scenario without(left) and with taxation (right).

second subplots (final wealth distribution) show. Figure 3 shows the results from the skill-driven scenario. As in Fig. 2, The four subplots on the left side show the outcomes obtained without taxation of inheritance, while the four subplots on the right side show results coming from the introduction of a tax rate of 0.5% on the amount of inherited wealth by each newborn agent. Compared to the results obtained from the wealth-driven scenario, the most remarkable change concerns the increase in total wealth when taxation of inheritance is present and a slightly decrease in the rate of population growth. In both scenarios, with and without taxation, there is a change in the evolution of total wealth when $t = 80$, i.e., when the maximum age (80 years) is reached, meaning that from $t > 80$ to the end of the simulation, every living agent is a heir. We also measured the Pearson correlation coefficient between individual skills and wealth levels and results show that it is always weak (around 0.4) in both scenarios, with or without capital taxation.

As we aim to study the effect of assortative mating (either driven by wealth similarity or by identical skills) in the way agents organize themselves into families, in the following we analyse the structure of the resulting networks of related agents by relative bonds.

3.1 Networks

At each time step (t), every pair of surviving agents (s_i and s_j) is visited.

In the **wealth-driven** scenario, the agents produce an offspring if and only if the age of each of them is in between 20 and 45 and the difference between their accumulated wealth is below a threshold τ_w :

$$|wealth(s_i, t) - wealth(s_j, t)| < \tau_w \quad (1)$$

where $\tau_w = \frac{0.1}{N} \sum_{k=1}^N (wealth_k)$, i.e., τ_w is set with 10% of the average value of wealth per capita at time t , being N the number of surviving agents.

In the **skill-driven** scenario, the agents are required to have the same skill level ($skill_i = skill_j$) in order to produce an offspring.

When the above conditions are satisfied, the two agents produce an offspring which is endowed with the parents' skill level and the average value of the parents' wealth, as Eqs. 2 and 3 show.

After each simulation, the surviving agents define the network nodes s where the links are defined by the parental bonds between each surviving agent and its living parents and by the agents that share a common offspring. Formally, two nodes s_i and s_j in the network S are connected, if and only if:

- s_i is an offspring of s_j or *vice-versa*
- s_i and s_j share at least one living offspring

At time t , a newborn offspring s_o is endowed with the parents' s_i and s_j attributes, as follows:

$$\text{skill}(s_o, t) = \text{skill}(s_i, t) \quad (2)$$

and

$$\text{wealth}(s_o, t) = \frac{\text{wealth}(s_i, t) + \text{wealth}(s_j, t)}{2} + \delta \quad (3)$$

where δ is a random value $\delta \in [0, 1]$.

Typical results end up with two-thirds of surviving nodes and around 10 families comprising four overlapping generations, each of them with a variable number of relatives. Most families comprise four generations, resulting from simulations of 150 time steps (years). Figure 4 shows the network resulting from a typical run of the wealth-driven scenario with taxation. There are 11 families (colors correspond to families), the size of each node corresponds to its amount of wealth while its label shows its skill level.

The typical network emerging from the skill-driven scenario with wealth taxation is presented in Fig. 5. There, a smaller contribution to the preservation of wealth levels (node size) along with consecutive generations inside families (nodes with the same color) is observed. Therefore and in accordance with the results presented in the subplots of Fig. 3, the introduction of taxation of inheritance in the skill-driven scenario contributes to a larger extent to the decrease of wealth inequality. Preservation of skills (node label) is observed in both skill- and wealth-driven scenarios, with or without capital taxation.

Table 1. Topological coefficients of each resulting network: size (N), average degree (k), diameter (d) and characteristic path length (PL).

Scenario	N	k	d	PL
wealth-driven without taxation	186	1.8	5	1.83
wealth-driven with taxation	178	1.7	5	1.60
skill-driven without taxation	177	1.8	4	1.53
skill-driven with taxation	170	1.7	5	1.93

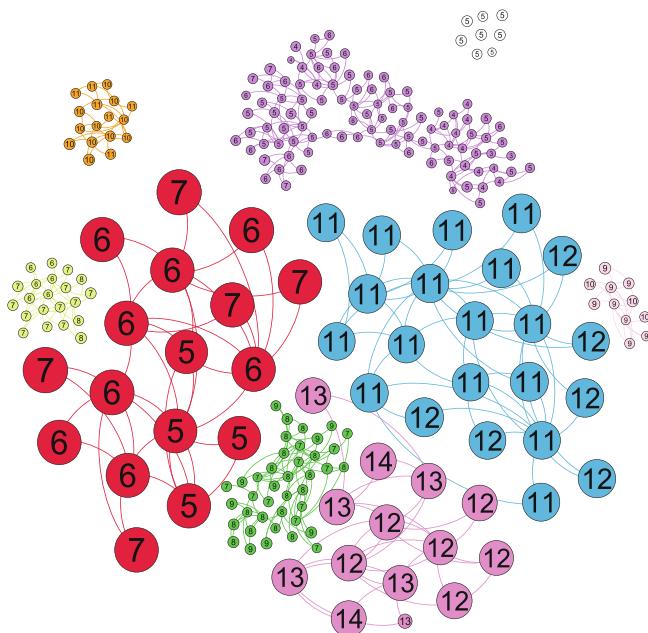


Fig. 4. Networks of relative bonds from the wealth-driven scenario with taxation.

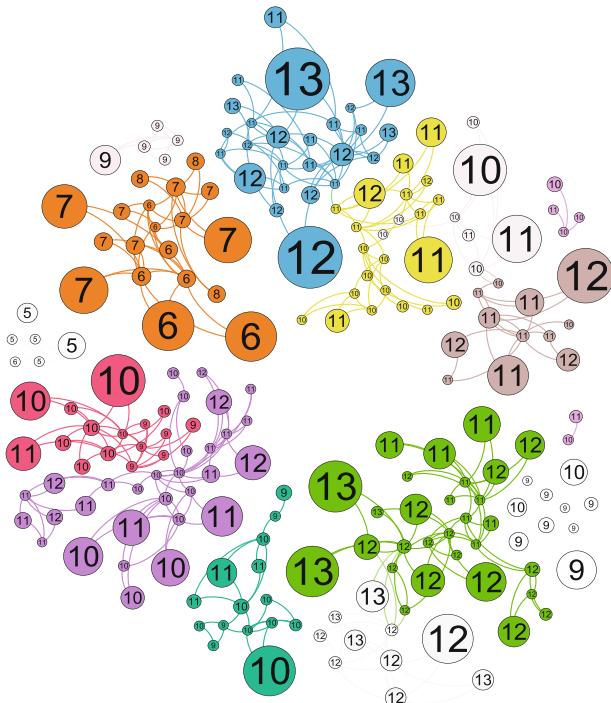


Fig. 5. Network of relative bonds from the skill-driven scenario with taxation.

Table 1 shows the values of some topological coefficients of each resulting network. Not surprisingly, the networks of the same size (N) show similar average degree (k), diameter (d) and characteristic path length (PL) values. Their main difference relies on the nodes' size inside each family. In the skill-driven scenario, mostly when taxation is present (Fig. 5), wealth mobility across generations is clearly stronger than in the wealth-driven one (Fig. 4).

4 Concluding Remarks

1. Initially, the emerging wealth distributions are highly skewed and mimic the one that it is usually found in empirical data.
2. The model also performs well in reproducing long-run stylized facts of capital accumulation.
3. Unexpectedly, the correlation coefficient between individual skills and wealth levels is always weak (around 0.4) in both scenarios, with or without capital taxation.
4. The network structures show that assortative mating based on skill tend to have a more pronounced role in promoting equality in inherited wealth.
5. We also find that intergenerational mobility is strongly related to inherited skills being reinforced by marital sorting based either on skill or wealth level.

The results confirm the usual policy rational for inheritance taxation: it promotes intergenerational mobility and reduces the weight of predetermined variables on wealth outcomes. Yet, our results innovate by showing that the classical rational for inheritance taxation critically hinges on the specific form of assortative matching arising in marriage markets: taxation of inherited wealth becomes an effective tool to induce intergenerational mobility in cases where assortative matching occurs mostly along wealth traits.

Acknowledgement. UECE (Research Unit on Complexity and Economics) is financially supported by FCT (Fundação para a Ciência e a Tecnologia), Portugal. This article is part of the Strategic Project (UID/ECO/00436/2019).

References

1. Goñi, M.: Assortative matching and persistent inequality: evidence from the world's most exclusive marriage market. Economic History Association Annual Meeting, Nashville, Tennessee, US (2015)
2. Buss, D.M.: Human mate selection. *Am. Sci.* **73**(1), 47–51 (1985)
3. Bisin, A., Verdier, T.: Beyond the melting pot: cultural transmission, marriage, and the evolution of ethnic and religious traits. *Q. J. Econ.* **115**(3), 955–988 (2000)
4. Ghidi, P.M.: A model of ideological transmission with endogenous parental preferences. *Int. J. Econ. Theory* **8**(4), 381–403 (2012)
5. Piketty, T.: Social mobility and redistributive politics. *Q. J. Econ.* **110**(3), 551–584 (1995)
6. Epstein, J., Axtell, R.: Growing artificial societies. *J. Math. Phys.* **4**(5), 701–712 (1963)



Inequality in Learning Outcomes: Unveiling Educational Deprivation Through Complex Network Analysis

Harvey Sánchez-Restrepo¹ and Jorge Louçã²

¹ University of Lisbon, Cidade Universitária, 1649-004 Lisbon, Portugal
Harvey_Restrepo@iscte-iul.pt

² Information Sciences, Technologies and Architecture Research Center,
ISCTE-IUL, 1649-026 Lisbon, Portugal
jorge.l@iscte-iul.pt

Abstract. Understanding which factors are determinant to guarantee the human right to education entails the study of a large number of non-linear relationships among multiple agents and their impact on the properties of the entire system. Complex network analysis of large-scale assessment results provides a set of unique advantages over classical tools for facing the challenge of measuring inequality gaps in learning outcomes and recognizing those factors associated with educational deprivation, combining the richness of qualitative analysis with quantitative inferences.

This study establishes two milestones in educational research using a census high-quality data from a Latin American country. The first one is to provide a direct method to recognize the structure of inequality and the relationship between social determinants as ethnicity, socioeconomic status of students, rurality of the area and type of school funding and educational deprivation. The second one focus in unveil and hierarchize educational and non-educational factors associated with the conditional distribution of learning outcomes. This contribution provides new tools to current theoretical framework for discovering non-trivial relationships in educational phenomena, helping policymakers to address the challenge of ensuring inclusive and equitable education for those historically marginalized population groups.

Keywords: Educational network · Large-Scale Assessments · Policy informatics

1 Introduction

With the establishment of the Sustainable Development Goals (SDGs), the 193 countries attached to Unesco promulgated that ‘education is a human right’ [1] and that the two pillars of quality in education should be learning and equity [1, 2], recognizing that all human beings have the right to learn and that the State is obliged to guarantee to all citizens equally [3]. Education is also a source for social mobility: an additional year of quality education can increase a person’s income up to 10% [4]. The lack of quality in education worldwide is of such magnitude that Unicef estimates that 250 million

children do not have the minimum learning and that more than half of them have been fooled systemically: despite having managed to attend school, many of them fail in developing the minimum learning such as reading, writing or performing basic operations. The causes for this deprivation of learning are multiple but they affect mainly to those belonging to historically marginalized population groups, almost always, the most impoverished [5–7]. Further, the modest improvements in learning achievements are usually by the hand of enormous inequalities among students, which has cast doubt on the government actions for improving the quality of education [8].

As a strategy to strengthening public policies, many countries have implemented national Large-Scale Assessments (LSA) to collect valid and reliable data on educational outcomes and Factors (variables) associated with Learning (FAL). The objective of LSA is to have evidence-based information and use quantitative methods to estimate educational changes when varying each factor [9] in many scales of time and across the territory [10], as well to modelling dynamic behavior and asymmetries at school and student level [11]. However, dominant models to study educational phenomena are based on tools that postulate that educational gaps can be explained just through the covariation between learning outcomes and each factor [12]. In this kind of studies, the linear analysis provokes a fragmented view of the system and dismiss very often the interactions between agents, FAL and educational phenomena [13, 14], dismissing the structure of the inequality in learning outcomes and its relationship with educational deprivation, which partially explains the lack in explanatory power of those models [15].

To represent the multiple interactions in a system and the emergence of collective properties at different scales from their constituents, several researchers have proposed the use of network theory to model social systems as a result of self-organized processes [17, 18]. Given that ‘networks are at the heart of complex systems’ [16], analyzing statistical and topological properties of the education system through network theory means studying the complexity of the system through its interactions.

Therefore, this research addresses empirical data for estimating educational deprivation in a Latin American country, as well as its relationship with the most relevant social determinants such as Socioeconomic status (SES) of the student and their families, Rurality in the area where the school is located (RA), the Type of school (TS) and Ethnic self-identification of the student (ET), for estimating topological features and order parameters of the network related to out-of-equilibrium states, providing a new kind of information about global and local properties of the structure of educational deprivation and helping to find those key factors driving inequality gaps in learning outcomes.

1.1 Dataset

In this model, a multivariate dataset integrates learning outcomes of every student who has completed the k-12 education process, estimated by the ability’s parameter θ^j through a LSA carried out in Ecuador in 2017 using a standardized computer-based test¹ and integrated with a robust dataset with more than 140 variables coming from

¹ Full dataset is available in <http://www.evalucion.gob.ec/evaluaciones/descarga-de-datos/>.

surveys to student's families and teachers. For building the scores, psychometric parameters were estimated by Item Response Theory through a 2P-Logistic model [19] following Eq. 1:

$$P(\theta_j) = \frac{e^{[z_j(\theta_j - \beta_j)]}}{1 + e^{[z_j(\theta_j - \beta_j)]}} \quad (1)$$

$$\text{Con } \theta_j \in (-\infty, \infty), z_j \in (-\infty, \infty) \text{ y } \beta_j \in (-\infty, \infty).$$

Raw scores of θ^j were re-scaled to a Learning index ($LI_j \in [4.0, 10.0]$), a monotonous transformation, where higher levels of learning are more likely to have higher scores [19]. For measuring relative deprivation, this model uses the sociological proposal that 'needs, thresholds and satisfactions are determined by each society', while absolute deprivation proposes that 'there is an irreducible nucleus of needs that are common to every human being' [20]. All students are classified in levels of achievement (L_k) based on a standard Bookmark process for establishing psychometrical cut points s_i [19]. Students suffering learning deprivation are those that did not meet the minimum learning standards at the end of the compulsory education, denoted by L_0 .

1.2 Deprivation Learning Index

To estimate this index, we use the family of scores $\{LI_j\}_{j=1,\dots,N}$, of those students with low level of achievement L_0 -class, s_1 is the first cut point —the minimum score to be located at level L_1 —. For the L_0 -class, absolute deprivation is given by $H = (n(L_0)/\sum n(LI_j))$, where $n(LI_j)$ represents the number of students below the first level of achievement, the intensity $\lambda(LI_j)$ is given by the distance to reach the first level L_1 and the Deprivation learning index (DLI) is given by $\delta_j = H \cdot \lambda(LI_j)$, which represents a measure of the collective learning deficit, which considers the magnitude —the number of students with low performance— and intensity —how much below the minimum performance level are located [20].

1.3 Model Specification

For analyzing topological properties and pointing out those nodes holding the system out of statistical equilibrium, a three sequential steps model was developed. The first step is focused on disaggregating L_0 -class by SES, each student is represented by a node and edges, weighted by $\lambda(LI_j)$, are directed to one of the SES-decile nodes $\{LI_j(\theta^j \rightarrow L_k^j \rightarrow (SES_d^j))\} \forall j$, a process which allows to analyze aggregated inequality at school level, as well as In-degree distribution for SES nodes. The second one is an extension for including RA, TS and ET to analyze their effects through the sequence $\{\theta^j \rightarrow L_k^j \rightarrow (SES_d^j) \rightarrow (RA_{C1}^j, TS_{C2}^j, ET_{C3}^j)\} \forall j$, where C denotes an index for each subcategory of the factors RA, TS and ET. Finally, the third step amplifies and strengthens the analysis through more than one hundred educational and non-educational factors associated with learning achievements trough the sequence $\{\theta^j \rightarrow L_k^j \rightarrow (SES_d^j) \rightarrow (RA_{C1}^j, TS_{C2}^j, ET_{C3}^j) \rightarrow FAL_{Cm}^j\} \forall j$, for integrating m different

educational and non-educational factors. Network analysis was carried out by Gephi 0.9.2 and statistical estimations and plots with R 3.5.0 and Orange 3.3.8.

2 Socioeconomic Status and Learning Deprivation

The first specification estimates the Weighted In-degree distribution of directed edges from social determinants nodes to those representing socioeconomic deciles, given by $\{SES_d\}d \in \overline{1, 10}$, where each edge represents one student in L_0 -class. As inequality implies asymmetries, in conditions of total equity —where socioeconomic factors would not produce differences— we might expect equal distribution of L_0 -edges over the network. Therefore, the study of equity can be deepened by analyzing the levels of absolute and relative deprivation experienced by different population groups and their relationship with the SES of the students.

According with LSA estimates, 8 438 of 39 219 students are in L_0 -class, a prevalence rate of 0.215, a $LI = 6.32$ and intensity of deprivation $\lambda = 0.225$, i.e., in average, L_0 -student lacks 0.68 standard deviations (SD) of the minimum learning. As shown in Fig. 1, distribution of In-degree $P(SES_d)$ is non-uniform and it decreases monotonically as the SES of the group increases.

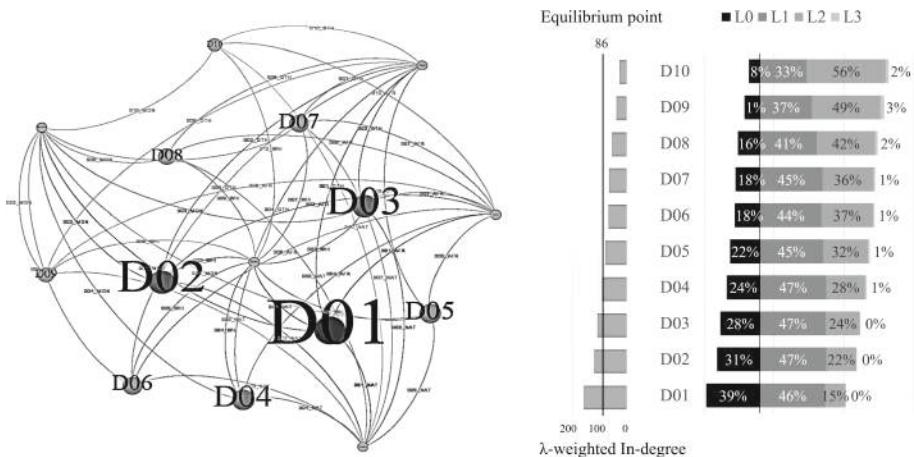


Fig. 1. Network of socioeconomic distribution with λ -weighted learning deprivation and distribution of students among levels of achievement.

For example, $P(SES_1) = 0.39$ and $P(SES_{10}) = 0.08$, this difference of 0.31 means that for each richest-family student who does not learn the minimum, there are 5 poorest-family students in the same situation. As we will see later, this unfortunate situation deepens in rural areas, where the ratio increases to one richest-student for every 7 poorest-students.

Differential centrality of deciles in Fig. 1 also shows a non-equilibrium system driven by SES where poorest students dominate the graph: nodes SES_1 , SES_2 and SES_3 have stronger connections and the highest prevalence rate ($DPR = 0.3860$), Hub parameter ($H = 0.4887$), Weighted In-degree ($WID = 158.9420$) and PageRank ($PR = 0.4289$). On the contrary, the richest students grouped by SES_{10} are relative irrelevant for the network with parameters $DPR = 0.0800$, $H = 0.1133$, $WID = 27.222$ and $PR = 0.0289$. As can be seen, in-degree parameter provides an estimation for deprivation rates of each SES_d , showing the size of the gaps among them through detecting SES effects in nodes grouping L_0 -students by deciles, pointed out by the negative correlation between LI and SES ($R = -0.58$, $p < 0.001$).

To estimate more accurately the cumulative effect of SES on learning outcomes, Fig. 2 shows LI (left plot) and intensity of deprivation (right plot), as functions of SES in two dynamical ways: (1) starting with the whole population of students and re-estimating LI and λ while excluding poor-students (black circles), and (2) starting with just poor-students and including richer students (white circles). As can be seen, the biggest circles indicate equal Global Average (GA) for LI (7.61) and λ (0.225), however, for case (1) the effect of removing poor students is that GA goes up immediately after removing SES_1 -students, reaching a $LI = 8.21$ (0.65 SD away from GA), as well as λ diminishes 14%. On the contrary, when the estimation process starts with just the poorest students, $LI = 6.52$ (-1.09 SD below from GA) and starts going down after including SES_3 students. In summary, these opposite behaviors show a learning gap of 1.82 SD, equivalent to almost two years of formal schooling between richest and poorest students, and pointing out that, even among those deprived students, the most impoverished get the worst part suffering 14% deeper intensity of deprivation.

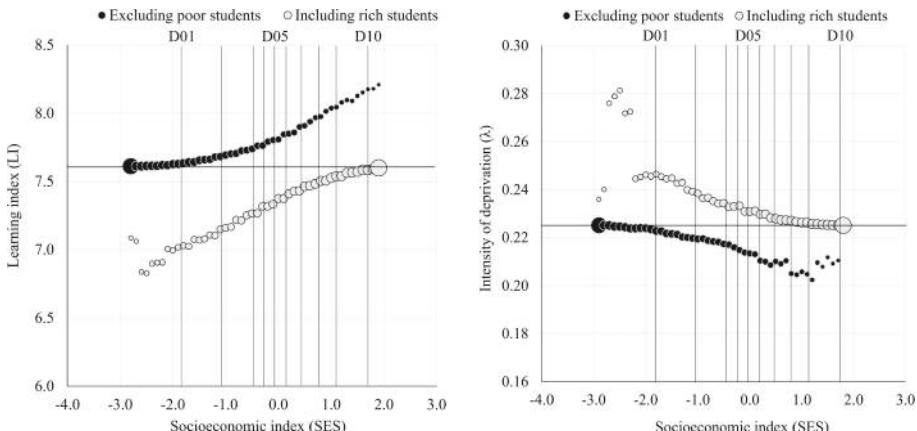


Fig. 2. Cumulative effect of SES in learning outcomes (*left side*) and intensity of deprivation (*right side*).

The previous results provide very clear information about learning inequalities at student level, however, to analyze an aggregated phenomenon, the school level shows local but systemic gaps to better understand the sources of structural deprivation. When studying schools as integrated units, the impact of *SES* becomes even more evident in learning outcomes, in left side of Fig. 3, schools are splatted between type of school —sources of funding—, here each school is represented by a circle whose size is proportional to its number of students. The average *SES* of students is located on the horizontal axis and *LI* is on the vertical one. As can be seen, private schools have higher *SES* (0.42) and *LI* (7.96) than the public ones (*SES* = -0.21) and (*LI* = 7.44), and there is also a strong positive relationship between *LI* and *SES*.

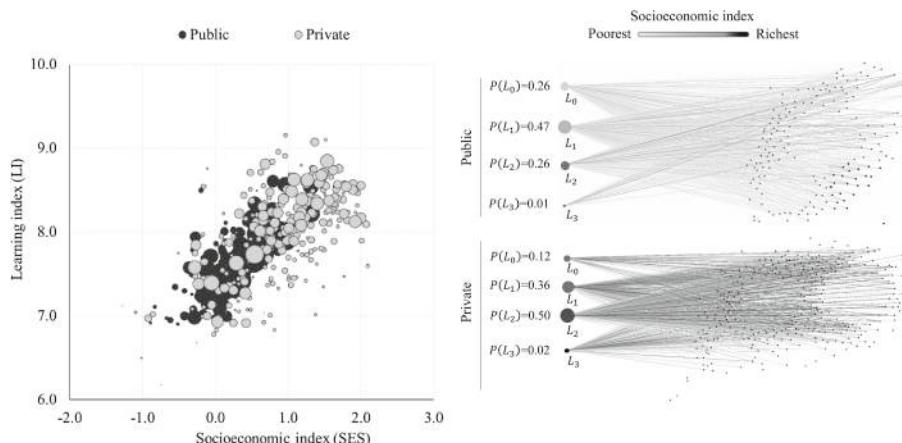


Fig. 3. Relationship between learning and socioeconomic indexes at school level (left side), and In-degree distribution of private and public networks (right side).

Moreover, right side of Fig. 3 shows a kind of bipartite data sources [21] for representing the networks of private and public schools, where darker edges refer to higher *SES* and the size of nodes are masses of probability for in-degree distribution $P(L_i)$ where Private-network has 309 nodes (schools) with 12 823 edges (41 per school) while Public-network has just 167 with 24 671 edges (148 per school), i.e., the private one has 85% more nodes than the public one, but its density is just 52%.

In addition, 70.6% of private schools are linked to L_0 through just 1 553 of their edges ($H = 0.12$ and $\lambda = 0.227$), while 97.0% of public are linked to the same node through 6 339 edges ($H = 0.26$ and $\lambda = 0.223$), a rate twice higher than in private sector, showing that *SES* is a key factor for educational deprivation due mainly to the influence of cultural capital, showing the lack in the capacity of the government to guarantee educational rights.

3 Ethnicity and Type of School Financing

The rurality of the area where the schools are located is also a factor that impacts on the learning gaps, its effect, combined with the type of school funding, indicates a huge variance among ethnic groups within the socioeconomic deciles. Figure 4 shows a Radviz plot, a non-linear multi-dimensional algorithm [22] and a dendrogram made with a cluster analysis based on $(SES_k^i, RA_k^i, TS_k^i, ET_k^i)$, this representation shows the hierarchical structure of inequality in learning deprivation for population groups, the points on the circle refer to three indicators attracting the students groups proportionally and the symbol size is proportional to the total deprivation given by δ .

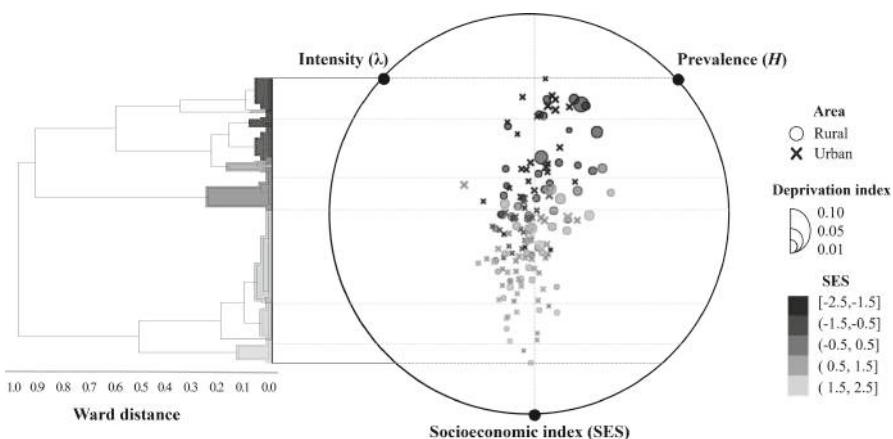


Fig. 4. Deprivation of learning by ethnicity, rurality and socioeconomic status.

As can be seen, the strong relationship between the SES and the prevalence distributes the groups in a vertical way showing borders between rural (circles) and urban areas (crosses).

Figure 4 shows that gaps between rural and urban areas are quite accentuated, especially among the poorest students in rural areas suffering the highest levels of prevalence and δ . In this sense, among the most deprived students in rural areas, Montubios exhibit the highest prevalence rates, followed by and Indigenous ($H = 0.64$) and Afro-Ecuadorians ($H = 0.55$). The clusters shown by the dendrogram points out that the highest SES students dominates private schools with the lowest levels of deprivation, with the exception of Montubios in rural areas, where no student attending private schools reaches the minimum level and where just 1 of 3 students in public schools is not deprived ($H = 0.63$), showing that social and ethnic classes are splatted into groups of students who have had different learning opportunities inside and outside of the schools, helping to understand how inequality evolves in a structural way [23].

4 Key-Factors for Public Policies

One of the most useful strategies for improving learning and closing gaps is micro-planning, however, to select which needs should be attended first for specific groups is always a great deal. For this reason, once the gaps in population groups are measured, we extend the network for introducing $\{FAL_{Cm}^j\}$ to the model and identify which variables are linked with specific population groups and recognizing those that should be intervened first, as well to simulating the variational aspects for establishing the order and control parameters for building more efficient and effective portfolios of policies at local level.

Prevalence rate of L_0 -students are related with networks strongly connected and this might be associated with Eigen-Centrality through measuring the influence of each factor for identifying how well connected the j -th factor (FAL_j) is and how many links have its connections. In this way, splitting L_0 -students in communities becomes in a very valuable tool for developing group-oriented strategies and avoid implementing the same actions for completely different needs. At this point, network simulation offers the advantage of building multiple scenarios across varying FAL parameters for recognizing and ranking the most relevant nodes to be attended by policymakers in a hypothetic but very specific situation.

Figure 5 shows a network build through the *ForceAtlas2* algorithm [24] to obtain the layout after a *Modularity process* (with parameter 0.073 at resolution of 0.254) for splitting richest from poorest students to find key factors for educational deprivation in both groups [25]. The Average Weighted Degree of the network is $AWD = 426.9$ and its density $D = 5.617$. Those large nodes appearing in the network as attractors correspond to two communities: the richest (SES_{10}), with degree of authority $A_{SES_{10}} = 0.158$ and just 20 factors —most of them showing very low centrality parameters— and poorest students (SES_1), with $A_{SES_1} = 0.987$ and 57 different factors associated with their deprivation.

For assessing the quality of its connections, PageRank is a well-known algorithm for providing accurate and clear information when looking for the factors dominating deprivation in each community [25, 26], which also might order nodes parametrically for selecting those who susceptible to be managed by policy and define ‘needs profiles’ for focusing actions on those variables susceptible to be managed by policy for specific zones or groups.

For example, a SES_1 profile based on Fig. 5 using the 10 most relevant factors might be:

‘Members of a household who currently receives the Human Development Bond and needs to work for a wage. Their parents have a very low level of education, they do not have a desktop computer neither Internet connection, and also have no books or just a very few. In their school, teachers arrive late to class, are not committed with learning and have low expectations about student’s future’.

As can be seen, this very detailed information is extremely helpful for developing based-evidence policies, to estimate and assign budget and increase the chances of having a successful deployment, the most wanted tool from policymakers.

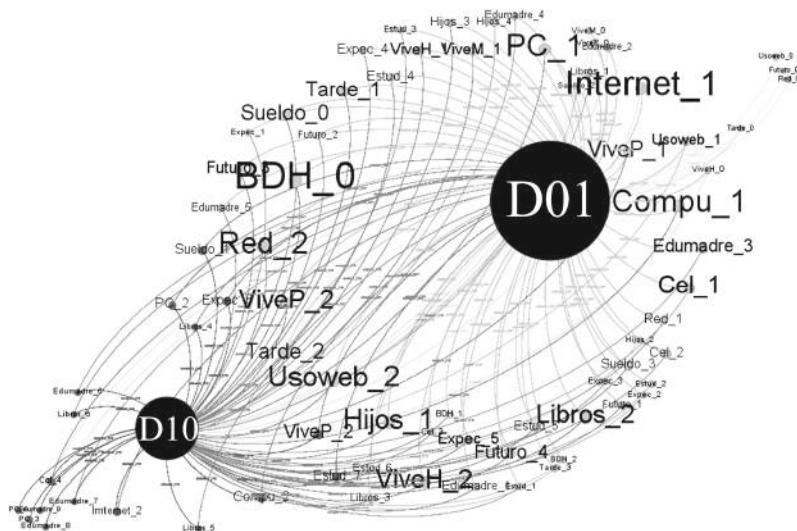


Fig. 5. Factor's network for richest and poorest students' communities. Centrality of nodes in each group points out the dissimilarities in factors provoking deprivation.

5 Discussion

This research finds a lot of evidence about the deep lack in equity in a Latin American country and its relationship with social determinants for synthetizing the complex structure of inequality on learning outcomes [27]. Measuring how far is each student of reaching the minimum level to ensure its right to education is crucial because the social deception of accessing education without guarantees of learning, converges at levels of precariousness similar to those found in people outside the scholar system [28]. Likewise, to identify a specific set of factors impacting a group of students that might not be relevant for others is a key issue for developing group-oriented policy [29].

This sequential network model allows to investigate the factors conditioning the exercise of the right to education and opens the possibility of reviewing how the type of school funding promotes inequality in different population groups at country level. Evidence found over deprivation among the socioeconomic groups shows the lack of attention to ethnicity, especially in the urban area. In this sense, when comparing the topologies of the networks of the richest versus the poorest was possible to estimate the gaps in educational outcomes that are associated to other factors like the influence of cultural capital, the context and the families of the students, beyond the schools.

Through this model is also possible to use LSA results for analyzing the interplay of socioeconomic status, rurality, type of school and ethnicity, bringing very useful information about the educational system at meso and macro levels from micro level interactions, offering valuable information for answering five main educational questions:

1. What is the level of deprivation of learning in students at the end of compulsory education?
2. Is learning deprivation associated with schools?
3. How deep is the inequality in the distribution of learning outcomes?
4. What factors are associated with educational deprivation in specific areas of the territory and population groups?
5. Considering limited resources and time, what factors and in what order should they be established in a group-oriented public policy?

Finding answers to these questions using a dataset from a LSA represents a tipping point for integrating thinking tools into current theoretical frameworks to find non-trivial relationships between social conditions and educational deprivation in historically marginalized population groups. In this sense, estimated gaps in deprivation due to rural area and ethnicity, point out the lack of effective inclusion policies, especially of those groups of students who have had different learning opportunities and highlights how inequality is structurally generated in the country.

However, though it is possible to infer the central role played by the schools in determining the more relevant factors for educational deprivation in different SES classes, more research is needed to understand the interactions between the school context and ethnic diversity. Nevertheless, the implications for public policies are: (1) programs aimed at indigenous self-identified students from the poorest quintile should be reviewed in all its pedagogical aspects to guarantee the achievement of meaningful learning; (2) the interculturality programs in the urban area should be broadened so that the ethnic groups can reach the minimum standards at same rates than the other ethnical groups and, (3) it is necessary to better regulate the private schools in the rural area to help close the gap in the absolute deprivation rate.

There is no doubt that the network analysis shows a great vein of scientific development that has not yet been explored to improve knowledge about educational systems, addressing the greatest challenges of educational research, and helping policymakers to develop strategies for an inclusive and equitable education for all.

References

1. OECD: Equity in education: breaking down barriers to social mobility. PISA, OECD Publishing, Paris (2018)
2. UN DESA: The sustainable development goals report 2018, UN, New York (2018)
3. Sánchez-Restrepo, H.: Equity: the focal point of educational quality. National Educational Evaluation Policy Gazette in Mexico, Year 4. (10), pp. 42–44 (2018)
4. United Nations Children's Fund (Unicef): The investment case for education and equity, Washington, DC (2015)
5. Keeley, B., Little, C.: The State of the World's Children 2017: Children in a Digital World. UNICEF, New York (2017)
6. Samman, E.: SDG progress: fragility, crisis and leaving no one behind: report (2018)
7. Gray, J., Kruse, S., Tarter, C.J.: Enabling school structures, collegial trust and academic emphasis: antecedents of professional learning communities. Educ. Manag. Adm. Leadersh. 44(6), 875–891 (2016)

8. Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación – LLECE. Agenda (2018). <http://www.unesco.org/new/en/santiago/press-room/newsletters/newsletter-laboratory-for-assesSSent-of-the-quality-of-education-llece/>
9. Motl, T.C., Multon, K.D., Zhao, F.: Persistence at a tribal university: factors associated with second year enrollment. *J. Divers. High. Educ.* **11**(1), 51 (2018)
10. Tracey, B., Florian, K. (eds.): *Educational Research and Innovation Governing Education in a Complex World*. OECD Publishing, Paris (2016)
11. Davis, B., Sumara, D.: *Complexity and Education: Inquiries into Learning, Teaching, and Research*. Routledge, New York (2014)
12. Fox, J.: *Applied Regression Analysis and Generalized Linear Models*. Sage Public, Thousand Oaks (2015)
13. Moll, K., Göbel, S.M., Gooch, D., Landerl, K., Snowling, M.J.: Cognitive risk factors for specific learning disorder: processing speed, temporal processing, and working memory. *J. Learn. Disabil.* **49**(3), 272–281 (2016)
14. van der Maas, H., Kan, K., MarSSan, M., Stevenson, C.: Network models for cognitive development and intelligence. *J. Intell.* **5**(2), 1–17 (2017). <https://doi.org/10.3390/intelligence5020016>
15. Gelman, A., Imbens, G.: Why high-order polynomials should not be used in regression discontinuity designs. *J. Bus. Econ. Stat.* 1–10 (2018)
16. Barabási, Albert-László, Pásfai, Márton: *Network Science*. Cambridge University Press, Cambridge (2016)
17. Johnson, J., Fortune, J., Bromley, J.: Systems, networks, and policy. In: *Non-Equilibrium Social Science and Policy*, pp. 111–134. Springer, Cham (2017)
18. Grandjean, M.: Complex structures and international organizations [Analisi e visualizzazioni delle reti in storia. L'esempio della cooperazione intellettuale della Società delle Nazioni]. *Memoria e Ricerca* **25**(2), 371–393 (2017)
19. Borsboom, D., Molenaar, D.: Psychometrics. In: Wright, J. (ed.) *International Encyclopedia of the Social & Behavioral Sciences*, vol. 19, pp. 418–422. Elsevier, Amsterdam (2015)
20. Bracho, T.: Índice de déficit en competencias? Avanzamos hacia la garantía del derecho a la educación? en *Reformas y Políticas Educativas*, No. 4, September–December 2017. FCE, México (2017)
21. van Borkulo, C.D., Borsboom, D., Epskamp, S., Blanken, T.F., Boschloo, L., Schoevers, R.A., Waldorp, L.J.: A new method for constructing networks from binary data. *Sci. Rep.* **4** (5918), 1–10 (2014)
22. Leban, G., Zupan, B., et al.: Vizrank: data visualization guided by machine learning. *Data Min. Knowl. Discov.* **13**(2), 119–136 (2006)
23. Koller, D., Friedman, N.: *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge (2009)
24. Jacomy, M., Venturini, T., Heymann, S., Bastian, M.: ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* **9**(6), 1–18 (2013)
25. Forbush, K., Siew, C., Vitevitch, M.: Application of network analysis to identify interactive systems of eating disorder psychopathology. *Psychol. Med.* **46**(12), 2667–2677 (2016)
26. Isvoranu, A.M., van Borkulo, C.D., Boyette, L., Wigman, J.T.W., Vinkers, C.H., Borsboom, D., Group Investigators: A network approach to psychosis: pathways between childhood trauma and psychotic symptoms. *Schizophr. Bull.* **43**(1), 187–196 (2017)
27. Sánchez-Restrepo, H., Louçã, J.: Topological properties of inequality and deprivation in an educational system: unveiling the key-drivers through complex network analysis. In: *International Conference on Human Systems Engineering and Design: Future Trends and Applications*, pp. 469–475. Springer, Cham, September 2019

28. Tsatsaroni, A., Evans, J.: Adult numeracy and the totally pedagogised society: PIAAC and other international surveys in the context of global educational policy on lifelong learning. *Educ. Stud. Math.* **87**(2), 167–186 (2014)
29. Menefee, T., Bray, T.M.: Education in the Commonwealth: Quality Education for Equitable Development. Commonwealth Secretariat (2015)



'I Ain't Like You' A Complex Network Model of Digital Narcissism

Fakhra Jabeen^(✉), Charlotte Gerritsen, and Jan Treur

Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
fakhraikram@yahoo.com, {cs.gerritsen, j.treur}@vu.nl

Abstract. Social media like Twitter or Instagram play a role of fertile ground for self-exhibition, which is used by various narcissists to share their frequent updates reflecting their narcissism. Their belief of saving and assisting others, make them vulnerable to the feedback of others, so their rage is as dangerous as their messiah complex. In this paper, we aim to analyse the behaviour of a narcissist when he is admired or receives negative critics. We designed a complex adaptive mental network model of the process of narcissism based on the theories of neuroscience and psychology including a Hebbian learning principal. The model was validated by analyzing Instagram data.

Keywords: Narcissistic rage · Narcissism · Complex mental network

1 Introduction

"My character has ever been celebrated for its sincerity and frankness, and in a cause of such moment as this, I shall certainly not depart from it. (Pride and Prejudice:56)".

Narcissism was addressed by various fiction characters like Lady Catherine (Pride and Prejudice), who always tried to be a savior and was overly concerned about how others think of her. Now in the age of technology, they can be observed by people who excessively use social media, like Twitter or Instagram [1], to share their lifestyle updates. These online communities are targeted as fertile grounds for self-presentation and getting the appraisal, especially by millennials, who tend to use whatever chrism and social skills to become quantifiers or influencers [1]. The exponential growth of audience makes narcissists more vulnerable to negative feedback, and usually make them unhappy about others [2]. Narcissistic rage is a psychological construct that addresses a negative reaction of a narcissist person, when he or she assumes that his self-worth is in danger.

Narcissistic rage is a common outcome due to lack of empathy, and it is the more related outcome to an ego-threatening action rather than self-esteem. It is to be made clear, that this is not related to self-esteem, but to the desire of self-admiration [3]. There are two types of narcissism, discussed in the literature: 'overt' or 'covert'. They both share grandiose fantasies, feeling of superiority, but covert narcissists, don't reveal their true self, thus may show aggression towards others [4]. In literature, different studies were conducted with respect to the psychology and neuroscience of a narcissist in certain surrounding conditions [3, 5]. Also, Artificial intelligence is used to identify a

narcissist [6]. However, no study has been presented in the domain of Mental Network Modeling, which can address: (a) the mental organization of a narcissist, (b) how his mental processes learn from experience, (c) how to relate a presumed narcissist with his or her cognitive behavior by taking social feedback into account, with respect to his social interaction.

In this paper, we present a complex-adaptive network model of a narcissist, based on psychological and neurological studies. We address non-trivial interaction of a narcissist brain (a) during a reward-seeking behavior, and (b) a consequence of an unwanted remark. We validated our model by case studies. Section 2 follows related work while Sect. 3 presents the model of a narcissist. Section 4 discusses the simulation scenarios, while Sect. 5 validates the model using public data through Instagram. Section 6 concludes the paper.

2 Related Work

This section explains the psychological and, neurological perspective of a narcissist. On the one hand, it provides literature: how a narcissistic behaves when (a) admired or (b) negatively criticized. On the other hand, it will address the problem in relation to complex networks and artificial intelligence.

Psychologically, a narcissist exhibits a higher tendency for self-presentation [2]. Many studies show an association between narcissism and reward-seeking behavior [3]. A survey indicates that Instagram is widely used service among other social networks, to exhibit grandiose narcissism. A narcissist who receives added appreciation often appears to be compassionate and happy [1]. A survey showed that people, who are not satisfied with their appearance are more vulnerable to anger, due to lack of empathy. As a result, narcissists are prone to bullying or violence [3, 5].

In a cognitive and neurological aspect, we would like to discuss cognitive parts of brain, hormones, along with the neurotransmitters, which process the self-relevant stimuli. A narcissist seeks admiration, which activates brain regions, like the Prefrontal Cortex, Anterior Cingulate Cortex (ACC), anterior Insula and Temporal lobe, which is strengthened during self-enhancement and mentalizing [7]. High activations in the anterior insula indicate focus on oneself or representing selfishness [5]. It is indicated that ventral striatum is involved with ACC, and get activated during reward-seeking behaviour. It depends on pre-synaptic and post-synaptic activations along with dopamine release. The facial attractiveness strengthens the synaptic transmission due to contingent feedback and dopaminergic projections [8]. ACC is also related to negative emotional valence which may lead to aggression. Also, stress faced in a social competitive environment makes the brain more vulnerable to experience anxiety [9].

Like in reward-seeking, hormones and neurotransmitters also play a role in aggression. For instance, noticeable levels of progesterone, testosterone and low levels of corticosterone help in mediating aggression along with γ -aminobutyric acid (GABA) receptors activation, due to anxiety [10]. Decreased levels in 5-serotonin (5-HT) leads to aggression. Further, vascular endothelial growth factor (VEGF) is a signal protein in the hippocampal region, which is effected in psychological stress. Damage in its microstructure and decrease in synaptic connections, influence the release of

neurotransmitters in the hippocampal region during stress. This decrease in synaptic plasticity hinders the message transfer to the central nervous system along with the changes in the brain structure, learning, and memory [10]. Long term effects of stress can lead to long-term genetic and epigenetic metaplastic effects [11]. Also, a presynaptic receptor 5-HT_{1A}, located in 5-HT has a greater density in people with high aggression, along with brain regions that are related to impulsive control [12].

A study was also presented, which incorporated machine learning techniques to identify a narcissist [6]. A temporal causal model is 'discussed in the context of esteem' [13]. However, no research was found, which could address the vulnerability in a narcissist and his reaction over certain feedback.

3 Complex Adaptive Mental Network Model of a Narcissist

In this section, a complex adaptive mental network of a narcissist is presented based upon studies addressed earlier, using a multilevel reified architecture [14, 15]. This is a layered architecture, where the temporal and adaptive dynamics of a model are represented in layers, from the base model to the evolution of the complex network by first and second order adaption principles, along with its mathematical representation.

A base temporal-causal model refers to a conceptual representation of a real-world scenario, depicted by *states* and *connections* where a *connection* designates a causal relationship among *states*. For example, consider two states X and Y , if Y is affected by X then $X \rightarrow Y$ is a causal relationship. More specifically, the activation value of Y is the *aggregated impact* of all influencing states along with X . The influencing states have different *activation levels* and *connection weights* to influence Y , that has a *speed factor* indicating the timing of influence. Such a temporal-causal model is characterized by [16]:

Connection weight $\omega_{X,Y}$ indicates how strong state X influences state Y . The magnitude varies between 0 and 1. A suppression effect is categorized by a negative connection.

Speed factor η_Y indicates that how fast a state Y changes its value upon a causal impact; values range from [0–1].

Combination function $c_Y(\cdot)$ is chosen to compute the causal (aggregated) impact of all incoming states ($X_i : i = 1 \text{ to } N$) for state Y . Certain standard combination functions are defined already to compute the aggregated impact of Y .

In Fig. 1, layer I presents the base model with 38 states, depicting the mental organization and psychology of a narcissist. It shows reactions of a narcissist when he receives admiration or negative criticism. The extended structure of the model is also explained by Layer II and III. A concise explanation of each state in figures is specified in Tables 1 and 2.

In layer I, a narcissist receives a feedback while using social media (ws_s, ss_s), after sharing his picture or a status. Feedback can be a positive ($ws_{pf}, ss_{pf}, srs_{pf}$) or a negative ($ws_{nf}, ss_{nf}, srs_{nf}$) remark which can make him happy or he can feel hurt. On receiving a compliment like 'you are awesome' his self-belief (bs_+) evaluates it as a positive remark (eval+), thus leads to a happy reaction (es_{happy} : i.e. a gratitude). Brain related

parts (striatum, pfc and insula) get activated more than usual, along with feelings of self-reward (fs_{reward}) and self-love (fs_{love}). This behavior increases by experiencing the same kind of feedbacks over the time (adaption/learning).

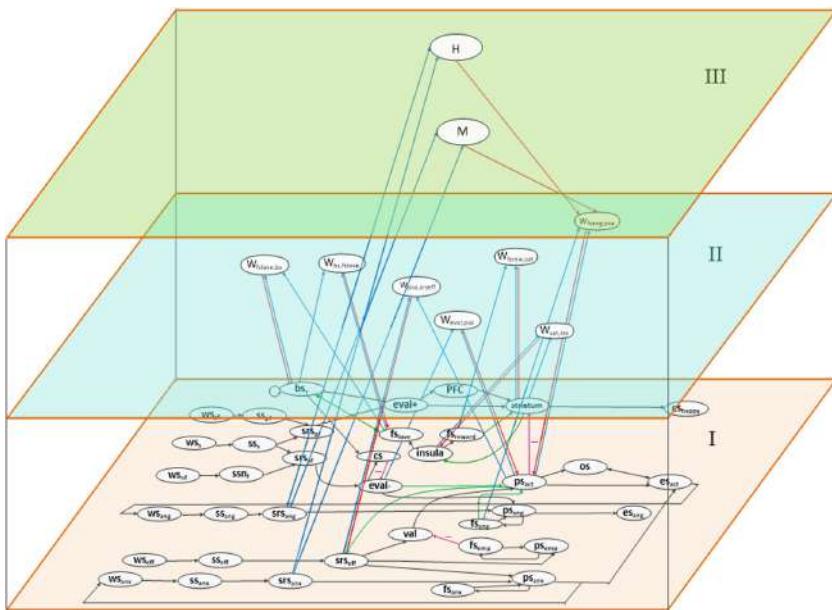


Fig. 1. Reified network architecture for a narcissist person.

Upon a negative critic (ws_{nf} , ss_{nf} , srs_{nf}), a narcissist usually disagrees due to high ego/self-belief, and gets angry, which also induces anxiety in him. To explain it further, consider a poor remark ‘you are ugly’, activates state ($eval-$). Self-belief (bs_+) tries to suppress $eval-$ through control state (cs). However, $eval-$ is too strong to be suppressed. It has dual influence: a) It stimulates anger (ps_{ang}) along with its body loop (ws_{ang} ; ss_{ang} ; srs_{ang} ; fs_{ang} ; ps_{ang} ; es_{ang}), e.g. ‘a raised eyebrow’ and, b) preparation state of action (ps_{act}) is also activated by an aggregated impact of $eval-$, val and fs_{ang} . Here ‘ val ’ is the valuation state which doesn’t get activated if a person has empathy (fs_{emp} ; ps_{emp}), which narcissist lacks. So, in turn ps_{act} activates the execution state (es_{act}), i.e. an angry reply. This reaction involves a thought process about predicted effect (ws_{eff} ; ss_{eff} ; srs_{eff}), and eventually anxiety is induced (ws_{anx} ; ss_{anx} ; srs_{anx} ; ps_{anx} ; fs_{anx}). However, anxiety still elevates reaction (es_{act}). Please note, as social media is not controllable, therefore es_{act} doesn’t influence ws_{nf} . Black horizontal arrows (layer I) show non-adaptive causal relations, while green show the adaptive ones. Purple arrows shows suppression from one state to another.

Table 1. Categorical explanation of states of base model (layer I).

Categories		References
<i>Stimulus states</i>		<i>Stimulus is sensed and leads to representation: p51 [16]</i>
ws _i	World state. <i>i</i> = stimulus (s); positive/negative feedback (pf/nf)	
ss _i	Sensory state. <i>i</i> = stimulus; pf/nf	
srs _i	Representation state <i>j</i> = pf/nf	
<i>Attribution/evaluation states</i>		<i>Narcissism involves states for self-enhancement and mentalizing [7]</i>
eval+	Positive evaluation of feedback	
eval-	Negative evaluation of feedback	
<i>Happiness related states</i>		<i>fMRI studies show activations at or near dopaminergic midbrain nuclei and the VS that correlate with both reward expectation and reward prediction errors... [8]</i>
bs ₊	Self-belief state	
striatum	Ventral Striatum: brain part	
PFC	Prefrontal Cortex: brain part	
fs _{reward}	Feeling state of reward (Amygdala)	
fs _{love}	Feeling state self-love (Amygdala)	
es _{happy}	Execution state of happiness	
insula	Anterior Insula: brain part	
<i>Anger related action states</i>		<i>... predictive and inferential processes contribute to conscious awareness ... action" p212 [16]</i>
os	Ownership state	
ps _{act}	Preparation state of action	
es _{act}	Execution state of action	
<i>Body loops (anger and anxiety)</i>		<i>Body loop via the expressed emotion is used to generate a felt emotion by sensing the own body state... the emergence of states [16]</i>
ws _i	World state <i>i</i> = anger/anxiety (ang /anx)	
ss _i	Sensor state <i>i</i> = anger/anxiety (ang/anx)	
ps _i	Preparation state of <i>i</i> = ang/anx	

(continued)

Table 1. (continued)

Categories		References
fs_i	Feeling state $i = ang/anx$	
es_{ang}	Execution state (Expression of anger)	
<i>Predicted effect of action</i>		<i>Sensory feedback provides more precise evidence about actions and their effects. p212 [16]</i>
ws_{eff}	World state of effect	
ss_{eff}	Sensor state of effect	
srs_{eff}	Representation state of effect	
<i>Control states</i>		<i>ACC become active in parallel with the insula when we experience feelings". p109 [16]</i>
cs	Control state	
val	Valuation state	

Layer II, presents the plasticity of the model, while layer III represents meta-plasticity. Each layer is connected by upward (blue) and downward (red) arrows. Layer II incorporates Hebbian principle [17] by upward (blue) and downwards (red) arrows with states at layer I [14]. For instance, for a narcissist receiving appreciation, feeling of reward (fs_{reward}) stimulates ventral striatum (striatum) [8], is represented by connection $fs_{reward} \rightarrow$ striatum. Its increases over the time, this increase is due to pre-synaptic (fs_{reward}) and post-synaptic (striatum) states depend on dopamine based activations (dopamine release) and this is represented through $W_{fs_{reward},\text{striatum}}$.

Layer III represents meta-plasticity, through states M and H, that can control the learning of state $W_{fs_{ang},ps_{act}}$ at layer-II. Former indicates persistence, while later specifies the learning rate of $fs_{ang} \rightarrow ps_{act}$. Usually, every W state at layer II has meta-plasticity, because of presynaptic and post synaptic states involvement in gaining experience [14]. However, to keep the complex network simple, we only meta-plasticized the angry reaction of a narcissist, which is due to changes in the synaptic connections [10, 11].

Table 2. Explanation of states in layer II and III.

States per layer		References
<i>Layer II (plasticity/omega states)</i>		
1. $W_{fs_{love},bs}$	For $fs_{love} \rightarrow bs$	1–4: Potentiation in the striatum depends not only on strong pre- and postsynaptic activation ... reward prediction ... modify behavior [8]
2. $W_{bs,fs_{love}}$	For $bs \rightarrow fs_{love}$	5–7: Presynaptic somatodendritic 5-HT1... people with a high level of aggression, there is a greater density ... with impulse control [12]
3. $W_{\text{striatum},\text{insula}}$	For striatum \rightarrow insula	
4. $W_{fs_{reward},\text{striatum}}$	For $fs_{reward} \rightarrow$ striatum	
5. $W_{eval- \rightarrow ps_{act}}$	For eval- \rightarrow ps_{act}	
6. $W_{ps_{act},srs_{eff}}$	For $ps_{act} \rightarrow srs_{eff}$	
7. $W_{fs_{ang},ps_{act}}$	For $fs_{ang} \rightarrow ps_{act}$	
<i>Layer III (meta-plasticity)</i>		<i>Damage to neurons in hippocampal CA3 area and microstructure of synapse indicates that anger... harms plasticity.... [10]</i>
H	Speed factor for $W_{fs_{ang},ps_{act}}$	
M	Persistence factor for $W_{fs_{ang},ps_{act}}$	

For computation of impacts most speed factors η and connection weights ω at layer I, have values between 0 and 1. Please note that for $\mathbf{W}_{fs_{ang},ps_{act}}$ the speed factor is adaptive, i.e. based on the reification state H , therefore we used adaptive combination function for computation. Here Δt is 0.5. We used three type of combination functions for the simulation of our model (Fig. 1):

- (a) For 24 states ($ws_s; ss_{pf}; ss_{nf}; ss_s; srs_{pf}; pfc; eval+; es_{happy}; eval-; os; es_{act}; ss_{ang}; srs_{ang}; ps_{ang}; es_{ang}; fs_{ang}; fs_{emp}; ws_{eff}; ss_{eff}; ws_{anx}; ss_{anx}; srs_{anx}; ps_{anx}; fs_{anx}$), we used Euclidian function, with order $n > 0$ and scaling factor λ as the sum of connection weights of a particular state:

$$\text{eucl}_{n,\lambda} \cdot (V_1, \dots, V_k) = \sqrt[n]{(V_1^n + \dots + V_k^n) / \lambda}$$

- (b) For 14 states ($srs_{nf}; bs; \text{striatum}; fs_{love}; fs_{reward}; \text{insula}; cs; ps_{act}; ws_{ang}; val; ps_{emp}; srs_{eff}; H; M$), we used the **alogistic** function with positive values of steepness σ and threshold τ less than 1:

$$\text{alogistic}_{\sigma,\tau}(V_1, \dots, V_k) = \left[\left(1 / \left(1 + e^{-\sigma(V_1 + \dots + V_k - \tau)} \right) \right) - 1 / (1 + e^{\sigma\tau}) \right] (1 + e^{-\sigma\tau})$$

where V is the single impact computed by product of state values and its connection weight i.e. $\omega_{X,Y} X(t)$.

- (c) Lastly, for 7 adaptation states ($\mathbf{W}_{bs,fs_{love}}; \mathbf{W}_{fs_{love},bs}; \mathbf{W}_{\text{striatum},\text{insula}}; \mathbf{W}_{fs_{reward},\text{striatum}}; \mathbf{W}_{ps_{act},srs_{eff}}; \mathbf{W}_{fs_{ang},ps_{act}}; \mathbf{W}_{eval-,ps_{act}}$) we used Hebbian learning:

$$\text{hebb}_\mu(V_1, V_2, W) = V_1 V_2 (1 - W) + \mu W$$

Mathematically, a reified-architecture based model is represented as [14]:

1. At every time point t , the activation level of state Y at time t is represented by $Y(t)$, with the values between $[0, 1]$.
2. Impact of state X on state Y at time t is represented by $\text{impact}_{X,Y}(t) = \omega_{X,Y} Y(t)$; where $\omega_{X,Y}$ is the weight of connection $X \rightarrow Y$.
3. Special states are used to model network adaptation based on the notion of reification network architecture. For example, $\mathbf{W}_{X,Y}$ represents an adaptive connection weight $\omega_{X,Y}(t)$ for the connection $X \rightarrow Y$, while \mathbf{H}_Y represents an adaptive speed factor $\eta_Y(t)$ of state Y . Similarly, $\mathbf{C}_{i,Y}$ and $\mathbf{P}_{i,j,Y}$ represent adaptive combination functions $c_Y(\cdot, t)$ over time and its parameters respectively. Combination functions are built as a weighted average from a number of basic combination functions $bcf_i(\cdot)$, which take parameters $P_{i,j,Y}$ and values V_i as arguments. Universal combination function $c^*_Y(\cdot)$ for any state Y is defined as:

$$c^*_Y(S, C_1, \dots, C_m, P_{1,1}, P_{2,1}, \dots, P_{1,m}, P_{2,m}, V_1, \dots, V_k, W_1, \dots, W_k, W) = W + S[C_1 bcf_1(P_{1,1}, P_{2,1}, W_1 V_1 \dots W_k V_k) + \dots + C_m bcf_m(P_{1,m}, P_{2,m}, W_1 V_1, \dots, W_k V_k) / (C_1 + \dots + C_m) - W]$$

where at time t:

- variable S is used for the speed factor reification $\mathbf{H}_Y(t)$
- variable C_i for the combination function weight reification $\mathbf{C}_{i,Y}(t)$
- variable $P_{i,j}$ for the combination function parameter reification $\mathbf{P}_{i,j,Y}(t)$
- variable V_i for the state value $X_i(t)$ of base state X_i
- variable W_i for the connection weight reification $\mathbf{W}_{X_i,Y(t)}$
- variable W for the state value $Y(t)$ of base state Y .

4. Based on the above universal combination function, the effect on any state Y after time Δt is computed by the following *universal difference equation* as:

$$Y(t + \Delta t) = Y(t) + [\mathbf{c}_Y^*(\mathbf{H}_Y(t), \mathbf{C}_{1,Y}(t), \dots, \mathbf{C}_{m,Y}(t), \mathbf{P}_{1,1}(t), \dots, \mathbf{P}_{1,m}(t), \mathbf{P}_{2,m(t)}, X_1(t), \dots, X_k(t), \mathbf{W}_{X_1,Y}(t), \dots, \mathbf{W}_{X_k,Y}(t), Y(t)) - Y(t)] \Delta t$$

which also can be written as a *universal differential equation*:

$$\frac{dY(t)}{dt} = \mathbf{c}_Y^*(\mathbf{H}_Y(t), \mathbf{C}_{1,Y}(t), \dots, \mathbf{C}_{m,Y}(t), \mathbf{P}_{1,1}(t), \mathbf{P}_{2,1}(t), \dots, \mathbf{P}_{1,m}(t), \mathbf{P}_{2,m}(t), X_1(t), \dots, X_k(t), \mathbf{W}_{X_1,Y}(t), \dots, \mathbf{W}_{X_k,Y}(t), Y(t)) - Y(t)$$

Our Simulation environment was implemented in MATLAB, and receives input of the characteristics of a network structure represented by role matrices. A role matrix is a compact specification with the concept of the role played by each state with specified information. Detailed information for the designed model can be found online [15, 18].

4 Simulation Scenarios

Simulation scenarios are used to verify the dynamic properties of the model by simulating real-world processes. Simulations of the model are addressed below by two kinds of comment: (a) appreciation or (b) a negative critic.

4.1 Reaction of a Narcissist When Appreciated

Twitter, Facebook [2] or Instagram [1] are a few popular platforms used by narcissists to gain recognition and prominence. Many scholars have been studying “messiah complex” of Donald Trump [19]. To illustrate his behavior related to reward-seeking and self-love, we can take his tweet as an example of self-love and self-reward:

“...my two greatest assets ... mental stability and being, like, really smart ... I went from VERY successful businessman, to top T.V Star....” (Tweeted: 1:27 PM – Jan 6, 2018)

Figure 2 shows that when positive feedback arrives, eval+ (purple) is activated, which activates PFC and reward seeking process through striatum. These activations along with self-belief (bs_+) make feelings of self-love (fs_{love}) and reward (fs_{reward}) high.

As a result, he expresses his gratitude (es_{happy}). Here, insula indicates the self-thinking process, which increases self-love and feeling of reward more and more.

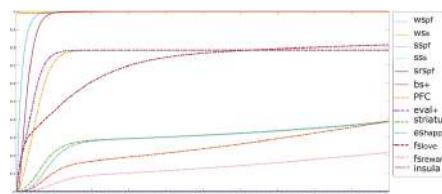


Fig. 2. Simulation of the model when $ws_{pf} = 1$ and $ws_{nf} = 0$.

4.2 Reaction of a Narcissist on a Negative Feedback

A narcissist, observing ego-threatening feedback, doesn’t hesitate to share his reactions. While studying Donald Trump, we can identify his overtly reactions easily on any of stances, which are evaluated as threat to his ego. For example, consider the tweet of Donald Trump as:

“... world class loser, Tim O’Brien, who I haven’t seen or spoken ... knows NOTHING about me ... wrote a failed hit piece book...” (Tweeted: 6:20 AM – Aug 8, 2019) [20].

Figure 3 explains the behavior of a narcissist over negative feedback (shaded region) in episodes: i.e. from 100–200; and 300–400. Reward related states in the white region (e.g. fs_{love} , fs_{reward} ,...) are suppressed, $eval$ - (purple) is raised due to srs_{nf} and cs . It is responsible for two more activations. Firstly, it activates body loop of anger in red (ws_{ang} , ss_{ang} , srs_{ang} , fs_{ang} , ps_{ang} and es_{ang}), which in turn activates the effect related states in green (ws_{eff} , ss_{eff} , and srs_{eff}). Secondly, it stimulates urge to respond by ps_{act} , os and es_{act} , however this leads him to experience anxiety in blue (ws_{anx} , ss_{anx} , srs_{anx}). Narcissists lacks empathy (fs_{emp} and ps_{emp}), therefore valuation (val) is not suppressed. As a result he replies to such feedback. Over the time (X-axis), it can be seen that such behavior continues to aggravate (by Hebbian learning) by similar experiences. Detailed explanation of each curve can be found online [18].

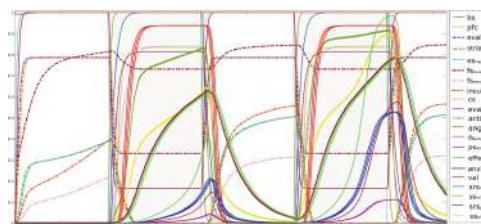


Fig. 3. Simulation of the model with alternative episodes of $ws_{nf} = 1$ or $ws_{pf} = 1$.

5 Model Validation and Analysis

The model is validated by three public Instagram users, with presumably narcissistic characteristics (names are not disclosed here). Reasons for choosing Instagram are: (a) users show more tendency towards narcissism [1] and, (b) this platform contain conversational elements, which can be helpful for our study.

5.1 Extraction and Analysis of Data

We extracted data from Jan 2017–July 10, 2019 and tested for three hypothesis: (A) *Frequency of sharing posts increases over a period of time.* (B) *Average number of likes increase with the number of shared posts.* (C) *They get happy when admired, but covert/overt behavior can be seen towards a critic.*

To come up with these hypotheses, data was analyzed in two steps. First, we analyzed a number of posts shared with the average number of likes per month. Figure 4-a shows that each profile user 1/2/3 tends to share more by the passage of time. Similarly, Fig. 4-b indicates that the number of average likes a profile receives increases. For example, Profile 1 (indicated by blue) 2 posts with avg. likes (817.5) (in Jan 2017), raised by 25 posts with avg. likes 7045.5 (July 2019). Please note, we didn't consider the number of followers the user had in mentioned duration, which makes average number of likes a bit subjective. However, the trend line of Fig. 4-a, indicates that Profile1 finds a good reason to share updates with increasing frequency of 25 posts per month.

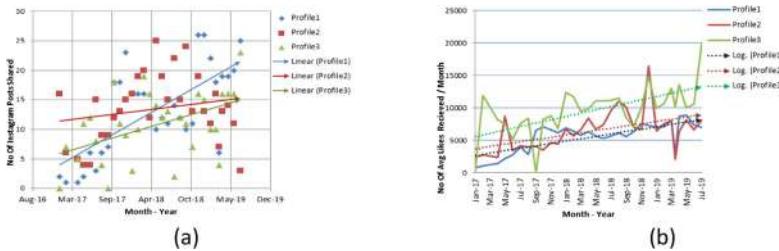


Fig. 4. Number of (a) Posts shared, (b) Average Likes received during a period of time

Our third hypothesis is related to the temporal analysis of data. Each post is deeply analyzed (Fig. 5). We selected conversations in each post, where profile user participated. We used Vader Sentiment Analysis [21] to analyze the first message in the conversation as positive, neutral or negative ones. Looking closely, it was revealed that various elements in negative conversation were an expression of positive assurance, but were misclassified (e.g. comment “🔥🔥 Fierce as fuck” score = -0.54). At this point, we used rule-based classification using further characteristics of data. Algorithm can be found online [18].



Fig. 5. Steps taken to interpret the behavior of each user

A stacked graph of three categories of conversation for each profile was plotted (Fig. 6). In general, it can be seen that most of the feedbacks received were positive or neutral, which provides a good reason for frequent posting on Instagram (Fig. 4). A mixed ratio of conversations can be seen. Like Profile 1, the users reaction was never negative in 2017. Similarly, Profile 2 has more in 2018 or 2019. On the contrary, Profile 3, had more negative conversations in 2017, a reason can be that he doesn't indulge into conversations anymore.

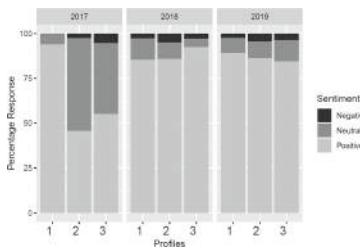


Fig. 6. Conversation ratio per profile from Jan 2017–July 2019

However, it can be concluded that narcissistic people do not hesitate to react over a negative comment in an overtly or covertly manner. Anxious behavior is not studied, as authors don't know them personally. However, it would not be wrong to say that learning from experience facilitates users to go further for sharing in their lifestyle, or responding to different followers. Our model also addresses the experience and reactions after experiences of the users.

5.2 Exhibition of Learning Experience in the Model

While looking into Fig. 4, we can see the complex learning behavior between all three layers over the time (in episodes). As addressed in Sect. 4, that an urge to respond (action: ps_{act} ; es_{act} ; os) increases on the basis of predicted effect (effect: ws_{eff} , ss_{eff} and srs_{eff}). Similarly, reward related states (striatum, fs_{reward} , fs_{love} , insula) are also elevated than before when a narcissist is admired. This can also be seen by the adaptive states (Layer II and III) shown in Fig. 7. For example, considering $W_{eval \rightarrow ps_{act}}$ (purple), we can see that it start increasing its value (e.g. from 0.2) in every negative episode (to 0.7). Similarly, M (brown) and H (blue) increase in every negative episode and suppressed otherwise. However, it can be seen that due to meta-plasticity, $W_{fs_{ang} \rightarrow ps_{act}}$ was not much raised (shaded region), which indicates that due to synaptic plasticity learning is effected [10], the similar pattern is observed during analysis of data, that action is followed if effect is assumed to be higher, i.e. if the feedback is really hurting the

esteem of a narcissist then he will go towards an angry reaction and a happy gesture is observed upon admiration, the reason can be he wants to show that he is loved by many individuals.

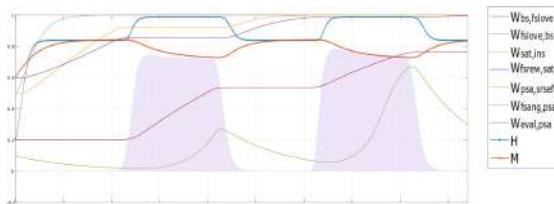


Fig. 7. Effects of plasticity (W states) and metaplasticity for $W_{fsang \cdot ps_{act}}$ (M and H)

6 Conclusion

In this paper, we discussed the complex adaptive mental network model of the processes of a narcissist, who reacts in different ways after having a positive, or negative feedback, and how prior experiences play role in learning and responding through different layers of reified network architecture. We tested our network model on Instagram data which is assumed to be a fertile ground for people with a narcissistic personality. Through the analysis of temporal data obtained from three users of Instagram, it was concluded that our model indeed depicts the behavior of a narcissist and reflects his/her joy or rage upon feedback.

In future research, we aim to extend our work, to incorporate how to react to a narcissistic person and explore his traits further. Moreover, we aim to detect and model how to support these complex behaviors among narcissists.

References

1. Moon, J.H., Lee, E., Lee, J.-A., Choi, T.R., Sung, Y.: The role of narcissism in self-promotion on Instagram. *Pers. Individ. Differences.* **101**, 22–25 (2016)
2. Wang, D.: A study of the relationship between narcissism, extraversion, drive for entertainment, and narcissistic behavior on social networking sites. *Comput. Hum. Behav.* **66**, 138–148 (2017)
3. Bushman, B.J., Baumeister, R.F.: Threatened egotism, narcissism, self-esteem, and direct and displaced aggression: does self-love or self-hate lead to violence? *J. Pers. Soc. Psychol.* **75**(1), 219 (1998)
4. Fan, C., Chu, X., Zhang, M., Zhou, Z.: Are narcissists more likely to be involved in cyberbullying? Examining the mediating role of self-esteem. *J. Interpers. Violence.* **34**(15), 3127–3150 (2019)
5. Fan, Y., Wonneberger, C., Enzi, B., de Greck, M., Ulrich, C., Tempelmann, C., Bogerts, B., Doering, S., Northoff, G.: The narcissistic self and its psychological and neural correlates: an exploratory fMRI study. *Psychol. Medicine.* **41**, 1641–1650 (2011)

6. Neuman, Y.: Computational Personality Analysis: Introduction. Practical Applications and Novel Directions. Springer, Cham (2016)
7. Olsson, J., Berglund, S., Annett, J.: Narcissism – brain and behavior: self-views and empathy in the narcissistic brain term, University of Skovde (2014)
8. Daniel, R., Pollmann, S.: A universal role of the ventral striatum in reward-based learning: Evidence from human studies. *Neurobiol. Learn. Memory.* **114**, 90–100 (2014)
9. Weger, M., Sandi, C.: High anxiety trait: a vulnerable phenotype for stress-induced depression. *Neurosci. Biobehav. Rev.* **87**, 27–37 (2018)
10. Sun, P., Wei, S., Wei, X., Wang, J., Zhang, Y., Qiao, M., Wu, J.: Anger emotional stress influences VEGF/VEGFR2 and its induced PI3 K/AKT/mTOR signaling pathway. *Neural Plast.* **2016**, 1–12 (2016)
11. Schmidt, M.V., Abraham, W.C., Maroun, M., Stork, O., Richter-Levin, G.: Stress-induced metaplasticity: from synapses to behavior. *Neuroscience* **250**, 112–120 (2013)
12. de Almeida, R.M.M., Cabral, J.C.C., Narvaez, R.: Behavioural, hormonal and neurobiological mechanisms of aggressive behaviour in human and nonhuman primates. *Physiol. Behav.* **143**, 121–135 (2015)
13. Jabeen, F.: How happy you are: a computational study of social impact on self-esteem. In: International Conference on Social Informatics, Social Informatics, vol. 11186, pp. 108–117. Springer, Cham (2018)
14. Treur, J.: Multilevel network reification: representing higher order adaptivity in a network. In: Complex Networks and Their Applications VII, pp. 635–651. Springer (2019)
15. Treur, J.: Modeling higher-order adaptivity of a network by multilevel network reification. *Netw. Sci. J.* (2019, to appear)
16. Treur, J.: Network-Oriented Modeling. Springer International Publishing, Cham (2016)
17. Hebb, D.: The Organization of Behavior. Wiley, Hoboken (1949)
18. https://github.com/MsFakhra/Narcissism_TemporalSpecifications (2019)
19. Nai, A.: Disagreeable narcissists, extroverted psychopaths, and elections: a new dataset to measure the personality of candidates worldwide. *Eur. Polit. Sci.* **18**, 309–334 (2019)
20. Folley, A.: Trump rips MSNBC, CNN for inviting guests who “have no idea what I am all about” (2019)
21. Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: AAAI Conference on Weblogs and Social Media, ICWSM (2014)



Text Sentiment in the Age of Enlightenment

Philipp Koncar^(✉) and Denis Helic

Graz University of Technology, Graz, Austria
{philipp.koncar,dhelic}@tugraz.at

Abstract. Spectator journals published during the Age of Enlightenment served to enhance morality of readers and focused on a plethora of topics, such as the image of women or men, politics and religion. Although spectator journals have been studied extensively, little is known about the sentiment that they express. In this paper, we analyze text sentiment of spectator journals published in four different languages during a time period of over one hundred years. For that, we conduct (i) a sentiment analysis and (ii) analyze sentiment networks, for which we compute and investigate various network metrics, such as degree distributions and clustering coefficients. Additionally, we study the commonalities and differences between negative and positive words according to the respective metrics. Our results depict a high variability in positive and negative word usage and their linking patterns and extend our knowledge of spectator journals published during the Age of Enlightenment.

Keywords: Sentiment networks · Spectator journals · Enlightenment

1 Introduction

During the Age of Enlightenment (starting in the 18th century), so-called *spectator journals* were a popular way of distributing information and providing a platform for debating a plethora of topics, such as politics, religion and literature. Originating from London, the idea of these journals quickly spread all across Europe either through translations or independent issues covering local matters in respective countries. As some journals questioned customs and traditions or even included public criticism, these journals were emotionally charged and contained opinions with polarizing sentiment.

While text sentiment on the Web and in social media has been studied extensively in recent years [21], we still lack a broader understanding of sentiment in texts originating from earlier, non-digital times and conventional media. Hence, we analyze a large dataset comprising spectator journals written in major European languages to shed more light on the text sentiment expressed in media dating back to the Age of Enlightenment.

Approach. In our work, we analyze a manually annotated dataset of more than 3 400 issues of spectator journals published in German, French, Italian and

Spanish between 1711 and 1822. Each issue deals with one or more topics (e.g., politics or marriage) and comprises multiple text passages, each following a specific stylistic scheme (e.g., letters to the editors, quotes or dream stories). To study text sentiment, we first utilize sentiment dictionaries to compute the overall sentiment of journals. Second, we construct and analyze sentiment networks, in which two words from a respective sentiment dictionary for a given language (i.e., words with either a positive or negative sentiment) are connected if they co-occur in the same text passage. Specifically, we investigate (i) degrees, (ii) clustering coefficients, (iii) assortativity, and (iv) centralities to assess commonalities and differences between words with positive and negative sentiment.

Findings and Contributions. We find that positive and negative words have distinguishable properties in sentiment networks. For example, we observe a higher local clustering coefficient for negative words as compared to positive ones, indicating higher transitivity of negative words. We report a tendency towards high and low degrees for negative words with a few negative words connecting to a high number of other words. On the contrary, positive words typically have mid-range degrees connecting frequently to a moderate number of other words. Centrality metrics reveal a majority of negative words among the top central words, hinting that prominent words frequently express a negative sentiment. Finally, our study shows non-assortative mixing between positive and negative words, signaling the absence of sentiment consistency in spectator journals.

To the best of our knowledge, our work is the first to investigate text sentiment in spectator journals published during the Age of Enlightenment. Further, we publish the code of our analysis¹, opening up possibilities to learn more about the characteristics of texts originating from the 18th and 19th century and to compare them to today's texts and media regarding, for example, the attitude towards important societal topics then and now.

2 Related Work

Spectator Journals. The foundation for spectator journals was laid by *The Spectator* (“spectator” implied to stand above party) which was published in England between 1711 and 1714 and combined until then separate fields of journalism, such as political, social and scholarly information, into one journal. Issues of *The Spectator* enjoyed great popularity (circulation of around 1,600 exemplars [17]) and were translated and imitated quickly and numerously all across Europe and, thus, created a new genre of periodicals and journalism [19, 26]. First (partial) translations were conducted by the French and Germans in 1714 [13, 24], followed by the Netherlands, Denmark and Sweden, and then by Italy, Spain and Portugal [14, 18, 26]. Journals significantly contributed to social development, such as the image of women [8, 25] or religious beliefs [1].

In our work, we focus on a manually annotated dataset including German, French, Italian and Spanish spectator journals published between 1711 and 1822.

¹ Code available at <https://github.com/philkon/sentiment-spectator>.

Sentiment Analysis. A common task in Natural Language Processing (NLP) is sentiment analysis, aiming to investigate emotions, attitudes and opinions expressed in textual data [28]. Basic methods rely on dictionaries comprising lists of words for which the sentiment or polarity (either positive or negative) is known. Popular dictionary-based methods include SentiStrength [34] or VADER [15], both specifically introduced for short texts originating from social media platforms, such as Twitter or Facebook. Another widely used dictionary-based method is LIWC [29], aiming to identify characteristics of authors by analyzing their texts. Further, machine learning approaches for sentiment classification have been studied [23, 27, 38], including neural networks [39]. As most of studies in the field of sentiment analysis (or, in general, all NLP areas) focus on the English language, we encounter a significant scarcity of dictionaries for non-English languages. Addressing this issue opened a whole new research area known as *cross-lingual sentiment classification* [9], trying to transfer existing English models to other languages. Commonly, this transfer happens by using machine translation techniques to project English models to the target language or vice versa [10, 30, 37].

In our work, we rely on a basic dictionary approach introduced by Chen and Skiena [10] as it allows for easy interpretation of results and, as opposed to other works, is available for the four languages contained in our dataset (cf. Sect. 3).

Networks to Represent Texts. Network representations of texts have been studied extensively [2, 3, 5, 11, 20, 33]. Depending on the application, there are multiple ways of how to model texts as networks. In cases where semantics are important, models connect words with semantic relations or words that co-occur in the same context, for example, a sentence or paragraph [2, 35, 36]. If structure or style is important, words are connected based on syntactical relations, with word adjacency networks [4, 32] being a well-known approach. Basically, this model links adjacent words in texts with each other, captures stylistic characteristics of texts and is language independent. Network representations of texts have also been used for topic modeling. For example, Gerlach et al. [12] introduced an approach based on bipartite networks of documents and words and used existing community detection methods to identify topics. Zuo et al. [40] introduced WTNM, a topic model based on co-occurrence networks and specifically designed for sparse and short texts. Further, Liu et al. [22] built the Topical PageRank for co-occurrence networks to extract keyphrases that summarize documents.

In this paper, we combine the network representations of spectator journals with sentiments. While most of the mentioned works focused on English, we investigate metrics of networks based on German, French, Italian and Spanish.

3 Text Sentiment in Spectator Journals

Dataset. We work with a collection of journals that has been manually annotated by experts working in the fields of humanities and has already been published in a digital edition. Annotated journals follow the TEI standard² and are

² <https://tei-c.org/>.

accessible to the public³. Overall, our dataset contains 3 446 issues of 59 different spectator journals written in four different languages, comprising German (35 issues), French (1 479 issues), Italian (1 308 issues) and Spanish (624 issues), each published during a distinct time period. Besides the issue text (with annotated levels of representation and narrative forms), the dataset includes the following additional information for each issue: author, date of publication, country of publication, one or multiple topics (out of a list comprising 38 distinct topics) and mentioned persons, places or works. In Table 1 we list a detailed comparison between languages and summarize statistics of the dataset.

Preprocessing. We start our analysis by parsing and processing the TEI encoded XML files. For each issue included in our dataset, we extract and aggregate text according to levels of representation and narrative forms into *text passages* to combine texts with relative subjects. We further extract authorship information, dates of publication and manually assigned topics. Next, we normalize imprecise dates of publication (e.g., 1711 – 1712 ⇒ 1711) and duplicate author names (e.g., *Eliza Haywood* and *Eliza Fowler Haywood* ⇒ *Eliza Fowler Haywood*). Lastly, we exclude 14 issues with missing data from our analysis.

Sentiment Dictionaries. We determine words expressing a positive or negative sentiment by using dictionaries that have already been evaluated for German, French, Italian and Spanish in a previous work [10]. Here, authors extracted most frequent words of Wikipedia articles and created a knowledge graph to combine similar words of different languages through using Wiktionary, automatic machine translation (via Google translate), transliteration links and WordNet. Starting with sentiments of English vertices based on a dictionary with 1 422 (32%) positive and 2 956 (68%) negative words, authors propagated sentiments to vertices of other languages and created dictionaries for 136 languages, each including a list of words with both a negative and a positive sentiment.

Computing Sentiment. Using the respective German (3 974 words, 38% positive), French (4 653 words, 35% positive), Italian (4 491 words, 36% positive)

Table 1. Basic Dataset Statistics. This table lists the number of issues, text passages, unique authors, anonymous publications, number of topics, and the time spans for which our dataset contains publications, respectively for each language.

	German	French	Italian	Spanish
# Issues	35	1 479	1 308	624
# Text passages	150	7 556	6 965	3 914
# Unique authors	3	13	11	20
# Anon. publications	27	311	2	108
# Topics	26	37	34	32
Time span (years)	1723–1765	1711–1795	1727–1822	1761–1804

³ <https://gams.uni-graz.at/mws>.

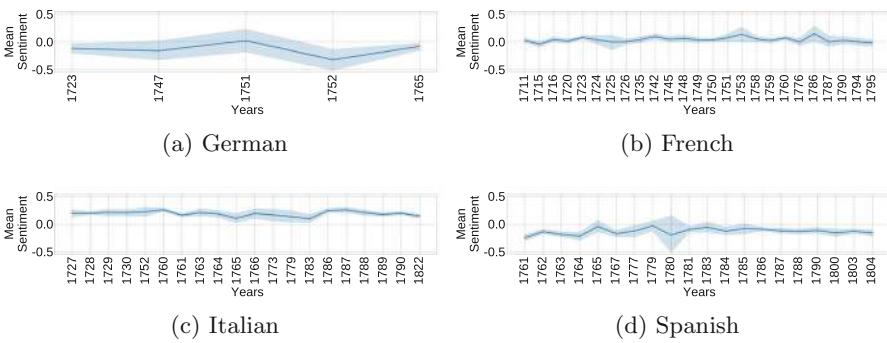


Fig. 1. Mean Sentiment Over Years. This figure illustrates the mean sentiment of text passages (with 95% confidence level) over the years in which issues were published in respective languages.

and Spanish (4 275 words, 36% positive) dictionaries, we compute the sentiment score s of each text passage in our dataset with $s = (W_p - W_n)/(W_p + W_n)$, where W_p is the number of positive words in a text passage and W_n is the number of negative words in a text passage. Hence, the sentiment score is a value ranging between -1 and $+1$, where values close to -1 are considered as negative, values close to $+1$ as positive, whereas values close to zero indicate a neutral sentiment.

Sentiment Results. We report mean sentiment of text passages over the years in which issues were published in respective languages in Fig. 1. Overall, sentiment varies for all of the four languages over time. For German and Spanish, the mean sentiment is mostly negative, indicating that German and Spanish journals express more negative emotions. In contrast, for French and Italian journals, mean sentiment is positive throughout the years.

In Fig. 2 we depict mean sentiment over all languages, respectively for each topic contained in our dataset. Again, sentiment varies significantly over topics. Topics regarding *Spain* are among the most negative. As 86% of issues dealing with *Spain* and the *Apologetic of Spain* are written in Spanish, for which sentiment is mostly negative throughout the years (cf. Fig. 1d), we conclude that local Spanish topics were negatively perceived in Spain during the Age of Enlightenment. Another topic with a significant negative sentiment is *Foreign Societies* (57% French, 40% Spanish, 2% Italian and 1% German issues), possibly reflecting closed societies in Europe at the observed time. On contrary topics, *Italy* (76% of issues are written in Italian) and, in general, all Italian journals are rather positive throughout the investigated time span. This indicates that Italian authors had a positive stance towards topics during the Age of Enlightenment.

Further, topics such as *Nature* (50% Italian, 45% French and 5% Spanish issues), *America* (24% Italian, 68% French and 8% German issues) and *Charity* (62% Italian, 26% French and 12% Spanish issues) are positively perceived all across Europe at the time. Together with further substantially positive topics, such as *Reason*, *Idea of Man*, or *Happiness*, this finding signals a positive

perception of intellectual and humanistic topics during the Age of Enlightenment in Europe. Interestingly, *Friendship*, the most positive topic on average, was only covered by French journals, signaling the importance of interpersonal relationships and communication for the French. This observation is further corroborated by the fact that the topics *Love* and *Family* are most represented in French journals (68% and 67% ratio of French issues, respectively).

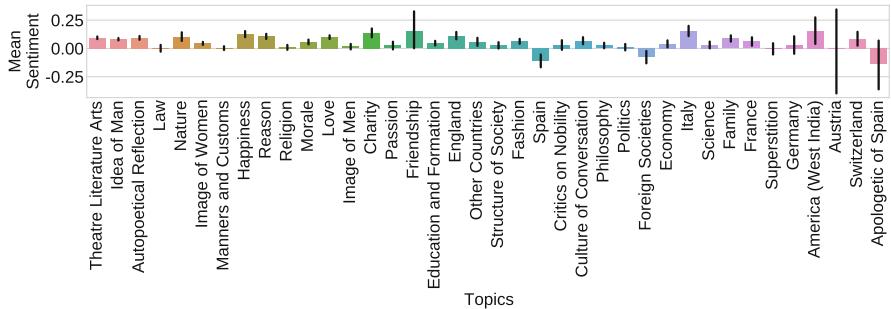


Fig. 2. Mean Sentiment Over Topics. This figure illustrates the mean sentiment (with 95% confidence level) over all languages for each topic contained in our dataset.

4 Sentiment Networks

Before creating sentiment networks, we use Spacy⁴ and its respective language models to lemmatize texts of journals in order to combine inflected word forms into single representations. Additionally, we filter 38 stop words⁵ across all languages (potentially introduced by machine translation of English dictionaries) and do not include them in our networks. We then construct sentiment networks by linking positive and negative words from our sentiment dictionaries if they co-occur at least once in the same text passage of an issue. As words can co-occur multiple times over text passages, for each language we create two versions of the networks: (i) unweighted networks in which we ignore edge multiplicity and (ii) weighted networks in which we use edge multiplicity as link weights.

4.1 Network Metrics

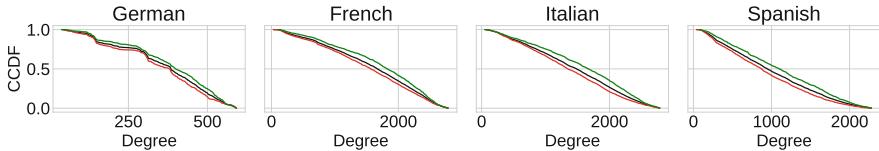
For all weighted and unweighted networks, we analyze: (i) degree distribution, (ii) local clustering coefficient distribution, (iii) degree and sentiment assortativity, and (iv) degree, betweenness, and closeness centrality. In our analysis, we distinguish between distributions of positive and negative words for all metrics except assortativity, for which we consider the networks as a whole.

⁴ <https://spacy.io> (version used: 2.1.3).

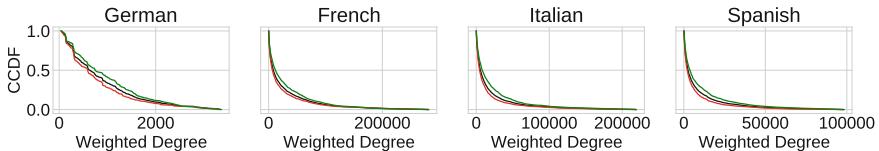
⁵ <https://github.com/Alir3z4/python-stop-words>.

Table 2. Network Sizes. This table lists the number of nodes and edges of networks for each language. Numbers in brackets are ratios of positive sentiment nodes.

German		French		Italian		Spanish	
# Nodes	# Edges	# Nodes	# Edges	# Nodes	# Edges	# Nodes	# Edges
593 (45%)	107028	2 786 (38%)	2 204 235	2 787 (39%)	2 019 043	2 280 (41%)	1 142 144



(a) Weighted Networks



(b) Unweighted Networks

Fig. 3. Degree Distributions. This figure depicts the CCDF of degrees of all nodes (black lines), positive nodes (green lines) and negative nodes (red lines) in unweighted (a) and weighted (b) networks, respectively for each language.

Basic Networks Statistics. In Table 2 we list the number of nodes and edges of sentiment networks including the percentage of positive and negative nodes in each network. In all networks positive words represent the minority with ratios ranging from 38% for the French network to 45% for the German network. Regarding network connectivity, we find that networks for all languages are well-connected, having only one connected component.

Degree. We depict complementary cumulative distribution functions (CCDFs) for unweighted degrees of all nodes, positive nodes only and negative nodes only, respectively for each language in Fig. 3a. For all languages, we observe that probability for mid-range degrees is higher for positive nodes while negative nodes have more probability for low and high degrees. The difference of distribution between positive and negative nodes is significant for all languages according to the two sample Kolmogorov-Smirnov test (p -values < 0.05). We observe similar differences in weighted degree distributions (cf. Fig. 3b). However, weighted degree distributions are more heterogeneous, indicating that the majority of words have lower numbers of co-occurrences, whereas only a few words co-occur very often in our dataset. Again, the two sample Kolmogorov-Smirnov tests reveals a significant difference between positive and negative distributions (p -values < 0.005). With respect to average degrees, positive words have a higher mean degree (1390.56 compared to 1233.86 over all languages) and a higher

mean weighted degree (19814.45 compared to 14476.8 over all languages). This suggests that a few negative words connect to many other words and that a substantial number of negative words connect only to a few other words, shifting the mean degree of negative words towards lower degrees and making it smaller as compared to mean degrees of positive words.

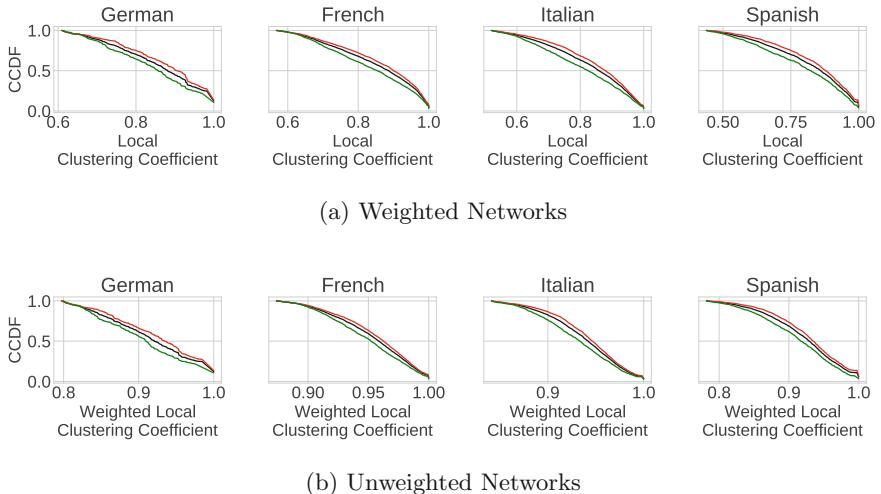


Fig. 4. Local Clustering Coefficient Distributions. This figure depicts the CCDF of local clustering coefficients of all nodes (black lines), positive nodes (green lines) and negative nodes (red lines) in unweighted (a) and weighted (b) networks, respectively for each language.

Local Clustering Coefficients. We illustrate CCDFs for local clustering coefficients of unweighted sentiment networks in Fig. 4a and for weighted [6] sentiment networks in Fig. 4b. The difference between positive nodes and negative nodes distributions is significant (p -values < 0.0005) for unweighted and weighted networks of all languages. As expected, negative nodes have a higher probability to have a higher local clustering coefficient because, on average, they have lower degrees (following previous observations about a negative correlation between the degree and the local clustering coefficient [31]).

Assortativity. To further investigate the patterns of linking between positive and negative words we compute assortativity of sentiment networks for each language based on (i) degree and (ii) sentiment (i.e., positive or negative). Similarly to recent findings on assortativity in word networks based on part-of-speech tags [7], node degrees are slightly disassortative for all languages (ranging between -0.355 and -0.233), indicating that high degree nodes have a tendency to connect to low degree ones.

Assortativity based on sentiment is close to zero for all languages (ranging between -0.006 and -0.002), signaling that positive and negative words co-occur almost randomly in spectator journals. These observations could be due to the original intention of spectator journals which was to report impartially and neutrally, meaning that authors critically discussed topics in journals by always considering positive and negative characteristics. Another hypothesis of non-assortative sentiment networks is that spectator journals were inconsistent, potentially leading to ambivalence. We leave this analysis to future work.

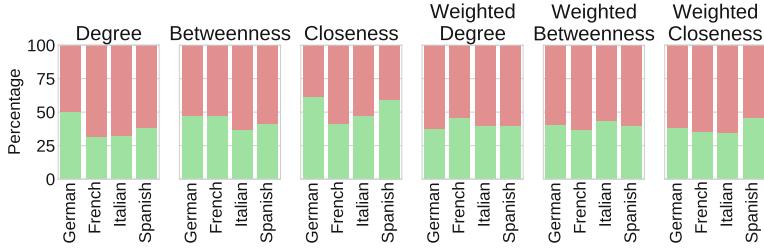


Fig. 5. Ratio of top central positive and negative words. This figure illustrates the ratio of positive (green color) and negative (red color) words among the top 100 central words, respectively for each centrality metric and language.

Table 3. Centrality Correlations. This table lists the Spearman rank-order correlation coefficients between each pair of centralities, respectively for each language. Gray cells indicate non-significance as p -values of the test are above 0.05.

	German	French	Italian	Spanish
Degree & Betweenness	0.38	-0.22	-0.16	0.01
Degree & Closeness	0.78	0.04	-0.15	-0.37
Betweenness & Closeness	0.54	0.26	0.24	0.32
W. Degree & W. Betweenness	0.29	-0.03	0.05	0.12
W. Degree & W. Closeness	0.29	0.02	0.1	0.08
W. Betweenness & W. Closeness	0.28	-0.4	-0.3	-0.41

Centralities. We list ratios between positive and negative words among the top 100 central words, respectively for each centrality metric and each language in Fig. 5. Overall, the majority of top central words is negative according to all three centrality measures and all languages. Exceptions are unweighted degree centrality for German and unweighted closeness for German and Spanish. As the positive nodes constitute the minority in our sentiment networks (cf. Table 2),

these results may be simply an artefact of the distribution of nodes in networks. To further investigate this observation we compute proportion tests based on one-sample z-tests. Our test results are inconclusive as we find that in four cases, ratios of negative words in top words are significantly smaller than ratios in networks and in all other cases we do not find statistical significance (with seven cases of smaller ratios and 13 of greater ratios of negative words). These test results even suggest a weak over-representation of the minority nodes, corroborating similar results for heterophilic networks found previously [16].

Among the negative top words we find, for example, “verboten” (German for “prohibited”), “scandaleux” (French for “scandalous”), “doloso” (Italian for “intentional”) or “difunto” (Spanish for “dead”)⁶. An exact analysis of these top words could be used to infer how topics were perceived in earlier times, opening a promising opportunity for future work.

Finally, in Table 3 we list the Spearman rank-order correlation coefficients for each pair of centrality metrics, depicting discrepancies among centralities, except for German, where we have positive correlations between metrics (all $r_s \geq 0.28$). Another exception are correlation of betweenness and closeness for unweighted French, Italian and Spanish networks.

5 Conclusion

Summary. In this paper, we investigated text sentiment expressed in spectator journals published during the Age of Enlightenment. We constructed sentiment networks based on positive and negative sentiment words found in journals written in four different languages. Computing basic metrics of sentiment networks, we investigated differences between positive and negative word usage. Particularly, we observed higher transitivity and a tendency towards low and high degrees for negative words as compared to positive ones. The majority of top central co-occurring words is negative for all languages, indicating an excessive usage of a small number of negative sentiment words throughout the journals.

Limitations. Due to the absence of dedicated and comparable approaches for 18th century texts written in the four languages, we fall back on existing methods (lemmatization and sentiment dictionaries) intended for modern day texts. Training adjusted models could further improve the quality of our analysis but is out of scope for this work. We still argue that our results are reasonable as the ratio of text passages for which we could determine a sentiment with the respective dictionaries is 76% (out of 18 585 text passages) over all languages. Further, we present results based on networks with edges requiring only one co-occurrence of two words in text passages. The construction of sentiment networks can be altered by introducing a threshold of minimum co-occurrences in order to connect two words. We tried additional thresholds of 5 and 10 co-occurrences for which results varied only minimally and leave an extensive analysis of the impact of sentiment network construction to future work.

⁶ See our GitHub for complete lists of top hundred central words per language.

Future Work and Impact. Future work includes the development of NLP techniques dedicated to the language of the 18th century, as well as the consideration of additional languages, such as Danish or Dutch spectator journals. Our work, including published code for reproducibility, lays the foundation for extended studies of spectator journals published during the Age of Enlightenment and for comparative analysis with modern day texts.

Acknowledgements. Parts of this work were funded by the goldigital programme of the Austrian Academy of Sciences. Further, we want to thank Alexandra Fuchs, Bernhard Geiger, Elisabeth Hobisch, Martina Scholger and further members of the DiSpecs project for their fruitful input during the conduction of our experiments.

References

1. Allan, D., Virtue, L.: *The Scottish Enlightenment*, Edinburgh (1993)
2. Amancio, D.R., Oliveira Jr., O.N., Costa, L.d.F.: Unveiling the relationship between complex networks metrics and word senses. *EPL* **98**(1), 18002 (2012)
3. Amancio, D.R.: A complex network approach to stylometry. *PLoS ONE* **10**(8), e0136076 (2015)
4. Amancio, D.R., Nunes, M.G.V., Oliveira Jr., O., Pardo, T.A.S., Antiqueira, L., Costa, L.d.F.: Using metrics from complex networks to evaluate machine translation. *Phys. A Stat. Mech. Appl.* **390**(1), 131–142 (2011)
5. Antiqueira, L., Oliveira Jr., O.N., da Fontoura Costa, L., Nunes, M.d.G.V.: A complex network approach to text summarization. *Inf. Sci.* **179**(5), 584–599 (2009)
6. Barrat, A., Barthelemy, M., Pastor-Satorras, R., Vespignani, A.: The architecture of complex weighted networks. *Proc. Nat. Acad. Sci.* **101**(11), 3747–3752 (2004)
7. Cantwell, G.T., Newman, M.: Mixing patterns and individual differences in networks. *Phys. Rev. E* **99**(4), 042306 (2019)
8. Carr, R.: *Gender and Enlightenment Culture in Eighteenth-Century Scotland*. Edinburgh University Press, Edinburgh (2014)
9. Chen, Q., Li, C., Li, W.: Modeling language discrepancy for cross-lingual sentiment analysis. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 117–126. ACM (2017)
10. Chen, Y., Skiena, S.: Building sentiment lexicons for all major languages. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 383–389 (2014)
11. Cong, J., Liu, H.: Approaching human language with complex networks. *Phys. Life Rev.* **11**(4), 598–618 (2014)
12. Gerlach, M., Peixoto, T.P., Altmann, E.G.: A network approach to topic models. *Sci. Adv.* **4**(7), eaauq1360 (2018)
13. Gilot, M., Sgard, J.: Le journalisme masqué. Le journalisme d'ancien régime (Lyon), pp. 285–313 (1981)
14. Gustafson, W.W.: The influence of the “Tatler” and “Spectator” in Sweden. *Scand. Stud. Notes* **12**(4), 65–72 (1932)
15. Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International AAAI Conference on Weblogs and Social Media (2014)
16. Karimi, F., Génois, M., Wagner, C., Singer, P., Strohmaier, M.: Homophily influences ranking of minorities in social networks. *Sci. Rep.* **8**(1), 11077 (2018)

17. King, R.S.: All the news that's fit to write: the eighteenth-century manuscript newsletter. In: Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, pp. 95–118. Brill (2018)
18. Krefting, E.: News versus opinion: the state, the press, and the northern enlightenment. In: Travelling Chronicles: News and Newspapers from the Early Modern Period to the Eighteenth Century, pp. 299–318. Brill (2018)
19. Krefting, E., Nøding, A., Ringvej, M.: Eighteenth-Century Periodicals as Agents of Change: Perspectives on Northern Enlightenment. Brill, Boston (2015)
20. Kulig, A., Drożdż, S., Kwapienie, J., Oświęcimka, P.: Modeling the average shortest-path length in growth of word-adjacency networks. Phys. Rev. E **91**(3), 032810 (2015)
21. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining Text Data, pp. 415–463. Springer (2012)
22. Liu, Z., Huang, W., Zheng, Y., Sun, M.: Automatic keyphrase extraction via topic decomposition. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 366–376 (2010)
23. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 142–150. Association for Computational Linguistics (2011)
24. Martens, W.: Die Botschaft der Tugend: die Aufklärung im Spiegel der deutschen-moralischen Wochenschriften. Springer, Heidelberg (2017)
25. Messbarger, R.: Reforming the female class: “il caffè”s” defense of women”. Eighteenth-Century Stud. **32**(3), 355–369 (1999)
26. Pallares-Burke, M.L.: The spectator, or the metamorphoses of the periodical: a study in cultural translation. In: Cultural Translation in Early Modern Europe, pp. 142–159. Cambridge (2007)
27. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86 (2002)
28. Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Found. Trends® Inf. Retr. **2**(1–2), 1–135 (2008)
29. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001, vol. 71. Lawrence Erlbaum Associates (2001)
30. Prettenhofer, P., Stein, B.: Cross-language text classification using structural correspondence learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1118–1127 (2010)
31. Ravasz, E., Barabási, A.L.: Hierarchical organization in complex networks. Phys. Rev. E **67**(2), 026112 (2003)
32. Roxas, R.M., Tapang, G.: Prose and poetry classification and boundary detection using word adjacency network analysis. Int. J. Mod. Phys. C **21**(04), 503–512 (2010)
33. Silva, T.C., Amancio, D.R.: Word sense disambiguation via high order of learning in complex networks. EPL (Eur. Lett.) **98**(5), 58001 (2012)
34. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. J. Am. Soc. Inform. Sci. Technol. **61**(12), 2544–2558 (2010)
35. Véronis, J.: HyperLex: lexical cartography for information retrieval. Comput. Speech Lang. **18**(3), 223–252 (2004)

36. Widdows, D., Dorow, B.: A graph model for unsupervised lexical acquisition. In: Proceedings of the 19th International Conference on Computational Linguistics, vol. 1, pp. 1–7. Association for Computational Linguistics (2002)
37. Xiao, M., Guo, Y.: Semi-supervised representation learning for cross-lingual text classification. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1465–1475 (2013)
38. Ye, Q., Zhang, Z., Law, R.: Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Syst. Appl.* **36**(3), 6527–6535 (2009)
39. Zhang, L., Wang, S., Liu, B.: Deep learning for sentiment analysis: a survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **8**, e1253 (2018)
40. Zuo, Y., Zhao, J., Xu, K.: Word network topic model: a simple but general solution for short and imbalanced texts. *Knowl. Inf. Syst.* **48**(2), 379–398 (2016)



A Gravity-Based Approach to Connect Food Retailers with Consumers for Traceback Models of Food-Borne Diseases

Tim Schlaich¹✉, Hanno Friedrich¹, and Abigail Horn²

¹ Kuehne Logistics University, 20457 Hamburg, Germany
tim.schlaich@the-klu.org

² Keck School of Medicine, University of Southern California,
Los Angeles, CA 90033, USA

Abstract. Computational traceback models are important tools for investigations of widespread food-borne disease outbreaks as they help to determine the causative outbreak location and food item. In an attempt to understand the entire food supply chain from farm to fork, however, these models have paid little attention to consumer behavior and mobility, instead making the simplifying assumption that consumers shop in their home location. This paper aims to fill this gap by modelling food-flows from supermarkets to consumers in a large-scale gravity model for Hesse, Germany. Modelling results show that on average, groceries are sourced from two to four postal zones with half of all goods originating from non-home postal zones. The results contribute to a better understanding of the last link in the food supply chain. In practice, this allows investigators to relate reported outbreak cases with sourcing zones and respective food-retailers. The inclusion of this information into existing models is expected to improve their performance.

Keywords: Food supply network · Supply-chain network · Gravity model · Grocery shopping · Food-borne diseases

1 Introduction

Incidents like the E. coli outbreak in Germany 2011 have shown that food-borne diseases can have a considerable economic and public health impact and therefore remain a challenging topic for food safety authorities. While most food-related illnesses occur locally and can be directly attributed to a food item or location, more severe and widespread outbreaks often require a multi-disciplinary team to localize the epicenter and remove the contaminated food item [1]. Since conventional investigation processes require a high degree of manual work and deal with complex food supply systems, authorities often identified the contaminated food item only weeks after the outbreak or not at all [2]. Emerging technologies and more readily available supply chain data enable authorities to complement investigation processes with computational models [3]. In the case of widespread outbreaks these data-scientific approaches promise particular utility in outbreak investigations, given the complex task of tracing through the massive food supply network. Data scientific approaches support the investigation process in

three parts: (i) detecting that an outbreak is occurring, (ii) identifying the location source of an outbreak at an early stage of the supply chain like a farm or food processor and (iii) identifying the contaminated food item that caused an outbreak [4]. The findings of this paper are relevant for both (ii) and (iii) of the investigation process.

Network-theoretic location source models (ii) aim to find the location source of an outbreak in the supply network given the underlying food distribution network and reported outbreak locations. In this type of model, reported outbreak cases are connected to the food supply network by linking to a node representing the point of sale of the contaminated food product, assuming the purchase happens in the district of residence of the infected person. Given the reported cases and the food network, these models aim at identifying the likeliest source node that could have caused the observed outbreak pattern [4, 5].

Another type of model seeks to determine the contaminated food item that caused an outbreak (iii) by relating reported outbreak locations with retail sales data from supermarkets [6, 7]. In principal, these models assume that outbreaks are caused by food items that are sold in the postal zones where infected people live. Hence, food items with a high relative share of sales in these areas compared to other areas are more likely to be the contaminated food item that caused the outbreak.

While both model types show promising results in identifying the contaminated food item and location, they do not appropriately consider the last link in the supply chain – the consumer. Both model types' assumption that consumers only shop in their home area is oversimplifying. Several studies indicate that consumers travel across postal zones for their grocery shopping [8–10]. Moreover, estimating travel behavior is complex and may depend on a set of socio-economic factors like income, availability of cars, marital or gender diversity [11, 12]. Hence, there is a need to gain a better understanding of consumer shopping behavior to explain the last mile flow of food products. It is expected that integrating the last mile flow from retailer to consumers into existing source detection models will improve speed and accuracy [4, 6, 7].

One way to reproduce the last mile shopping behavior and tackle this gap is to apply trip distribution models. Such models use quantitative and/or qualitative factors to estimate shopping behavior [13]. Among them, gravity-based models belong to the most commonly used trip distribution models [14]. Suhara et al. [12] have shown that gravity models yield particularly good results for grocery shopping estimations. Gravity models estimate the flow of goods or persons between two zones based on two factors: distance and attractiveness. By simulating the last mile flow of food products from retailers to consumers, this paper aims to answer the following research question: *How does a consumer's place of residence differ from the place of food purchasing?* The findings to this question are mainly investigated in the context of food-borne diseases.

To answer this question, we build a gravity model for the state of Hesse to simulate the flow of food products on a postal zone level ($n = 547$). We generate a large-scale flow matrix with high resolution zoning data that matches the short-distance activity grocery shopping while at the same time covering a wider geographical area as required for large-scale outbreaks. The gravity model is based on mobility survey data provided by the Federal Ministry of Transport and Digital Infrastructure, supermarket locations and revenues as well as consumption data from open source and commercial data

sources. The model parameters are calibrated with the Furness and Hyman algorithms to ensure a realistic flow distribution [15, 16]. We place particular focus and introduce methodological innovations in our approach to estimate intra-zonal distances. The resulting model estimates monetary grocery flows between postal zones and can be used as a proxy for the strength of consumer-retailer interactions between zones.

The remainder of this paper is structured as follows: Sect. 2 introduces the chosen gravity model form and its formulation. In Sect. 3, we explain the required data inputs and outline the calibration procedure. In Sect. 4, modeling results are presented and implications for food-borne diseases are discussed. Lastly, we refer to limitations and provide an outlook for future research.

2 Gravity Model

Our objective is to develop a model of the last mile flow of groceries between retailers and consumers that can be used to supplement traceback models of foodborne disease outbreaks. Multiple gravity model forms exist for simulating flows between retailers and consumers, and the choice of model depends on the purpose for its use, as well as the data available for model fitting. An important factor in modeling choice is the level of aggregation. Shopping interactions between consumers and retailers can be represented in a disaggregate model that estimates the behavior of individuals, or an aggregate model if retail outlets within a zone are jointly evaluated. In aggregated models, individual store characteristics and exact distances between consumer and supermarket are lost [17]. However, the aggregation per zone reduces the complexity considerably as the set of destinations decreases and may even be beneficial if the model still fulfils the anticipated purpose. Because the purpose of our gravity model is to link to traceback models that are aggregated to a zonal level, we choose an interzonal gravity model form.

To build a large-scale gravity model where consumer-retailer interaction is estimated by revenue flows between zones, we rely on Wilsons's (1970) entropy maximizing gravity model. In this model, revenue flows F between two given postal zones i and j are defined as:

$$F_{ij} = A_i O_i B_j D_j e^{-\beta c_{ij}} \quad (1)$$

where O_i denotes the total retailer revenue generated by a zone i and D_j consumption potential of a zone j . A_i and B_j are balancing factors to ensure that the modeled revenue distribution matches the given zonal revenue generated by retailers and zonal revenue attracted by consumers per zone. They are defined as:

$$A_i = \frac{1}{\sum_j B_j D_j e^{-\beta c_{ij}}} \quad (2)$$

$$B_j = \frac{1}{\sum_i A_i O_i e^{-\beta c_{ij}}} \quad (3)$$

Further, to ensure modeling consistency $\sum_i F_{ij} = O_i$ and $\sum_j F_{ij} = D_j$. The frictional impact of distance is incorporated by an exponential deterrence function with deterrence factor β and distance c between two zones i and j .

3 Experimental Setting

This section describes the inputs necessary and procedure used to fit the gravity model. First, the level of aggregation and corresponding geo-spatial units - so called *traffic analysis zones* (TAZ) - must be chosen as the origin and destination. Corresponding *distances* between zones representing trip lengths are calculated. The flow intensity between two zones is then estimated as a function of the *revenue* and *consumption potential*, and the spatial distance separating these zones. We analyze mobility survey data to determine the average shopping distance of consumers. Lastly, we calibrate the model and generate revenue flows that match the observed mean shopping distance and the zonal revenue and consumption constraints.

3.1 Model Inputs

Due to the data available, most of the model inputs are not available as raw data and require modeling approaches. We therefore describe the raw data together with the modeling choices used to derive model input and calibration.

Area of Analysis. We generate the gravity model for Hesse, a medium-size state in middle Germany with both urban and rural areas and 547 postal zones.

Traffic Analysis Zones (TAZ). The delineation of zones has a direct impact on the modelling outcome and needs to match the modelling purpose [18]. Shopping trips, for instance, require a relatively fine zoning grid since the average trip length is relatively short. Otherwise, a high share of trips will originate and end in the same zone which typically indicates a too-coarse zoning system [19]. Nevertheless, TAZ delineation underlies certain restrictions and is usually not freely defined. The configuration of TAZs needs to match the data sources used for transport modelling [20]. It is suggested to choose the TAZ in line with the administrative zoning systems used in population, employment or marketing data [21]. Regardless of the zone size, there will always be a certain share of intra-zonal flows even for fine-grid zoning systems.

For our gravity model, we use German postal zones as TAZs [22, 23]. Postal zones are comparable to municipalities that represent the highest resolution unit LAU2 within the European NUTS (*Nomenclature des unités territoriales statistiques*) zoning system. Postal zones size varies from 0.0044 to 219.86 km² with a mean size of 39.14 km².

Inter-zonal Distance Estimation. We calculate inter-zonal distances between centroids – a common point that bundles all people and activities of a zone. For the sake of simplification, this center is often assumed to be equal to the geographical center [24]. The aggregation of all activities and people to a single point can lead to inaccurate estimations of separation [25]. First, the geographical center might not represent the actual center of population or activity. And second, the aggregation per se leads to an

error as all retailers/consumers in a zone are assumed to be located in the centroid [19]. The larger the zones are, the stronger the aggregation effect and the higher the potential estimation error. This error is limited by our zoning choice of the high-resolution postal zone level.

Intra-zonal Distance Estimation. In practice, transportation modelers mainly focus on centroid-to-centroid flows and tend to exclude intra-zonal trips [19, 25]. However, this leads to a biased sample and impedes proper model calibration [26, 27]. In our case this is especially true, since despite the high resolution on the postal zone level, many shopping trips are very short and a large proportion of all shopping trips are expected to be intra-zonal. We therefore develop an estimation approach to account for intra-zonal trip distances.

We follow two distinct approaches to model intra-zonal distances. First, we implement a geometrical approach, which is to calculate the average distance between two randomly distributed points within the zone. We use the formula derived by CZuber [28], which gives that within a circle shaped zone with radius r , the average distance d_{intra} between two randomly distributed chosen points, i.e. a consumer and a retailer, can be calculated by:

$$d_{intra} = \frac{128}{45 \times \pi} \times r \quad (4)$$

Since this mathematical formulation is a function only of the zone size and does not take store density or distribution into consideration, it is expected to overestimate the true intra-zonal distance especially for large and high-density zones. Therefore, we introduce a second estimation method, accounting for store density and distribution, to serve as a lower bound on intra-zonal shopping distances. In this approach, we assume retailers to be arranged in a lattice. This order maximizes the distance between retailers and minimizes the mean average distance to the nearest neighbor for randomly distributed consumers [29] (Fig. 1).

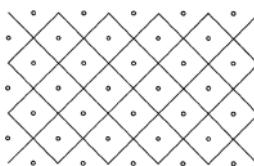


Fig. 1. Retailer in a lattice-arranged grid

In this setting, the mean distance $E[D]$ of a randomly distributed consumer to the nearest retailer within a postal zone r of area A_r with n_r retail stores and entity density rate $\lambda_r = \frac{A_r}{n_r}$ can be calculated as:

$$[D] = \frac{2}{3} \sqrt{\frac{1}{2\lambda}} \approx 0.427\lambda^{-0.5} \quad (5)$$

Retailer Revenue Estimation. We define a zone's potential to attract consumers as the sum of all store revenues of a given zone. Since individual store revenues are not available on a postal zone level, we generate a food retailing network in order to simulate these revenues as follows:

We incorporate 13 supermarkets and discounter including the major players Edeka, REWE, Aldi and Lidl. Store locations (degrees of latitude & longitude and addresses) are mainly sourced from a publicly available point of interest (POI) platform Pocket-navigation. The modeled store revenue was calculated based on the latest yearly revenue figures from Lebensmittel Zeitung (LZ) [30]. To calculate the revenue of an individual store, the yearly revenue of a retail chain reported by LZ was divided by the total number of stores found in the POI dataset. The chosen approach implies that all stores of a certain retailer generate equal revenues and ignores potential differences in size and/or purchasing power of customers. Edeka and Rewe stores were modeled separately with commercial retailer data as these two full-range retailers are predominantly operated by independent traders and vary considerably in revenue and size [31].

Consumption Potential Estimation. The consumption potential of a zone is expected to be proportional to its population size (Eq. 6), i.e.

$$c_r = \frac{pop_r}{\sum_r pop_r} \times REV \quad (6)$$

where c_r denotes the grocery consumption in a postal zone r with population pop_r and REV represents the total revenue of all food retailers over postal zone r . This consumption estimation implies that the mean food consumption is equal across different zones.

Observed Trip Data. We analyze mobility data from the *Federal Ministry of Transport and Digital Infrastructure* to find the mean shopping distance of consumers between their home and supermarkets for the calibration process.

The most recent mobility survey, *Mobilität in Deutschland 2017*, encompasses about 316,000 individuals from 156,000 households across Germany. Their mobility patterns are gathered in almost 1 million trips [32]. While the mobility data does not contain origin and destination information, trip purpose (e.g. work) and travel distances are assigned to each trip. Consequently, we only extract trip chains *home-shopping* and *shopping-home* to derive the distance between consumers' home and supermarkets. After data processing 78,754 shopping trips yield a mean distance of $\bar{x} = 4.65$ km (Fig. 2).

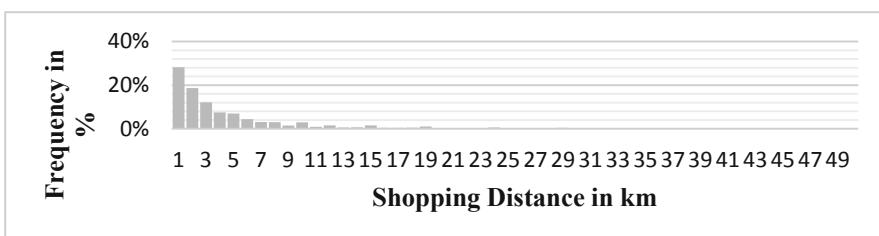


Fig. 2. Trip length distribution of shopping

3.2 Calibration Procedure

Gravity models need to be calibrated to ensure that the model successfully reproduces observed or estimated properties. In a doubly constrained gravity model, the production ($\sum_i T_{ij} = O_i$) and attraction constraints ($\sum_j T_{ij} = D_j$) ensure that the modelled sum of flows within a row or column of the OD-matrix matches the given production or attraction constraint of each zone. To reach a flow distribution that satisfies these constraints the balancing factors A_i for each row and B_j for each column need to be calculated. An additional parameter for the frictional impact of distance needs to be adjusted. An appropriate beta value is calibrated to ensure that the modelled average flow distance is equal to the target average flow distance. Consequently, in a matrix with n zones a total of $2n + 1$ parameters are required to calibrate a doubly-constrained gravity model [19].

We use a combined calibration method after Furness [16] and Hyman [15] to find adequate model parameters. The former method applies an iterative algorithm to resolve the interdependent balancing factors A_i and B_j while the latter method helps finding a deterrence factor β that matches the modeled flow distance with the target flow distance.

4 Results

4.1 Calibration Results

The gravity models were implemented in KNIME, an open-source analytics platform. For both gravity models – (a) geometrical approach (DC_G) and (b) nearest neighbor approach (DC_{NN}) – the KNIME-workflow was stopped after 100 Furness and 10 Hyman iterations. The target average distance of 4.65 km was fulfilled by the modelled distribution, while minor deviations regarding the given revenue and consumption constraints remained. In the DC_G , the modelled flow distribution on average deviates by 4.5% from the given consumption constraints while the maximum deviation was found to be 12.7% (compared to 2.9 and 7.7% in the DC_{NN}). The challenging calibration can be explained by the properties of the base matrix and is in line with the findings from literature where zero-cell values and a low level of entropy in the base-matrix are described as problematic [33, 34].

4.2 Food Flow Distribution

The modelled food flow distribution is expected to provide an answer to the following questions: (i) From how many postal zones are groceries sourced? (ii) What proportion of goods are expected to be bought intra-zonally by consumers? For the interpretation of the flow distribution we assume the consumer perspective (grocery inflow to a zone).

Given the matrix with flows FLOW_{ij} where groceries flow from a retailer zone i to a consumer zone j , the proportion of a zone's grocery supply p_{ij} can be calculated by:

$$p_{ij} = \frac{\text{FLOW}_{ij}}{\sum_{i=1}^n \text{FLOW}_{ij}} \in [0; 1] \quad (7)$$

To only consider meaningful consumer-retailer interaction, we define two percentage thresholds at 5 and 10%. Results show that on average groceries are sourced from two to four postal zones. Regardless of the chosen threshold, considerably more sourcing zones are estimated if intra-zonal flows are modelled with the nearest neighbor approach (Table 1).

From a practical standpoint, it is not only important which postal zones to look at, but also how likely a zone is to be visited by consumers. Quantifying the intensity of revenue flows therefore adds value for investigation purposes. If Eq. 7 is calculated for $ij = jj$, the proportion of intra-zonal consumption p_{jj} (i.e. share of groceries that were bought in the home zone) is obtained. In this regard, considerably higher proportions of intra-zonal consumption are estimated by the geometric approach. Since the mean intra-zonal distance for DC_{NN} is shorter, a lower proportion of intra-zonal flows is required to reach dist^* . This explains why the NN approach estimates intra-zonal revenue flows by 10% lower than the geometrical approach. Nevertheless, both the lower and the upper bound estimate that a major share of all groceries are expected to

Table 1. Mean number of sourcing zones and proportion of intra-zonal flows

Indicator	Intra-zonal approach			
	Geometrical approach		Nearest neighbor	
	5%	10%	5%	10%
Number of sourcing zones	2.45	1.96	3.50	2.20
Proportion of intra-zonal flows	55.7%		45.03%	

be bought from outside the home zone. The relatively consistent modelling results suggest that *independently from the chosen estimation approach for intra-zonal distances, a major part of all groceries is consumed from external zones*.

We explain further properties of our model with a food flow example for *Friedberg*, a small town with 30.000 inhabitants. For this specific zone, the DC_G estimates meaningful interaction (with 5% threshold) for two postal zones. The major part (56.4%) is expected to be bought from within while another larger proportion (41.7%) is estimated to origin from postal zone *Bad Nauheim*. This zone is located 6.6 km north of Friedberg and has a relatively high revenue compared to other neighboring zones (Fig. 3).

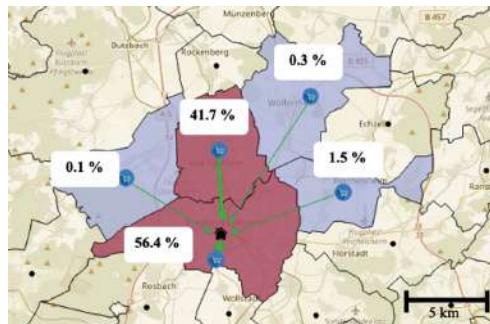


Fig. 3. Flow distribution of DC_G for Friedberg

Estimation of Retailer Market Share in Affected Zones. In addition to sourcing zones and flow intensities, the constructed gravity model provides an additional dimension of information for food-borne disease outbreak source identification. While the model operates on an aggregate level where supermarket revenues are added up to a zone's revenue, the information about individual retailers and their estimated revenue remains accessible. For the above given example of Friedberg, the following retailer and revenues can be identified (Table 2):

Table 2. Total revenue per retailer and sourcing zone (Friedberg)

Consumer Zone: 61169 Friedberg	Sourcing Zone 1: 61169 Friedberg	Sourcing Zone 2: 61231 Bad Nauheim
	9.59 Mio	26 Mio
	8.63 Mio	22.48 Mio
	7.38 Mio	19.19 Mio
	4.28 Mio	8.63 Mio
	2.67 Mio	3.69 Mio

The disaggregation of a zone's total revenue into revenues per retailer is value adding. Retailers that have no market share in the sourcing zones of postal zones where infected individuals live are unlikely to be the point of sales of a contaminated food product. In contrast, retailers that are active in all sourcing zones of reported cases and show high market shares deserve special attention from investigators if high selling volumes are associated with a higher likelihood to sell the contaminated item.

4.3 Interpretation of Results

From a FB investigation perspective, the findings about the last mile flow has a series of impacts. Firstly, they demonstrate that the last link of the food supply chain from farm to fork should not be neglected. Results from the analysis of the survey data show that supermarkets are on average 4.65 km away from consumers and that groceries are sourced from two to four postal zones. This implies that ignoring a consumer's mobility and simply assuming the home postal zone to be the zone of shopping leads to a sizeable estimation error for traceback models. The gravity model allows to bridge the gap between retailer and consumer. Given the place of residency of an infected individual, the gravity model points to a subset of postal zones and retailers where major food flows can be expected and quantifies them.

For the sales-frequency traceback models as discussed in Sect. 1, Norström et al. [6] account for a consumer's shopping mobility by assigning weights to neighboring zones based on shopping behavior gathered from questionnaires. As recognized by the authors, such information may not always be available, especially at an early investigation stage. If only limited information is given (e.g. only the place of residence) a modelled approach can be a valuable solution to account for a consumer's mobility. It would be possible to assign weights based directly on gravity-model-derived inflow proportions of a zone.

With respect to the network-theoretical approach, Horn and Friedrich [4] model food flows along the food supply network from the producer to the retailer stage. The network can be translated into a multi-path network structure, where flow volumes along network paths are depicted as probabilities. In this respect, the modelled food network appears to be compatible with the gravity model. It seems intuitive to expand the network by an additional layer and connect retailers with potential consumer locations identified by the gravity model.

5 Conclusion

The purpose of this paper was to investigate the flow of food products from retailers to consumers and assess its relevance for traceback models in the investigation process of food-borne diseases. A particular focus was put on the modelling of intra-zonal distances. Two gravity models with distinct estimation approaches for intra-zonal distance were constructed to simulate the flow of groceries in monetary terms.

Visited supermarkets were found to be located on average at a distance of 4.65 km according to data from consumers' home. Overall, it results that on average meaningful interaction between retailer und consumer is expected for two to four postal zones. The gravity models estimate that about half of all groceries are bought from retailers outside a consumer's home zone.

This paper dealt with large OD-matrices in an attempt to guarantee a relatively high resolution that matches the short-distance activity grocery shopping while at the same time covering a wider geographical area as required for large-scale outbreaks. Given the expected high proportion of intra-zonal flows, a particular focus was put on the length estimation of intra-zonal flows.

From a practical perspective the gravity model allows to identify zones where meaningful food flows can be expected. Further, it quantifies grocery inflow proportions and reveals retailers where consumers most probably have bought food items. The findings suggest that the last mile flow of food products deserves more attention from traceback modelers. Especially if models perform on a disaggregate level such as postal zones, a consumer's mobility needs to be integrated to avoid estimation errors. For this purpose, the gravity model is considered a useful tool that links retailers with consumers and estimates flow intensities. The benefit of an aggregate, modelled consumer-retailer interaction is particularly given at an early investigation stage where only limited socio-economic information about infected individuals are given and questionnaires on shopping habits are not available.

Nevertheless, we encourage further research on disaggregate shopping models for occasions where such information is available. Further, results need to be validated and tested on a national scale to ensure usability for widespread outbreaks. Lastly, shopping mobility data should be integrated into traceback models to assess potential performance improvements.

Acknowledgement. The project this report is based on was supported with funds from the German Federal Ministry for Education and Research (BMBF) in the context of the call "Civil Security - Critical structures and processes in production and logistics" under project number 13N15072.



References

1. World Health Organization: Foodborne Disease Outbreaks: Guidelines for Investigation and Control. WHO Library Cataloguing-in-Publication Data. World Health Organization, Geneva (2008)
2. Tinga, C., Todd, E., Cassidy, M., Pollari, F., Marshall, B., Greig, J., et al.: Exploring historical Canadian foodborne outbreak data sets for human illness attribution. *J. Food Prot.* **72**(9), 1963–1976 (2016)
3. Marvin, H.J.P., Janssen, E.M., Bouzembrak, Y., Hendriksen, P.J.M., Staats, M.: Big data in food safety: an overview. *Crit. Rev. Food Sci. Nutr.* **57**(11), 2286–2295 (2017)
4. Horn, A.L., Friedrich, H.: Locating the source of large-scale diffusion of foodborne contamination. *J. R. Soc. Interface* **16**(151), 1–11 (2019)
5. Manitz, J., Kneib, T., Schlather, M., Helbing, D., Brockmann, D.: Origin detection during food-borne disease outbreaks - a case study of the 2011 EHEC/HUS outbreak in Germany. *PLoS Curr.* (2014)
6. Norström, M., Kristoffersen, A.B., Görlach, F.S., Nygård, K., Hopp, P.: An adjusted likelihood ratio approach analysing distribution of food products to assist the investigation of foodborne outbreaks. *PLoS ONE* **10**(8), 1–13 (2015)

7. Kaufman, J., Lessler, J., Harry, A., Edlund, S., Hu, K., Douglas, J., et al.: A likelihood-based approach to identifying contaminated food products using sales data: performance and challenges. *PLoS Comput. Biol.* **10**(7), 1–10 (2014)
8. Infas: Mobilität in Deutschland - Ergebnisbericht (2017)
9. Veenstra, S.A., Thomas, T., Tutert, S.I.A.: Trip distribution for limited destinations: a case study for grocery shopping trips in the Netherlands. *Transportation (Amst)* **37**(4), 663–676 (2010)
10. Jonker, N.J., Venter, C.J.: Modeling trip-length distribution of shopping center trips from GPS data. *J. Transp. Eng. Part A Syst.* **145**(1), 04018079 (2019)
11. McFadden, D.: Disaggregate behavioral travel demand's RUM side a 30-year retrospective (2000)
12. Suhara, Y., Bahrami, M., Bozkaya, B., Pentland, A.(S.), Suhara, Y., et al.: Validating gravity-based market share models using large-scale transactional data (2019)
13. Cascetta, E., Pagliara, F., Papola, A.: Alternative approaches to trip distribution modelling: a retrospective review and suggestions for combining different approaches. *Pap. Reg. Sci.* **86**(4), 597–620 (2007)
14. Drezner, T.: Derived attractiveness of shopping malls. *IMA J. Manag. Math.* **17**, 349–358 (2006)
15. Hyman, G.M.: The calibration of trip distribution models. *Environ. Plan.* **1**, 105–112 (1969)
16. Furness, K.P.: Time function iteration. *Traffic Eng. Control* **77**, 458–460 (1965)
17. Suel, E., Polak, J.W.: Development of joint models for channel, store, and travel mode choice: grocery shopping in London. *Transp. Res. Part A Policy Pract.* **99**, 147–162 (2017)
18. Viegas, J.M., Martinez, L.M., Silva, E.A.: Effects of the modifiable areal unit problem on the delineation of traffic analysis zones. *Environ. Plan. B Plan. Des.* **36**(4), 625–643 (2009)
19. de Dios Ortúzar, D., Willumsen, L.G.: Modelling Transport. Wiley, Chichester (2011)
20. Martin, W., McGuckin, N.: Report 365: Travel Estimation Techniques for Urban Planning. Washington, DC (1998)
21. Huff, D.: Calibrating the huff model using ArcGIS business analyst (2008)
22. Open Street Map: OpenStreetMap Deutschland: Die freie Wiki-Weltkarte (2019). <https://www.openstreetmap.de/>
23. Statistische Ämter des Bundes und der Länder: ZENSUS2011 - Homepage (2018). https://www.zensus2011.de/EN/Home/home_node.html;jsessionid=8A55DF20B6CB474A1DB6DEFDD94B4949.1_cid389
24. Khatib, Z., Ou, Y., Chang, K.: Session #10 GIS and Transportation Planning (1999)
25. Kordi, M., Kaiser, C., Fotheringham, A.S.: A possible solution for the centroid-to-centroid and intra-zonal trip length problems. In: Gense, J., Josselin, D., Vandenbroucke, D. (es.) Multidisciplinary Research on Geographical Information in Europe and Beyond, Avignon, pp. 147–152 (2012)
26. Bhatta, B.P., Larsen, O.I.: Are intrazonal trips ignorable? *Transp. Policy* **18**, 13–22 (2010)
27. Manout, O., Bonnel, P.: The impact of ignoring intrazonal trips in assignment models: a stochastic approach. *Transportation (Amst)*, 1–21 (2018)
28. CZuber, E.: Geometrische Wahrscheinlichkeiten und Mittelwerte. T.B. Teubner, Leipzig (1884)
29. Larson, R., Odoni, A.: Urban Operations Research. Prentice Hall, New Jersey (1981)
30. Lebensmittel Zeitung: Ranking: Top 30 Lebensmittelhandel Deutschland 2018 (2018). <https://www.lebensmittelzeitung.net/handel/Ranking-Top-30-Lebensmittelhandel-Deutschland-2018-134606>

31. Edeka: Edeka Einzelhandel (2019)
32. Infas: Mobilität in Deutschland - Wissenschaftlicher Hintergrund (2019). <http://www.mobilitaet-in-deutschland.de/>
33. Mekky, A.: A direct method for speeding up the convergence of the furness biproportional method. *Transp. Res. Part B* **17B**(1), 1–11 (1983)
34. Cesario, F.J.: Parameter estimation in spatial interaction modeling. *Environ. Plan. A Econ. Sp.* **5**(4), 503–518 (1973)



The Effect of Social Media on Shaping Individuals Opinion Formation

Semra Gündüz^(✉)

Department of Computer Engineering, Faculty of Engineering,
Ankara University, 06345 Gölbaşı, Ankara, Turkey
gunduc@ankara.edu.tr

Abstract. In this paper, the influence of the social media on the opinion formation process is modeled during an election campaign. In the proposed model, peer-to-peer interactions and targeted online propaganda messages are assumed to be the driving forces of the opinion formation dynamics. The conviction power of the targeted messages is based on the collected and processed private information. In this work, the model is based on an artificial society, initially evenly divided between two parties. The bounded confidence model governs peer-to-peer interactions with a value of confidence parameter which leads to consensus. The targeted messages which was modeled as an external interacting source of information convert some weakly committed individuals to break this evenness. Both parties use the same methods for propaganda. It is shown that a very small external influence break the evenness of the opinion distribution which play significant role in the election results. Obtained opinion fluctuation time series have close resemblance with the actual election poll results.

Keywords: Opinion formation · Social media · Fake news · Fabricated news

1 Introduction

During the last decade internet and particularly online news services and social media networks have been the dominant information sharing channels. In the social media large groups of individuals, sharing similar interests form networks [1] which create mutual trust among the members of the network. Information coming from a member of the network is accepted and propagated by the members of the group without much criticism [2–4]. Such a free environment make the users vulnerable since as well as expressing their opinions they also reveal some personal data. Third parties may collect such personal information, process it and use for their purposes. Advertising agencies and political parties are willing use the personal information to convince or to convert individuals. Hence the freedom of expressing opinion and spreading information may be misused to propagate rumor, gossip and misleading or false information [5–7].

The question, whether such interventions effect the public opinion or not [8] is still an open debate. The effect of an information system such as TV, newspaper, blogs, on the evolution of public opinion is studied in [9]. Also, the analysis of the frequency of interaction is considered in [10]. It is also shown that, to use/support some concepts which are valuable or very sensitive for the society, such as religion, nationality, cultural issues, collective beliefs can also make a profit for politicians [11] and even the ideas adopted by the minority of the society can be supported by the majority, after such a process. The effect of the social network utilizing degree-dependent fitness and attributes when there are competing opinion diffusion is introduced in [12]. The role of the social media on the social polarization of the society is studied through both data [1,13] and model studies [3,6]. Although the social studies still do not have clear evidence on the influence of such misleading information flow on the social preferences, it is shown that false news propagation is faster and broader than the spread of true news due to the attractiveness of the false news [14].

Recently, two very important social events, namely Brexit – the British referendum to leave the European Union – and 2016 US presidential election campaigns are the striking examples of such social phenomena. During the campaign the fake news are not only fabricated but more than that the issues are carefully selected by using personal information of the social media users [15,17]. The collected personal information is used to convince and convert individuals. It is shown that in the US Presidential election campaigns some nodes and bots, i.e. automated accounts, in the social networks spread fabricated, fake, biased information, distort actual news, disseminate deceptive information [4,8,14,16]. The severity of the outside interruption can be seen by comparing the numbers: Only on 20 election stories, the number of Facebook engagements are 8.7 m fake news versus 7.3 m mainstream news starting from the beginning of August to the election day [18]. During a recent survey, nearly 85% of respondents stated that they believed fake news is a serious social problem [19].

The aim of this work is to build a model to study the effects of varying sources of fake and biased news on the opinion formation during an election campaign. The model society consist of N individuals with multi-component opinion on the election issues. On the decision-making process information, originating from different sources, play a vital role to build up opinion. For the simplicity election issues are limited in the model. The individuals exchange opinion on only three different issues (such as the economy, health services, security). Hence, each individual is identified by three real opinion values, but when it comes to making a decision on the vote, the choice is a result of combined opinion formed on all three issues. In this sense, the opinion structure of the model resembles Axelrod mode [20]. The individuals exchange opinion according to the bounded confidence model (BCM) [21]. Each individual is also subject to information flow through public services, social media, and online news sources. Some of this information may be fake, fabricated and even targets a particular individual. Targeted news specifically designed by considering the individual preferences [15] which is more effective on the non-committed individuals.

A new interaction is introduced to mimic the effects of the public services, social media, targeted, false or biased information spread.

The model is tested on four different cases:

1. Individual interact with others through peer-to-peer interactions no external sources affects the dynamics.
2. One of the parties send messages to convert less convinced individuals.
3. Both of the parties send messages to convert less convinced individuals.
4. Both parties send messages but one sends more convincing (fabricated) messages.

In all four cases individuals exchange opinion through social media channels. The acceptability of the messages of online news sources are controlled by using different probability values which are discussed in detail in the results section.

The rest of this paper is organized as follows. Section 2 provides a background for the proposed model with bounded confidence opinion dynamics. Section 3 is devoted for presentation of the simulation results. Finally Sect. 4 concludes the paper.

2 The Model

The proposed model is based on an artificial society with N individuals. The communication network is a fully connected network with nodes, $i = 1, \dots, N$. This topology allows every individual to interact with every other mutually. Even though it looks too simple; it eliminates the artifacts of more complicated network topology which is essential for our discussion. The position of each member of the society is labeled by Latin alphabets i, j, \dots . Each individual carries a three opinion component which is labeled by Greek alphabet, $\alpha = 1, 2, 3$.

Equation 1 defines the opinion of an individual who has three opinion components in the matrix form.

$$O = \begin{pmatrix} o_{11} & o_{12} & o_{13} \\ o_{21} & o_{22} & o_{23} \\ \vdots & \vdots & \vdots \\ o_{N1} & o_{N2} & o_{N3} \end{pmatrix} \quad (1)$$

Here N is the number of individuals, $o_{i\alpha}$, ($o_{i\alpha} \in \mathbb{R}^w$), are the opinion values of i^{th} individual on the α^{th} issue. All three opinion components are assigned randomly to each individual. Each opinion component has a Gaussian distribution with mean value $\langle O \rangle = \pm 1$, indicates two opposing views. In order to have some interaction between different views (different opinion individuals), variance of the Gaussian is used as the control parameter of the overlapping region.

At every interaction a randomly chosen pair of individuals exchange opinion on a randomly chosen issue. In each interaction individuals discuss on any one of

the three issues in concern. Opinion exchange is realized according to BCM [21] given in Eq. 2.

$$\text{if } |o_{i\alpha}(t) - o_{j\alpha}(t)| \leq \Delta \quad (2)$$

$$o_{i\alpha}(t) = \omega o_{i\alpha}(t-1) + (1-\omega)o_{j\alpha}(t-1)$$

$$o_{j\alpha}(t) = \omega o_{j\alpha}(t-1) + (1-\omega)o_{i\alpha}(t-1)$$

Here Δ is the tolerance threshold, ω is the opinion exchange factor, and t indicates discrete time steps.

Equation 3 defines the interaction of external influences with the individuals.

$$\text{if } o_{i\alpha} \leq \eta_{i,\alpha} \text{ then } o_{i\alpha} = \begin{cases} o_{i\alpha} + M_{i\alpha} & \text{if } P_{i\alpha} > r \\ o_{i\alpha} & \text{otherwise} \end{cases} \quad (3)$$

Here, $P_{i\alpha}$ is the probability of i^{th} individual receiving an information on the issue α , r is a uniform random number between 0 and 1, $o_{i\alpha}$ and $\eta_{i,\alpha}$ are the opinion and the tolerance level of the i^{th} individual on the issue α respectively.

At any discrete time step each individual, i , may receive a message on the issue α if he/she is supporting his/her idea less than a threshold value $\eta_{i,\alpha}$. The individual adopts the incoming message, $M_{i\alpha}$ with a probability of $P_{i\alpha}$. Hence the opinion value is replaced with a new value which is modified by the message content. Messages can be sent by either one of the two parties or by any existing friend. If the i^{th} individual is a supporter of one of the parties but not a committed follower of its political stands on some of the issues the message may convert the individual as a new supporter of the opposite party.

The society can have different sensitivities on different issues which are represented as s_α in the numerical simulation. To deal with this scenario three opinion weights can be assigned, s_α $\alpha = 1, 2, 3$, to each three opinion components (Eq. 4).

$$RO_i(t) = s_1 o_{i1}(t) + s_2 o_{i2}(t) + s_3 o_{i3}(t) \quad (4)$$

where $RO_i(t)$ is the resulted opinion of the i^{th} individual, s_α and $o_{i\alpha}(t)$ are the weights, $\alpha = 1, 2, 3$, and the opinions on different issues respectively.

A binary decision of the individual proceed continuous opinion components which is calculated by using the sign of the Eqs. 4 and 5.

$$D_i = \begin{cases} 1 & \text{if } RO_i > 0 \\ -1 & \text{if } RO_i < 0 \end{cases} \quad (5)$$

If the overall opinion of the individual is positive we say an individual is the supporter of the first view, otherwise the second view.

Both parties use regular media and social media communications to locate weakly committed individuals and try to win them over by sending messages. The system evolves a one-time step in discrete time as follows;

1. Choose randomly an individual, i , from the set $\{1, 2, \dots, N\}$
2. Choose randomly a neighbor, j , from the set $\{1, 2, \dots, N\}$
3. Choose randomly an issue, α , to discuss from the set $\alpha = 1, 2, 3$
4. Check the opinion component difference between individual i and individual j on issues α
5. If $diff = o_{i\alpha} - o_{j\alpha}$ is less than tolerance threshold, Δ , exchange opinion with the rule;

$$\begin{aligned} o_{i\alpha}(t) &= \omega o_{i\alpha}(t-1) + (1-\omega)o_{j\alpha}(t-1) \\ o_{j\alpha}(t) &= \omega o_{j\alpha}(t-1) + (1-\omega)o_{i\alpha}(t-1) \end{aligned}$$

6. If $o_{i\beta} < \eta$, where β is the issue on which external observer send messages,
7. Individual i receive a targeted message $M_{i\beta}$
8. Choose a random number, $r \in (0, 1)$
9. If $P_\beta > r$ where P_β is the probability to adopt the message, individual i accepts the message and update opinion $o_{i\beta} = o_{i\beta} + M_{i\beta}$
10. Repeat starting from the first step and continue N times.

The above steps describe 1 discrete time step. The system is followed until the final date of the campaign.

In the next section simulation results, obtained by applying the proposed model is introduced with figures.

3 Results and Discussion

The proposed model, described in Sect. 2 contains two different but complementary interactions among the members of an artificial society which consists of $N = 40000$ fully connected individuals. Simulations are carried on discrete time steps. A time step is defined as the number of interactions, $\mathcal{O}(N)$ which is sufficient for each individual to interact with at least with one neighbor and one outside news source. At each time slice the averages are taken over the opinion configurations. In the simulations, 500 different initial opinion configurations are created. The time span of the election campaign is chosen as 200 time steps.

The society, initially, consists of equally divided group of individuals. Each opinion component has a Gaussian distribution with mean value ± 1 , indicates two opposing views, and variance $\sigma^2 = 0.5$. With these choices, the Gaussian opinion distributions overlap at the origin. As the variance becomes closer to 1 the overlapping opinions increases. The individuals who constitutes the overlap region (uncommitted supporters of opposing ideas) are the targeted individuals by the external influences to persuade to their view.

The interaction parameters are grouped into two. The first group is related to peer-to-peer interactions while the second one is external influences.

1. Peer-to-Peer Interaction Parameters Two parameters, the tolerance limit, $\Delta_{i\alpha}$ and opinion exchange parameter ω controls the peer-to-peer interactions. $\Delta_{i\alpha}$ is taken as a constant for all members of the society and all issues, $\Delta_{i\alpha} = \Delta$.

The choice of tolerance parameter $\Delta = 1.216$ allow the individuals to interact with a wide range of opinion holders only excludes extremists. The opinion exchange parameter is taken as, $\Omega = 0.8$ which controls the speed of the opinion formation process.

2. The influence of the external sources

The sign of the resultant opinion (Eq. (4)) is the indicator of the vote where the relative weights of the issues are taken equal fir the simplicity of the discussions. The conviction parameter, η , is considered as a small value in the simulation studies the value is used as $\eta = 0.3$. The news acceptance probability, P_β , take different values according to the alignment of the opinion of the individual and the incoming news item. The message size, $M_{i\beta}$, is also an other parameter which changes at each interaction. It is taken as random value, $0 \leq M_{i\beta} < 0.5$.

Four different situations are considered: (a) averaged opinion with only peer-to-peer interactions ($P_\beta = 0$; for $\beta = 0, 1$), (b) one of the opinion supporters spread information by using mainstream and social media while the other opinion spread only by peer-to-peer interactions, ($P_0 = 0.5$; and $P_1 = 0.0$) (c) both opinion followers use the same means of external influences, ($P_\beta = 0.5$; for $\beta = 0, 1$) (d) both parties use influential sources together with peer-to-peer interactions, but one put more convincing arguments forward, $P_0 = 1.0$; and $P_1 = 0.5$.

3.1 Individual Interact with Others Through Peer-to-Peer Interactions No External Sources Affects the Dynamics

Figure 1 shows simulation results starting from different initial opinion configurations which may be interpreted as the opinion changes in a region during an election campaign (Paths of opinion). As it can be observed from the figure, (Fig. 1) each initial configuration converges a different final state. This situation resemble election results in different election zones. In different election zones, majority support may be on different parties but the overall votes are the decisive factor for the result of the election.

In fact non of these individual paths has much meaning, the result of the campaign is the average of all these paths. Figure 2(a) shows that if there is no external influences in the average both parties share the population almost equally.

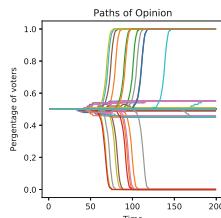


Fig. 1. Paths of opinion change starting from statistically independent initial configurations.

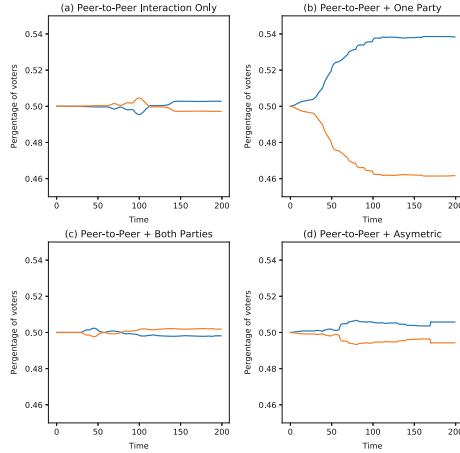


Fig. 2. Time evolution of the configuration averaged opinion distributions.

The dynamics of opinion formation without external influences can be better understood by observing the changes of the opinion distributions. Figure 3, show the averaged opinion distributions for four instances of the election campaign.

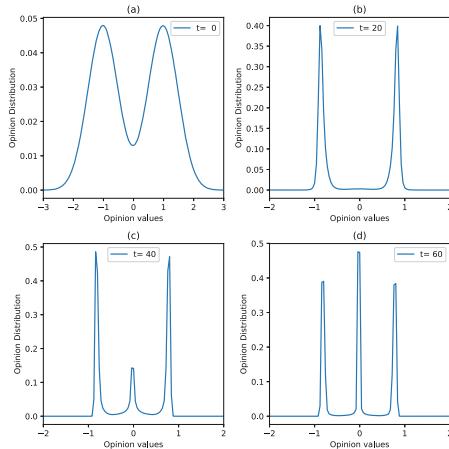


Fig. 3. The snapshots of the averaged opinion distribution without external influences.

At the initial stages of the campaign, $t = 0$ (Fig. 3(a)), the society is assumed to be equally divided on two opposing opinions. As soon as the campaign starts, bounded confidence dynamics unite individuals around the opposing opinions which sharpens the Gaussian opinion distributions. This situation is not stationary, a third peak start to appear around the origin, $t = 40$ (Fig. 3(c)).

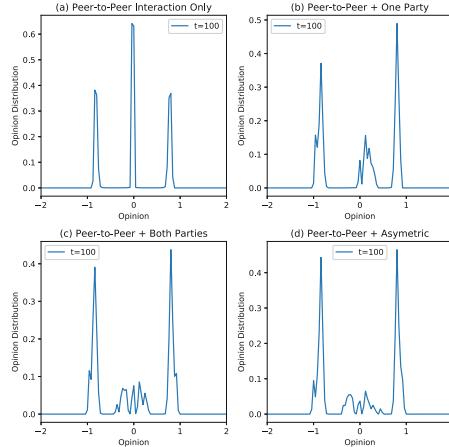


Fig. 4. Configuration averaged opinion distributions.

As the time passes, the supporters of both parties, apart from some extremists, converge towards a moderate opinion ($t = 60$, (Fig. 3(d))) and the distribution remains the same (Fig. 4(a)).

3.2 One of the Parties Send Messages to Convert Less Convinced Individuals

In any election campaign, ideally, both parties use the same means to convince individuals. Never the less, it is not always possible to maintain the same level of publicity or use of media for both parties. The uncommitted ($|O_{i,\beta}| < \eta$) electors remain under one-sided news bombardment which may change opinion of some of the individuals who aim to vote for one party without a deep conviction. The average daily progress of the opinion formation results are presented by Fig. 2(b).

If the outside sources can convince even a small group of uncommitted individuals it may be sufficient to win the election. The time dependent variations of the opinion distributions also give clear picture of the process of opinion changes. Figure 4(b) show that at $t = 100$, peer-to-peer interactions sharpen the Gaussian while the external influences change the heights of the Gaussian's. At the final stages of the election campaign the middle peak lean towards the opinion who use external sources to convince moderate individuals.

3.3 Both of the Parties Send Messages to Convert Less Convinced Individuals

Figures 2(c) and 4(c) show that if both parties are using external news sources and the social media to convince the less committed individuals, the picture is quite similar to the one seen for only peer-to-peer interaction case. The moderate individuals fluctuate between two opposing opinions, hence, the final election result is unpredictable, Fig. 2(c).

3.4 Both Parties Send Messages But One Sends More Convincing (Fabricated) Messages

The final consideration is that both parties use media and social media. One of the parties increase activities, send targeted, fake or fabricated messages, on the social media during the election period. This resemble the situation during 2016 US Presidential elections [22]. The situation is not as sewer as the one party usage of the social media (Discussed in Subsect. 3.2) but even such an effort difference can be sufficient for winning the election. For the fake news case, the message acceptance probability of the fake news is taken equal to the messages comming from friends, $0.5 < P_\beta < 1$. Figures 2(d) and 4(d) show that if one of the parties spreading fake or targetted news using external news sources and the social media to convince the less committed individuals, a small group of individuals are converted which is sufficient for a small majority.

4 Conclusions

Recently the internet is the primary source of acquisition of knowledge for the societies. This makes the internet a very powerful and unique. An exciting information, whether it is fake, fabricated or destructive propagate very fast among the members of the societies. Media can put forward some ideas or hide some information, by censoring, to make followers/members gain an advantage. In social interactions the spread of gossip and fabricated information has a very long history and it is not only limited with the online media [23]. Never the less the involvements of various data companies on the 2016 US presidential elections are publicly known and opened a debate on the violation of civil liberties [17]. Such a data-driven research needs of using an interdisciplinary approach. Using data science techniques to understand voter behavior on the segment of their ideas allow politicians to use digital- marketing strategies to reach individuals. Hence usage of powerful data analyzing techniques is becoming increasingly harmful to the civil liberties. As the 2020 US presidential election is approaching the studies on the effects of the fake and fabricated news and personalized, targeted messages gain upmost importance.

The present work aims to introduce a simple agent-based model to simulate the effects of using gathered information to send targeted messages during an election process. Recent studies show that societies are almost evenly divided on the main political issues. Hence such a targeted external information bombardment may be very effective to change the opinions. Since all parties may use the same techniques, a small percentage of, (1%–2%), opinion fluctuations may be decisive on the result of the elections. The above assumptions seem reasonably realistic considering voting processes such as Scottish referendum (55.3%–44.7%), 2016 US Election (46.1%–48.2%) and Brexit referendum (51.9%–48.1%). It is observed that if both parties compete equally the election results are unpredictable. If one of the parties use the technological power and social media more than the opponents, can easily gain the required small percentage

of the undecided voters. It is evident that in the next decade the use of artificial intelligence techniques to extract information from individuals social media history will be used more frequently unless some global legislative regulation on the use of private information.

References

1. Bessi, A., et al.: Homophily and polarization in the age of misinformation. *Eur. Phys. J. Spec. Top.* **225**, 2047–2059 (2016)
2. Mocanu, D., Rossi, L., Zhang, Q., Karsai, M., Quattrociocchi, W.: Collective attention in the age of (mis)information. *Comput. Hum. Behav.* **51**, 1198–1204 (2015)
3. Askitas, N.: Explaining opinion polarisation with opinion copulas. *PLoS ONE* **12**, e0183277 (2017)
4. Shao, C., et al.: Anatomy of an online misinformation network. *PLoS ONE* **13**, 1–23 (2018)
5. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015)
6. Del Vicario, M., Zollo, F., Caldarelli, G., Scala, A., Quattrociocchi, W.: Mapping social dynamics on Facebook: the Brexit debate. *Soc. Netw.* **50**, 6–16 (2017)
7. Schmidt, A.L., et al.: Anatomy of news consumption on Facebook. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3035–3039 (2017)
8. Bessi, A., Ferrara, E.: Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* **21** (2016). <https://doi.org/10.5210/fm.v21i11.7090>
9. Quattrociocchi, W., Caldarelli, G., Scala, A.: Opinion dynamics on interacting networks: media competition and social influence. *Sci. Rep.* **4**, 4938 (2014). <https://doi.org/10.1038/srep04938>
10. 1st Workshop on Social Media Analytics (SOMA '10), Washington, DC, USA, 25 July 2010
11. Gunduc, S.: The role of fanatics in consensus formation. *Int. J. Mod. Phys. C* **26**(03), 1550029 (2015)
12. Hu, H.: Competing opinion diffusion on social networks. *R. Soc. Open Sci.* **4**(11), 171160 (2017). <https://doi.org/10.1098/rsos.171160>
13. Bessi, A., et al.: Users polarization on Facebook and Youtube. *PLoS ONE* **11**, 1–24 (2016)
14. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**, 1146–1151 (2018)
15. cambridgeanalytica.org
16. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. *Commun. ACM* **59**, 96–104 (2016)
17. Bovet, A., Makse, H.A.: Influence of fake news in Twitter during the 2016 US presidential election. *Nat. Commun.* **10**, 7 (2019). <https://doi.org/10.1038/s41467-018-07761-2>
18. <https://www.statista.com/chart/6795/fake-news-is-a-real-problem/>
19. <https://www.statista.com/statistics/829314/fake-news-outside-groups-mainstream-media/>
20. Axelrod, R.: The dissemination of culture - a model with local convergence and global polarization. *J. Confl. Resolut.* **41**(2), 203–226 (1997)
21. Deffuant, G., Neau, D., Amblard, F., Weisbuch, G.: Mixing beliefs among interacting agents. *Adv. Compl. Syst.* **3**, 87–98 (2000)

22. Bovet, A., Morone, F., Makse, H.A.: Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Sci. Rep.* **8**, 8673 (2018)
23. Soll, J.: The long and brutal history of fake news. Politico (2016). <https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535>



A Network-Based Analysis of International Refugee Migration Patterns Using GERGMs

Katherine Abramski^(✉), Natallia Katenka, and Marc Hutchison

University of Rhode Island, Kingston, RI 02882, USA
keabral0@gmail.com, nkatenka@cs.uri.edu,
mlhutch@uri.edu

Abstract. Understanding determinants of migration is central to anticipating and mitigating the adverse effects of large-scale human displacement. Traditional migration models quantify the influence of different factors on migration but fail to consider the interdependent nature of human displacement. In contrast, network models inherently take into account interdependencies in data, making them ideal for modeling relational phenomena such as migration. In this study, we apply one such model, a Generalized Exponential Random Graph Model (GERGM), to two different weighted-edge networks of international refugee migration from 2015, centered around Syria and the Democratic Republic of Congo (DRC), respectively. The GERGM quantifies the influence of various factors on out-migration and in-migration within the networks, allowing us to determine which push and pull factors are largely at play. Our results indicate that both push factors and pull factors drive migration within the DRC network, while migration within the Syria network is predominately driven by push factors. We suspect the reason for this difference may lie in that the conflict in Syria is relatively recent, in contrast to the conflict in the DRC, which has been ongoing for almost two decades, allowing for the establishment of systematic migration channels, migration networks, and resettlement, all which are related to pull factors, throughout the years.

Keywords: Networks · GERGM · Refugees · Migration

1 Introduction

1.1 Challenges in Migration Research

Being able to anticipate mass migrations before they occur is central to mitigating the negative effects of large-scale forced human displacement. In order to do so, it is essential to understand determinants of migration and their influence on migrant flows. Factors that influence migration can be broadly classified as either push factors or pull factors. Push factors are associated with the origin country/region and play a role in the decision to migrate (e.g. poverty, violence) while pull factors are associated with the destination country/region and play a role in determining where migrants go (e.g. political stability, economic prosperity) [18]. Intervening obstacles and personal factors such as migration policy and individual preferences can have a high degree of influence on migration as well [19].

One major challenge in migration research is taking into account all of these different factors. A large body of research has been dedicated to understanding determinants of migration, however many theories have focused on just one group of factors, such as economic factors or political factors, rather than the collective interplay and influence of many different factors together [6]. One model that has become popular for assessing the influence of many different determinants of migration simultaneously is the modified gravity model [7, 16, 23, 24]. It is essentially a regression model based on the original gravity model, which assumes that migration flows between two countries are proportional to their size (population or GDP) and inversely proportional to the geographical distance between them [23]. The modified gravity model, however, is adapted to include other factors that are thought to influence migration, and takes the following form [12]:

$$\ln M_{ij} = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln P_j + \beta_3 \ln D_{ij} + \sum \beta_k \ln X_k + \varepsilon_{ij} \quad (1)$$

where M_{ij} is migration between a pair of countries, P_i and P_j are the sizes of the two countries, D_{ij} is the distance between them, and the X_k term includes additional determinants of migration specified by the researcher such as economic, political, and social factors. The respective influence of all these factors on migration is measured by the parameters β_1 , β_2 , β_3 , and β_k , which can be easily estimated with ordinary least squares [22]. While the modified gravity model is powerful in its simplicity, it has a significant drawback. It assumes independence between observations, making it less than ideal for modeling relational phenomena that is interdependent in nature, such as migration. Since migration flows between pairs of countries have been shown to influence migration flows between other pairs of countries [11], models that do not consider this dependency factor may produce results that inaccurately reflect the relationships between dependent and independent variables [3].

Statistical network models offer a solution to this problem because they inherently take into account interdependencies in data. Among the most commonly applied network models are latent space models [13], the quadratic assignment procedure [17], Stochastic Actor Oriented Models [25, 26] and Exponential Random Graph Models (ERGMs) [14, 28] and their extensions. After a careful review of the strengths and weaknesses of these models [10, 20], we concluded that the ERGM family of models is most suitable for this analysis. ERGM family models are similar to regression models in that they are capable of measuring the influence of various factors simultaneously, but they are distinct in that they also consider interconnectedness as a significant influential factor in itself.

1.2 ERGM Family of Models

ERGM family models have two main features: they measure the influence of covariates (node and edge attributes) on the structure of the network, and they model the prominence and significance of structural dependencies (e.g. reciprocity, in-stars, out-stars) of the network [10]. This allows the researcher to test specific hypotheses about

how certain network structures drive the formation of the network. Even if the researcher is uncertain about the types of interdependencies that underlie the formation of the network, the model can shed light on specific relational patterns inherent in the data that may be intuitively difficult to recognize [10].

In the context of ERGMs, the observed network of interest, Y , can be thought of as a single observation from a multivariate distribution where many other realizations of the network are possible [3]. The goal is to select features of the observed network that differentiate it from a random draw from the uniform distribution of all other possible networks with the same number of nodes that could be observed [4]. These features, which can include both covariates and structural components, are incorporated in the model in a set of statistics computed on the network. The parameters of the model are then estimated to maximize the likelihood of observing the network of interest, Y . These parameters tell the researcher how the covariates and the inherent network structures drive the formation of the network [3].

The ERGM is a flexible tool because it relies on only two assumptions. First, it assumes that the network statistics calculated on the observed network are the expected values of those statistics across all possible graphs. This is a strong assumption, but in many cases the observed network is the only network we can possibly observe, so it is the best available estimate [3]. Second, the ERGM assumes that the model is specified correctly, meaning that only network statistics chosen by the researcher influence the probability that Y is observed [3]. Consequently, the challenge lies in choosing the covariates and network structures to include in the model. This decision should be informed based on prior knowledge about the behavior and dynamics of the network.

The classical ERGM is configured for static networks with binary edges. Weighted-edge networks can be coerced to binary-edge networks using a thresholding technique, but this can result in the loss of important information that is crucial to the analysis. In 2012, Desmarais and Cranmer addressed this problem and developed an ERGM family model that is adapted for weighted-edge networks, the Generalized Exponential Random Graph Model (GERGM), which we use in this analysis [9].

The decision to apply a model from the ERGM family for this analysis was in part motivated by two recent studies that used ERGM family models to model migration. In 2012, Desmarais and Cranmer applied a GERGM to investigate the effects of unemployment, temperature, distance, income, and population as well as structural components on interstate migration in the USA [9]. In 2017, Windzio applied a temporal ERGM adapted for longitudinal networks to a global migration network over four years in order to understand and quantify the influence of geographic, demographic, economic, religious, linguistic, and historical factors on international migration [30]. These studies lay the foundation for using ERGM family models for modeling migration and are useful for informing model specification.

In this analysis, we build upon that foundation by applying a GERGM to two different networks of international refugee migration composed of two different sets of countries. We have two main goals. First, we aim to explore the capabilities of the GERGM for modeling specific cases of international refugee migration. Second,

we aim to explore similarities and differences in refugee migration patterns between the two different networks. Specifically, we are interested in seeing how various determinants of migration and interdependencies influence out-migration and in-migration differently in the two networks.

1.3 Data Description

For this analysis, we construct two networks in which nodes represent countries and edges represent refugee migration. Refugees are defined according to the UNHCR population statistics database, from which the data for the edges were collected [27]. An edge from node A to node B represents the number of refugees originally from country A that reside in country B in 2015, thus the edges are both directed and weighted. It is important to note that these numbers are a cumulative snapshot, reflecting the total number of refugees that migrated *up to 2015* and remained there, rather than the number of refugees that migrated *in 2015*, so we cannot know when the refugees arrived in the receiving country. Because of the extreme skewness of the distribution of the edge weights, we took the log transformation of all the edges, so edges in this analysis represent the log number of refugees from one country residing in another.

The two networks are each composed of different sets of 12 countries, selected in such a way to reveal patterns in refugee migration from Syria and the Democratic Republic of Congo (DRC), respectively. Each network is composed of the country of interest and the 11 countries with the highest numbers of refugees from the country of interest as at 2015. While nodes were selected specifically to capture migration patterns related to Syria and the DRC, the edges in the networks do not represent migration exclusively from these countries, rather, they represent migration between *all* nodes. Network visualizations of the two networks can be found in Fig. 1.

We chose to investigate patterns in migration related to Syria and the DRC because they both have had very high levels of refugee migration out. Refugee migration from these countries is similar in that it is driven mainly by violent conflict in both cases, but the distinction between the history of the conflicts should be noted. The conflict in Syria began in 2011, making it relatively recent in comparison to the 2015 data used in this analysis, while the conflict in the DRC began over a decade earlier in the late 1990's. We also make the distinction that the Syria network is composed mostly of Middle Eastern and European countries while the DRC network is composed mostly of African countries. The combinations of the nodes in the networks are important to note since these countries have their own long histories of relations with each other which greatly influence the dynamics within the networks.

To investigate the influence of various push and pull factors on migration in the networks, we include several node attributes as well as one edge attribute. We included log population following from the assumption of the gravity model that higher population is associated with higher migration flows [23]. To investigate economic and political factors, which have been shown to be among the most significant drivers of migration [2], we included log GDP per capita, unemployment rate (percentage of total workforce), and political terror scale (level of political violence and terror that a country experiences: 1-low, 5-high). To investigate societal factors, we included excluded population (percentage of ethnic minority population excluded from

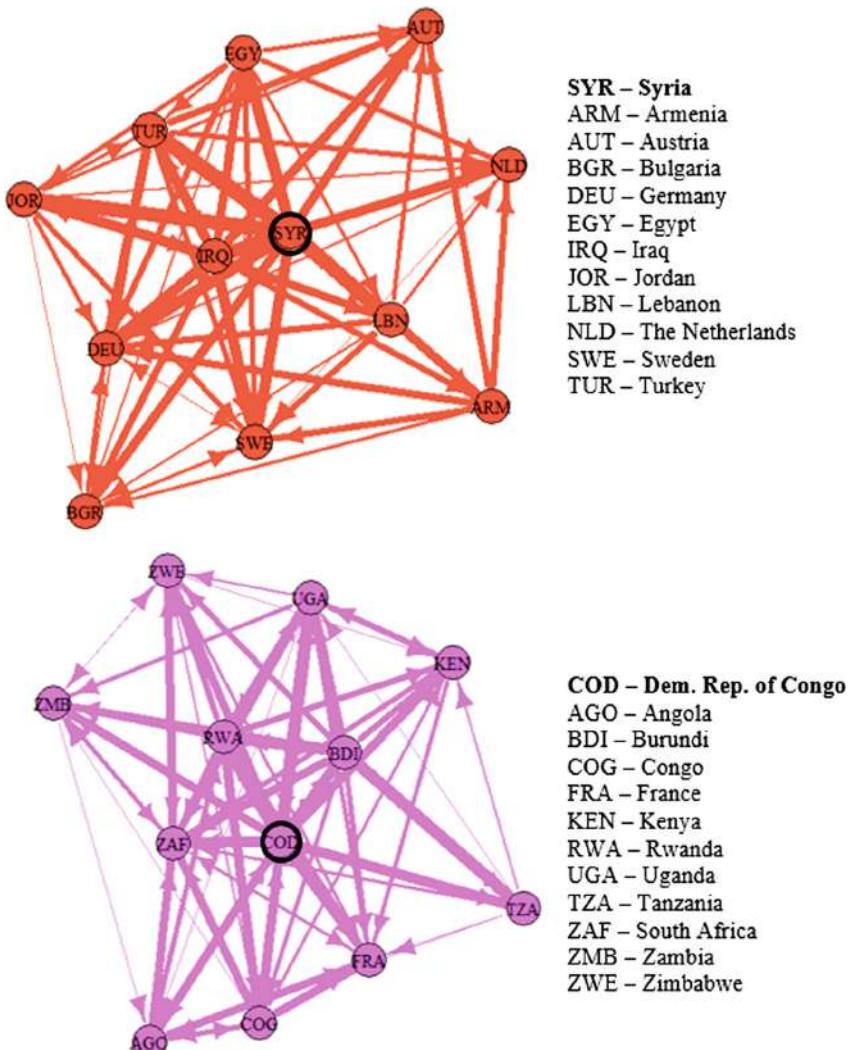


Fig. 1. The Syria network (top) composed of Syria and the 11 countries with the most Syrian refugees as at 2015, and the DRC network (bottom) composed of DRC and the 11 countries with the most Congolese refugees as at 2015. Legends provide keys for country codes.

government) and ethnic fractionalization (measure of ethnic diversity: the probability that two randomly chosen individuals will be of different ethnicity), as ethnic fragmentation has been found to be linked to increased levels of conflict [1]. Also following from the gravity model, we included geodistance – the geographical distance (km) between pairs of countries – as an edge attribute since geographical proximity has been shown to be positively related to migration between countries [15, 21]. All data are as at 2015. These node and edge attributes are included in the model as covariates, together with structural components, as features that are thought to drive the formation of the networks.

2 Methods

2.1 ERGM

The analytical form of the GERGM follows directly from that of the classical ERGM. If Y is the observed network of n nodes and Y^* is a random network of n nodes, then the probability of observing the network Y rather than all other possible realizations of Y^* can be expressed as a function of the set of statistics:

$$P_\theta(Y^* = Y) = \frac{\exp\{\theta^T \mathbf{h}(Y)\}}{\sum_{all\ graphs\ y^*} \exp\{\theta^T \mathbf{h}(Y^*)\}} \quad (2)$$

where θ^T is a vector of parameters to be estimated and $\mathbf{h}(Y)$ is a vector of networks statistics [4]. The statistics contain information about the covariates and the network structures, as specified by the researcher, and the parameters quantify their respective influence on the formation of the network.

Node and edge covariates, X_n and X_d respectively, are included in the vector of statistics as $\mathbf{h}_{X_n}(Y, X_n) = \sum_{i \neq j} X_i X_j Y_{ij}$ and $\mathbf{h}_{X_d}(Y, X_d) = \sum_{i \neq j} X_{ij} Y_{ij}$ [3]. The covariates should be chosen by the researcher such that high values of that covariate will either decrease or increase the probability of observing an edge. In this way, the ERGM is similar to a regression model because it can quantify the influence of some covariate on an outcome, i.e., the specific combination of edges that form the network. If some covariate X has a positive effect, then a higher value of X should increase the probability of an edge in Y .

The network structures are included in the vector of statistics similarly. For example, mutual dyads (reciprocity) is accommodated as $\mathbf{h}_R(Y) = \sum_{i < j} Y_{ij} Y_{ji}$ while in-two-stars and out-two-stars are accommodated as $\mathbf{h}_{in}(Y) = \sum_i \sum_{j < k \neq i} X_{ji} X_{ki}$ and $\mathbf{h}_{out}(Y) = \sum_i \sum_{j < k \neq i} X_{ij} X_{ik}$ [8]. The challenge lies in choosing the network structures to include in the vector of statistics. The researcher must choose the network structures that are believed to increase the probability of observing the observed network Y based on what drives its formation, thus distinguishing the model from a regression model (e.g. the modified gravity model) which does not account for interdependence.

2.2 GERGM

While the GERGM is derived directly from the classical ERGM and has the same assumptions, the procedures for specification and estimation of the GERGM are slightly different since the analytical form must account for weighted edges. Specification is a two-step process. First, a joint distribution that captures the dependencies of interest of the observed network Y is defined on a restricted network, $X \in [0, 1]^m$ where m is the total number of directed edges between nodes [9, 29]. Note that X has the same vertices as Y , but the edge values are continuous and bounded between zero and one. Then, X is transformed onto the support of Y through an appropriate transformation function, which creates a probability model for Y [9, 29].

In the first step, a set of network statistics, \mathbf{h} is defined to contain information about the covariates and the network structures, as with the ERGM. Then a probability distribution for X is defined by modifying the ERGM formula to have a convergent sum in the denominator for a bounded network [9, 29]:

$$f_X(X, \theta) = \frac{\exp(\theta^T \mathbf{h}(X))}{\int_{[0,1]^m} \exp(\theta^T \mathbf{h}(Z)) dZ} \quad (3)$$

where $\theta \in \mathbb{R}^p$ is the vector of parameters and $\mathbf{h} : [0, 1]^m \rightarrow \mathbb{R}^p$ is formulated to represent the joint features of Y in the distribution of X [9, 29]. This specification resembles that of the ERGM except the edges are now modeled as continuous, taking values between zero and one [29].

In the second step, the restricted network X is transformed onto the support of the observed network Y by applying a parameterized one-to-one monotone increasing transformation function $T^{-1} : [0, 1]^m \rightarrow \mathbb{R}^m$ to the m edges of the restricted network. Specifically, for each pair of distinct nodes i, j we have $Y_{ij} = T_{ij}^{-1}(X, \beta)$ where $\beta \in \mathbb{R}^k$ parameterizes the transformation to capture the marginal features of Y [9, 29]. This transformation allows for the specification of the GERGM such that the basic structure, strength, and flexibility of the classical ERGM are maintained, only now the vector of statistics \mathbf{h} is specified on a transformation of the network rather than the network in its observed form [9]. The GERGM, which is the pdf of Y , can be written [9, 29]:

$$f_Y(Y, \theta, \beta) = \frac{\exp(\theta^T \mathbf{h}(T(Y, \beta)))}{\int_{[0,1]^m} \exp(\theta^T \mathbf{h}(Z)) dZ} \prod_{ij} t_{ij}(Y, \beta) \quad (4)$$

where $t_{ij}(Y, \beta) = \frac{dT_{ij}(Y, \beta)}{dY_{ij}}$. When choosing the transformation T^{-1} , the distribution of the data should be considered, and while there is flexibility in this choice, it is wise to select a transformation such that T_{ij}^{-1} is an inverse cdf as it leads to beneficial properties [29]. The parameters of the GERGM are estimated by maximizing the likelihood function. The exact computation of the likelihood function is almost always too computationally demanding, so the likelihood must be approximated. This can be done with MCMC using Gibbs sampling [9] or alternatively Metropolis-Hastings methods can be used [29].

The estimated parameter values can be interpreted similarly to regression coefficients. If a parameter estimate is significantly different from zero, we can conclude that its corresponding statistic significantly effects the probability of observing a particular instance of that network, controlling for the other statistics included in the model, suggesting that the patterns observed in the network of interest did not occur by chance [9]. Significant parameters affect the width of the edge conditional on the rest of the network. For our analysis, this means that the parameter estimates allow us to understand which push and pull factors and structural dependencies drive out-migration and in-migration within the two networks.

2.3 Model Specification

The main challenge of the GERGM is the correct specification of the model which includes choosing the covariates and structural components to be included in the statistics vector, as well as the selection of the transformation function. In this study, the selection of these components was determined based on the properties of the data, previous research, and background knowledge about the problem of interest.

We applied the same GERGM model specifications to both networks. We included three structural components, six node covariates, and one edge covariate in the model. For structural components, we included mutual dyads, also called reciprocity [8], which measures the extent to which there is mutual migration between a pair of countries, in-stars, which measures “popularity” [8], the tendency of some countries to receive more refugees than others, and out-stars, which measures “sociality” [8], the tendency for some countries to have more out-migration than others. All previously mentioned node attributes were included as both sender and receiver effects to measure their respective influence on out-migration and in-migration, and the edge attribute geodistance was included to measure the influence of geographical proximity on migration within the networks.

For the transformation function we chose the log Cauchy function because it is suitable for highly skewed data, considering that the edge weight distributions of both networks are extremely skewed. Since the log Cauchy transformation is adapted for non-zero data and the networks contained some instances of non-edges, we added a negligible non-zero term to all edges.

3 Results

The results of the parameter estimates and standard errors for each of the statistics included in the models can be found in Table 1, visualizations of the parameter estimates and their corresponding credible intervals can be found in Fig. 2, and diagnostics plots for the models can be found in Fig. 3.

For the structural components, many of the parameter estimates are significant for both networks. There is a positive parameter estimate for mutual dyads for the Syria network. While this is somewhat surprising, it is possible that the conflicts or events responsible for this mutual migration took place at different times. There are positive parameter estimates for out-stars for both networks. This is unsurprising, since we

would expect Syria and the DRC to have more out-migration than the other countries in the networks. The in-stars parameter estimates are negative for both networks, and even more so for the Syria network. This suggests that within the networks, there are no dominating countries receiving many more refugees than the rest, which makes sense, since all the countries included in the networks are receiving high numbers of refugees from Syria and the DRC, respectively.

Significant covariate parameter estimates affect the width of the edge conditional on the rest of the network. We interpret the parameter estimates for the node covariates as sender effects (push factors) and receiver effects (pull factors), having either a positive or negative relationship to out-migration and in-migration respectively. So, if a covariate has a positive sender parameter estimate, then a node with a higher value of that covariate is expected to have a wider edge going out of it (more out-migration) compared to a node with a lower value of that covariate, all else equal. The extent to which the edge is narrower or wider depends on the magnitude of the coefficient and is a function of the transformation function.

Table 1. The parameter estimates and standard errors (SE) for the two networks. Asterisks indicate parameter estimates that are significant at the 10% (*), 5% (**), and 1% (***) level.

	Syria		DRC	
	Estimate	SE	Estimate	SE
Intercept	-36.0***	2.68	-31.6*	19.10
Mutual	6.98***	2.04	-0.13	1.47
Out-stars	1.56***	0.11	0.62***	0.22
In-stars	-3.00***	0.31	-0.72***	0.27
log population sender	0.69***	0.05	0.41	0.32
log GDP sender	-8.13***	0.16	-3.16***	0.69
Unemployment sender	1.37***	0.07	0.05	0.26
Excluded population sender	1.63***	0.05	2.00***	0.21
Ethnic fractionalization sender	2.22***	0.07	-1.66***	0.21
Political terror scale sender	-1.63***	0.13	1.83***	0.27
log population receiver	0.38***	0.10	1.81	1.36
log GDP receiver	-0.03	0.11	1.15*	0.64
Unemployment receiver	-0.05	0.04	1.38***	0.36
Excluded population receiver	0.04	0.03	0.53**	0.22
Ethnic fractionalization receiver	0.03	0.04	-0.33	0.21
Political terror scale receiver	-0.12*	0.07	-0.07	0.21
geodistance	0.03	0.03	0.50	0.56

The results of the covariate parameter estimates demonstrate very different patterns for the two networks. For the Syria network, there are positive sender effects for log population, unemployment, excluded population, and ethnic fractionalization,

negative sender effects for log GDP and political terror scale, positive receiver effects for log population, and negative receiver effects for political terror scale. For the DRC network, there are positive sender effects for excluded population, and political terror scale, negative sender effects for log GDP and ethnic fractionalization, positive receiver effects for log GDP, unemployment, and excluded population, and no negative receiver effects.

While many of these results are unsurprising considering what we know about migration dynamics, it is interesting to see opposite signs in sender effects for ethnic fractionalization and political terror scale for the two networks as well as positive receiver effects for unemployment and excluded population for the DRC network. We are also surprised that geodistance does not have a significant effect in either network. An in-depth analysis of the social, economic, and political history of all the countries included in these networks could provide insight into the specific patterns these results suggest.

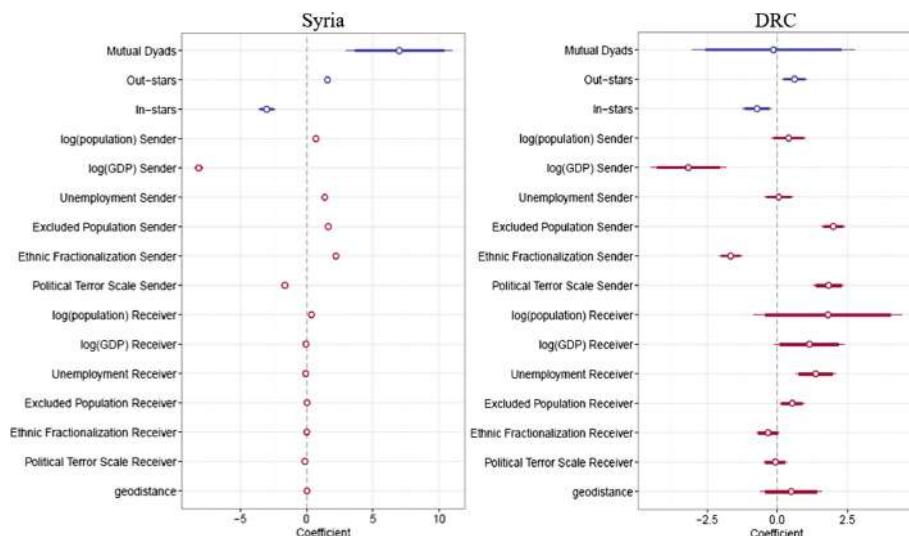


Fig. 2. The parameter estimates and credible intervals for the statistics for the Syria network (left) and the DRC network (right).

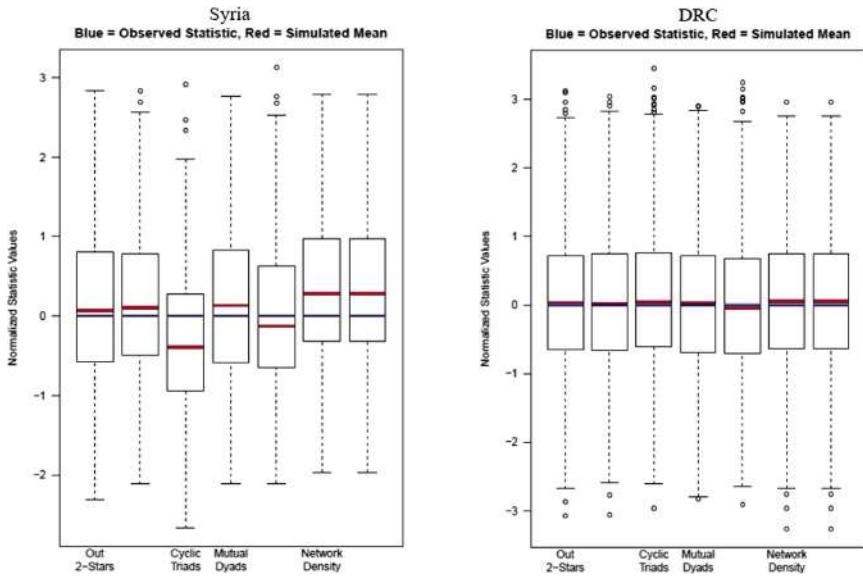


Fig. 3. Diagnostics plots display the network statistics calculated on the observed network and the expected values of those statistics across simulated networks. We see that the two are very close for both models indicating a good fit.

4 Discussion

4.1 Interpretation of Results

Considering the results for the sender and receiver effects for the two networks separately, there are considerably more significant sender effects (push factors) than receiver effects (pull factors) for the Syria network, while there are both sender and receiver effects for the DRC network, suggesting that migration within the Syria network is predominately driven by push factors while migration within the DRC network is driven by both push factors and pull factors. This pattern would have been even stronger had some of the credible intervals of the parameter estimates for the DRC network been slightly narrower.

We suspect the reason for this pattern has to do with when the conflicts in Syria and the DRC began relative to 2015. Migration within the DRC network reflects a snapshot of the refugee situation almost two decades after the start of the conflict in the DRC. This has allowed time for NGOs to step in and establish organized migration routes in the DRC and neighboring countries as well as the development of community networks, which have been shown to be a significant pull factor for refugees [18]. Also, since migration is often a process with two stages – first migrants tend to flee to a neighboring country to seek immediate safety, and then they seek a more permanent settlement [5] – many DRC refugees may have resettled from their original destinations before 2015. In contrast to the conflict in the DRC, the conflict in Syria began in 2011, so most of the edges from Syria within the network reflect migration that occurred over

a period of just 4 years, leaving little time for the establishment of organized migration routes, community networks, and refugee resettlement. Most Syrian refugees have fled to other countries in a non-systematic fashion, prioritizing safety regardless of what the destination country has to offer. This may offer some explanation for the different patterns in push and pull factors that we see from the results.

Since the model considers the networks collectively rather than migration exclusively from Syria and the DRC, the results reflect the migration dynamics between all countries within the networks. A thorough analysis of the histories and relational dynamics between all countries in the networks could provide valuable insight about which nodes and edges are largely driving these results. In future research, it could be useful to look at the number of refugees that migrated in a given year, rather than a cumulative snapshot as we did in this analysis, allowing researchers to link migration patterns to specific events that occurred at certain points in time.

4.2 Limitations

The GERGM has proven to be a powerful and effective tool that is useful for modeling migration because of its inherent capacity to account for interdependencies within the data, however there are some important limitations that should be considered when applying this model. GERGMs are computationally intensive and take a very long time to estimate. They are also difficult to fit to larger networks and sparse networks. GERGMs also rely on the assumption that the model is specified correctly, meaning it includes all variables that play a role in the formation of the network – both covariates and network structures. This assumption is difficult to satisfy considering that some influential variables are difficult or impossible to measure (migration policy, migrant preferences, other extraneous factors). Finally, the interpretability of the results depends largely on the properties of the nodes and edges that comprise the network. In this analysis, the networks are essentially sub-networks of the global migration network of all countries in the world. By choosing to look at only sub-networks, we essentially excluded part of the story, which introduces inherent biases. However, applying the model to the global migration network of all countries would be problematic nonetheless because patterns and dynamics unique to specific groups of countries would be lost, and it would be difficult to pinpoint which nodes and edges are largely responsible for the patterns suggested by the results.

With these challenges in mind, when applying GERGMs, we emphasize the importance of strategically choosing the nodes and edges to include in the network in order to facilitate model estimation and to maximize the interpretability of the results.

References

1. Akee, R.K.Q., Basu, A.K., Chau, N.H., Khamis, M.: Ethnic fragmentation, conflict, displaced persons and human trafficking: an empirical analysis. IZA Discussion Paper 5142, Institute for the Study of Labor (IZA) (2010)
2. Altai Consulting: Migration trends across the Mediterranean: connecting the dots. IOM MENA Regional Office, June 2015

3. Cranmer, S.J., Desmarais, B.A.: Inferential network analysis with exponential random graph models. *Polit. Anal.* **19**(1), 66–86 (2011)
4. Cranmer, S.J., Desmarais, B.A., Menninga, E.J.: Complex dependencies in the alliance network. *Confl. Manage. Peace Sci.* **29**(3), 279–313 (2012)
5. Day, K., White, P.: Choice or circumstance: the UK as the location of asylum applications by Bosnian and Somali refugees. *GeoJournal* **55**, 15–26 (2001)
6. de Haas, H.: The determinants of international migration: conceptualizing policy, origin and destination effects. IMI Working Paper No 32. International Migration Institute, Oxford (2011)
7. Dedeoğlu, D., Deniz Genç, H.: Turkish migration to Europe: a modified gravity model analysis. *IZA J. Dev. Migr.* **7**, 17 (2017)
8. Denny, M.: The importance of generative models for assessing network structure. Pennsylvania State University (2016)
9. Desmarais, B.A., Cranmer, S.J.: Statistical inference for valued-edge networks: the generalized exponential random graph model. *PLoS ONE* **7**(1), e30136 (2012)
10. Desmarais, B.A., Cranmer, S.J.: Statistical inference in political networks research. In: *The Oxford Handbook of Political Networks*. Oxford University Press (2017)
11. Görlich, J.S., Motz, N.: Refuge and refugee migration: how much of a pull factor are recognition rates? Bocconi University and Universidad Carlos III de Madrid (2017)
12. Greenwood, M.J.: Modeling migration. In: Kempf-Leonard, K. (ed.) *Encyclopedia of Social Measurement*, vol. 2, pp. 725–734. Elsevier, New York (2005)
13. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* **97**(460), 1090–1098 (2002)
14. Holland, P.W., Leinhardt, S.: An exponential family of probability distributions for directed graphs. *J. Am. Stat. Assoc.* **76**(373), 33–50 (1981)
15. Iqbal, Z.: The geo-politics of forced migration in Africa, 1992–2001. *Confl. Manage. Peace Sci.* **24**, 105–119 (2007)
16. Karemera, D., Iwuagwu Ogueledo, V., Davis, B.: A gravity model analysis of international migration to North America. *Appl. Econ.*, 1745–1755 (2010)
17. Krackardt, D.: QAP partialling as a test of spuriousness. *Soc. Netw.* **9**(2), 171–186 (1987)
18. Langley, S., Vanore, M., Siegel, M., Roosen, I., Rango, M., Leonardelli, I., Laczko, F.: The push and pull factors of asylum related migration: a literature review. European Asylum Support Office (2016)
19. Lee, E.S.: A theory of migration. *Demography* **3**(1), 47–57 (1996)
20. Leifeld, P., Cranmer, S.J.: A theoretical and empirical comparison of the temporal exponential random graph model and the stochastic actor-oriented model. In: *The 7th Political Networks Conference*, McGill University, May 2014
21. Neumayer, E.: Bogus refugees? The determinants of asylum migration to Western Europe. *Int. Stud. Quart.* **49**(3), 389–410 (2005)
22. Poot, J., Alimi, O., Cameron, M.P., Maré, D.C.: The gravity model of migration: the successful comeback of an ageing superstar in regional science. Institute for the Study of Labor. Discussion Paper No. 10329 (2016)
23. Ramos, R.: Gravity models: a tool for migration analysis. *IZA World Labor* **239**, 1–10 (2016)
24. Ramos, R., Suriñach, J.: A gravity model of migration between ENC and EU. Research Institute of Applied Economics (2013)
25. Snijders, T.A.B.: The statistical evaluation of social network dynamics. *Sociol. Methodol.* **31**, 361–395 (2001)

26. Snijders, T.A.B., van de Bunt, G.G., Steglich, C.E.G.: Introduction to stochastic actor-based models for network dynamics. *Soc. Netw.* **32**(1), 44–60 (2010)
27. UNHCR. The UN Refugee Agency. Population Statistics, 30 August 2019. <http://popstats.unhcr.org/en/overview>
28. Wasserman, S., Pattison, P.: Logit models and logistic regressions for social networks: I. An introduction to markov graphs and p^* . *Psychometrika* **61**(3), 401–425 (1996)
29. Wilson, J., Denny, M., Bhamidi, S., Cranmer, S.J., Desmarais, B.A.: Stochastic weighted graphs: flexible model specification and simulation. *Soc. Netw.* **49** (2016)
30. Windzio, M.: The network of global migration 1990–2013: using ERGMs to test theories of migration between countries. *Soc. Netw.* **53**, 20–29 (2017). SON-1045

Social Networks



Friendship Formation in the Classroom Among Elementary School Students

Raúl Duarte-Barahona^(✉), Ezequiel Arceo-May,
and Rodrigo Huerta-Quintanilla

Departamento de Física Aplicada, Centro de investigación y de Estudios Avanzados
del Instituto Politécnico Nacional, Mérida, Yucatán, Mexico
raledubar@gmail.com
<https://www.mda.cinvestav.mx>

Abstract. We used Exponential Random Graph Models (ERGMs) to determine how important is the role played by homophily, transitivity and preferential attachment effects in the formation of friendship networks obtained from data collected in 3 elementary schools; two are located in the rural area and the other is in the urban area of Yucatán, México. Structural terms were considered, as well as individual attributes (gender and scholar grade) of each student in the network. We hypothesize three network effects thought to contribute to the observed structure: homophily, triad closure and preferential attachment. Assessment model was done using p-value, Akaike information criteria (AIC) and a graphical goodness of fit (GOF). Results from exponential random graph models support our hypothesized homophily and triad closure effects. All friendship networks in our investigation had positive and significant triad closure and homophily effects. On the other hand, we do not find evidence for preferential attachment processes. Well connected students are not more prone to gain additional links than sparsely connected students.

Keywords: ERGM · Graph theory · Friendship networks

1 Introduction

A few decades ago, the accumulation of data on empirical networks progressed slowly thanks to the arduous work of a few dedicated to the collection task. While it becomes relatively less difficult to acquire network data than years before (thanks to the internet and social media), there are still population segments in blind spots. Such is the case of minors, who cannot have accounts on social networks before reaching a threshold age: for example, Facebook, Twitter and Instagram required its users to have at least 13 years. A work oriented towards that population sector is the paper entitled *Modeling Social Network Topologies in Elementary Schools* [1]. In that work, the authors determined the topological structures of three elementary schools, without considering individual attributes of students. Thus, they ignored social forces like gender homophily

(the preference for interaction with individuals of the same gender) and how this force interacts with other driving forces. Another point worth mentioning about that work is that the interactions among the students are face-to-face instead of using technology to communicate among them. These kinds of interactions are frequent in closed settings such as hospitals, offices and particularly in our case schools. A considerable effort has been done to understand the mechanisms that give rise to face-to-face networks structures [2–4]. The aim of this study is to explore social mechanisms that might explain friendship formation in Yucatan's elementary schools recorded by Huerta et al. [1]. We will focus on homophily, transitivity and preferential attachment. In doing so, this research contributes to deepening and widening knowledge about these empirical children's friendship network. To achieve this aim, we will test the next hypotheses, by using ERGMs:

1.1 Hypotheses

Hypothesis (H1): Students of the same gender are more likely to form friendship links than students of different genders. Hypothesis (H2): Students of the same grade are more likely to form friendship links than students of different grades. Hypothesis (H3): Our friendship networks reflects a tendency for edges to concentrate on popular students. Hypothesis (H4): There are significant transitivity structure effects in our friendship networks. We consider H1 because there are studies showing that during childhood and early adolescence interaction takes place predominantly in groups of the same gender [5, 6]. We consider H2 because the spatial confinement was a phenomenon taken into account into the models of [1]. We consider H3 because studies report a tendency for popular individuals to disproportionately attract more connections than less popular individuals [7, 8]. This is formally described by theory as preferential attachment [9, 10]. We consider H4 because friendship networks (as well as other social networks) have a strong tendency towards transitivity levels higher than those of random networks with comparable size [11].

2 ERGM

We used Exponential Random Graph models (ERGMs) to test our hypotheses. ERGMs is a framework for studying complex networks, and has the capacity to represent the effects of observables measured in empirical networks. With these models we can estimate the effects of observables on the characteristics of real networks, through models that seek to parsimoniously describe the social forces that give shape to a network. The preceding makes it possible to formulate theoretical models and learn about the properties of empirical networks, allowing to address the complex dependencies within relational data. The previous advantages have motivated the use of ERGM models in various disciplines like Social Sciences [12, 13], Physics [14], Biology [15, 16], Information Science [17], Statistics [18]. With applications ranging from construction of consumer recommendation systems [19], disease prevention and control [20], agricultural research [21], conservation projects [22] and many others [23–26]. For many real world networks

we only have one copy or instance: there is only one internet, and only one world wide web. It is essential to ask, could we analyze how it would be a process of diffusion information on the internet with a slightly (or completely) different structure from the real internet? Many times, we would like to know how a process behaves in social networks in a general way, instead of just knowing how it behaves in the network that we have measured [27]. Therefore, it is convenient to have a set of possible networks (ensemble) with a probability distribution that represents a specific network, so that we can calculate or estimate the effects of the process of interest. The methods that use an ensemble idea are useful, not only in the field of networks and in statistical physics, but also in others areas such as machine learning, since there are methods which combine multiple learners (or models) to produce the expected output [28]. In this context, statistical mechanics allow us to define an ensemble of random graphs with expected values equal to those of the network of interest. It can be demonstrated, by maximizing Gibbs entropy, that the probability distribution of a network g is given by:

$$P(g) = \frac{e^{H(g)}}{Z}. \quad (1)$$

Where Z is the partition function, and H is the Hamiltonian, defined as:

$$H(g) = \sum_{i=1} \theta_i x_i(g). \quad (2)$$

Here θ_i is the coefficient associated to the observable x_i in the network of interest. Calculating the partition function Z in Eq. 1 is inconvenient: the number of possible graphs with n nodes scale as $2^{\binom{n}{2}}$. Thus calculating Z represents a severe difficulty in practice, so Z is approximated by generating a large number of random graphs. To generate the random graphs, we used the MCMC (Markov Chain Monte Carlo). The idea is to use our real network as the sample of a distribution in which $P(g)$ has been defined, but without calculating the partition function Z . A network can be considered in the form of the adjacency matrix so that each position a_{ij} is a random variable. By adding or removing a link to the network, we can change the value of these variables to 0 or 1. Repeating this process, we generate a sequence of graphs, each one dependent only on the previous one.

2.1 Choice of Observables

Newman and Park [29] studied observables known as two-stars and triangles. Using the mean field approximation, they found two physically possible solutions for the same parameter value, so it is said that we have a degenerate state [30]. In practice, degeneration is a problem that leads to bimodal distributions between extreme edge densities. To avoid degeneration we include two types of weighted distributions, GWESP (geometrically weighted edgewise shared partner) which can be defined as the Weighted sum of the number of connected nodes having

exactly i shared partners weighted by the geometric sequence $(1 - e^{-\alpha})^i$ where α is a decay parameter.

$$v = e^\alpha \sum_{i=1}^{n-2} [1 - (1 - e^{-\alpha})^i] EP_i(y). \quad (3)$$

Where $EP_i(y)$ is the distribution that show how many connected dyads have one or more shared partners. Another term is the GWD (Geometrically Weighted degree) which is simply a weighted sum of the counts of each degree i weighted by the geometric sequence $(1 - e^{-\alpha})^i$ where α is a decay parameter [31].

$$u = e^\alpha \sum_{i=1}^{n-1} [1 - (1 - e^{-\alpha})^i] D_i(y). \quad (4)$$

Where $D_i(y)$ represents the number of nodes in the network y with *degree* = i .

3 Building Models

For each school, four models were built. We begin with a simple Bernoulli model and then add variables. As we add variables we assess model improvement and determine what to keep in the model.

Model 1 (Edges) consists of a random network where each possible link in the network has the same probability of being or not. Usually, this assumption is unrealistic for empirical social networks, but it serves as a reference for comparison with more elaborate models.

Model 2 (Edge+attributes) takes into account the effects of the attributes of each node to determine if there are effects of homophily. One reason why two people could form a friendship link is by sharing similar activities or similar tastes. It is important to emphasize or highlight that results obtained through an ERGM such as the Model 2 (i.e., one that only takes into account links and attributes of the nodes) would be the same as if we had used a conventional logistic regression analysis.

Model 3 (Edges+attributes+transitivity) is based on Model 2 plus transitivity. Real world networks usually have more transitivity than random networks. That is, members of the network that are connected to each other tend to be connected to common members. To consider the transitivity in the model, we proceed to include the geometrically weighted edgewise-shared partners (GWESP), this is a structural term. Models without structural terms tends to overestimate the effects that are given by the common features of the students (homophily), whereas a model that contains only structural terms would estimate the unique effects of these factors.

Model 4 (Edge+attributes+GWESP+GWD) is based on Model 3 plus the term geometrically weighted degree (GWD) which is just a weighted degree distribution. So it will consider the endogenous plus the exogenous part given by the attributes of the nodes. As a result, estimates of the homophily effects or the

structural effects could be smaller in Model 4 than in a model in which only one of the two separate parts (either exogenous or endogenous) were considered. Model 4 also includes the term GWESP for transitivity, and the term geometrically weighted degree (GWD). The term GWD is an anti-preferential attachment term [32–34]. We expect Model 4 will fit the observed network better than models containing only homophily effects, since at least some structural effects, such as transitivity and popularity, are important social processes that influence the formation of friendship links. The standard Monte Carlo error and the p-value are used for selecting which observable to keep. Finally, we proceed to make a comparison of Models 4 among all these schools and interpretation each one.

4 Estimation

The estimation of parameters of the ERGM can be performed using the maximum likelihood principle: given an observed network x , θ is estimated with the value that maximizes the log-likelihood of the model:

$$l(\boldsymbol{\theta}, \mathbf{x}) \equiv \sum_{i=1} \theta_i x_i(g) - \ln(Z). \quad (5)$$

Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) is the preferred method to estimate the maximum likelihood fit to ERGMs. An alternative method is called maximum pseudolikelihood estimation (MPLE). MPLE is substantially less computationally difficult than MCMCMLE. However, properties of the MPLE estimators are not well understood, and the estimates tend to be less accurate than those of MCMCMLE. From the existing research it is not clear when the estimates of pseudo-likelihood can be acceptable. So we decided to use the maximum likelihood estimation of the Monte Carlo Markov (MCMCMLE) chain in this study. The MCMCMLE method seems to be the standard approach in literature [31,35]. We used the software packages statnet [36] and ergm [35] under the statistical computational environment R to fit our ERGMs.

5 Results

With the results shown in Tables 1, 2 and 3 we can spot individual node characteristics that strongly influence the network formation process. From Model 2, we find clear evidence of grade and gender homophily (positive and statistically significant coefficients) on each school.

From Model 3 we note that, on each school, the coefficients related to the effects of homophily (grade and gender) have lower estimates with respect to Model 2. Such a change is a direct consequence of adding the term GWESP. This means that not all link formation in the network are explained by grade and gender homophily: some of the links among students with similar attributes (gender and school grade) are driven by a tendency to form links with others

Table 1. Estimated parameters for the models applied to school 1.

Model	Observable	Estimate	Std. Err.	AIC
Model 1	Edges	-2.350***	0.0466	3419
Model 2	Edges	-3.3627***	0.094	3037
	Nodematch(grade)	1.838***	0.099	
	Nodematch(sex)	0.838***	0.102	
Model 3	Edges	-4.816***	0.124	2766
	Nodematch(grade)	1.285***	0.071	
	Nodematch(sex)	0.650 ***	0.078	
	Gwesp(0.8)	0.836***	0.054	
Model 4	Edges	-5.291***	0.172	2725
	Nodematch(grade)	1.092***	0.064	
	Nodematch(sex)	0.572***	0.065	
	GWESP(0.8)	1.023***	0.069	
	GWD(0.8)	5.126***	1.176	
	esp(2)	-0.456***	0.105	

Four models are run on school 1. Each term has a parameter estimate, a standard error and a p-value indicated by asterisks. The last column shows the AIC for each model.

*** Significant at $p < 0.001$.

Table 2. Estimated parameters for the models applied to school 2.

Model	Observable	Estimate	Std. Err.	AIC
Model 1	Edges	-3.2113***	0.0325	8338
Model 2	Edges	-4.174***	0.064	7232
	Nodematch(grade)	2.319***	0.067	
	Nodematch(gender)	0.611***	0.069	
Model 3	Edges	-5.049***	0.068	6646
	Nodematch(grade)	1.496***	0.050	
	Nodematch(gender)	0.496***	0.054	
	GWESP(0.8)	0.791***	0.032	
Model 4	Edges	-5.79***	0.1015	6514
	Nodematch(grade)	1.244***	0.045	
	Nodematch(gender)	0.464***	0.052	
	GWESP(1.0)	0.925***	0.035	
	GWD(1.0)	5.086***	0.611	

Four models are run on school 2. Each term has a parameter estimate, a standard error and a p-value indicated by asterisks. The last column shows the AIC for each model. *** Significant at $p < 0.001$.

Table 3. Estimated parameters for the models applied to school 3.

Model	Observable	Estimate	Std. Err.	AIC
Model 1	Edges	-4.000***	0.02543	15781
Model 2	Edges	-6.181***	0.071	10457
	Nodematch(grade)	4.037***	0.064	
	Nodematch(gender)	1.036***	0.059	
Model 3	Edges	-6.200***	0.062	10103
	Nodematch(grade)	2.902***	0.080	
	Nodematch(gender)	0.850***	0.049	
	GWESP(0.8)	0.561***	0.030	
Model 4	Edges	-5.86***	0.107	9997
	Nodematch(grade)	2.226***	0.107	
	Nodematch(gender)	0.780***	0.046	
	GWESP(1.0)	0.571***	0.0306	
	GWD(1.0)	1.770***	0.3153	
	esp(2)	-0.206***	0.0556	
	absdiff	-0.0321***	0.0046	

Four models are run on school 3. Each term has a parameter estimate, a standard error and a p-value indicated by asterisks. The last column shows the AIC for each model. *** Significant at $p < 0.001$.

students having mutual friends. A different analysis of the ERGM as linear or similar regression would not have detected this distinction between these two driving forces. On the last column of the tables, we showed the AIC value. The AIC approach is a common way to choose ERGMs that adequately describes the experimental data. Such an approach selects the set of metrics that produce the estimated distribution most likely to have resulted in the observed data with a penalty for additional metrics to ensure parsimony. Our last and more complex model, Model 4, yield the smallest AIC factor on each school as we move from Model 1 to Model 4. This AIC decreasing behavior indicates an improved adjustment as we add more (statistically significant) terms, those on hypothesis H1, H2 and H4. One thing to keep in mind is that the strength of the different network patterns is not comparable because the scale varies for each observable. The results are then interpreted as follows: For Model 4 applied to school 1 (see Table 1) all coefficient are statistically significant ($p\text{-value} < 0.001$) and all positive but the coefficient of `edges`. The negative value of the edges coefficient indicates network sparsity (i.e. a low edges density), a common feature of real social networks. Positive coefficients evidenciate several tendencies. The first one is the tendency of gender homophily (i.e. ties are more likely between students sharing the same gender), supporting H1, in agreement with previous studies showing a tendency among students to form friendship links based on gender. The second tendency is scholar grade homophily

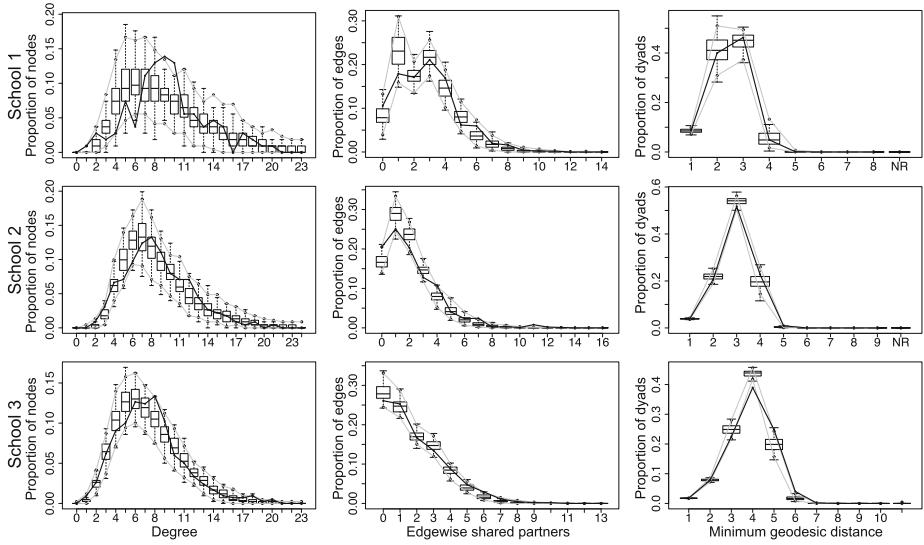


Fig. 1. Goodness of fit (GOF). Degree, edgewise shared partners and geodesic distance distributions for the Model 4 applied to the 3 elementary schools. The results obtained in our real friendship networks are represented by the solid black line; Box plots include the median and the interquartile range; and light gray lines represent the range in which 95% of the simulated observations fall.

(i.e. ties are more likely between students sharing the same grade), supporting H2. The third tendency is the transitivity (positive GWESP coefficient), supporting H4, meaning that some of the links among students, even with similar attributes gender and school grade, are driven by a tendency to form links with other students having mutual friends. The fourth tendency is that in all three schools, GWD estimates were significantly positive, indicating a relatively homogeneous degree distribution. In addition this is consistent with the degree distributions shown in Fig. 1. This is opposite the preferential attachment mechanism that we proposed in H3; it means that popular students should be less attractive to create new friendship links in urban and rural areas of our study. Thus, H3 was not supported. For Model 4 applied to school 2 (see Table 2) and to school 3 (see Table 3) all coefficient are statistically significant ($p\text{-value} < 0.001$) and all positive but the coefficient of edges. In both schools, we observe the same tendencies as in school 1. Additionally, for schools 1 and 3 the coefficient of esp(2) is significant and negative, indicating that edges with exactly two shared partners are less frequent than expected by chance. A more meaningful way to assess fit is to examine how well the model reproduce important properties in our observed networks. This may be done through the goodness of fit plots (GOF), which is shown in Fig. 1. The first plot in each row shows the degree distribution, the number of connections among friends. The second plot in each row represents the distribution of shared partners, the number of friends in common

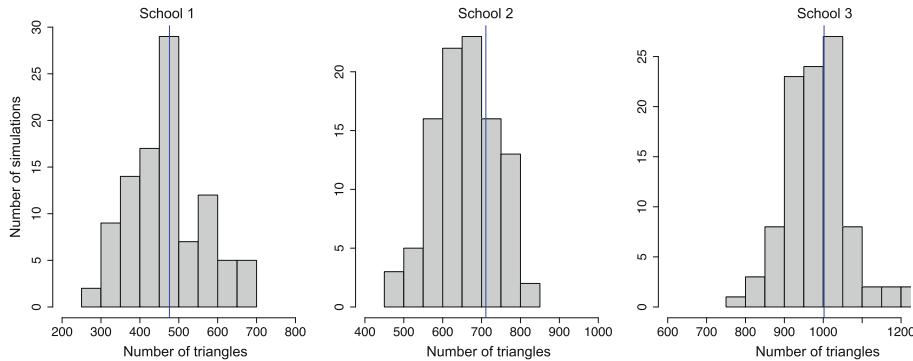


Fig. 2. Triangle distribution on Full Models. Histograms on triangles distribution per network in 100 simulations with Model 4. The blue vertical line shows observed number of triangles in the three elementary schools.

among linked nodes. The third plot shows the distribution of geodesic distances. Model 4 does a good job capturing the degree distribution, the shared partner distribution and the geodesic distance in urban and rural areas. For school 3, Model 4 has an additional term capturing absolute difference in grade, aimed to improve the geodesic distance fit. Figure 2 shows the histograms of variation in the number of triangles for networks simulated with Model 4. There are 476, 711 and 1002 triangles in the schools 1, 2 and 3 respectively. Simulated networks has triangle count modes close to these values. This is yet another indicative of Model 4 viability to describe our friendship networks.

6 Discussion and Conclusion

This work contribute to prior studies on friendship formation. Our findings revealed that boys and girls tends to create friendship ties with other students with the same gender. This is not necessarily true for all friendship networks on other stages of life. During adolescence and the beginning of romantic relationships, girls shows an earlier evolution in their attitude toward other gender mates than boys [37]. Additionally, this study revealed that children were more likely to have friends within their classroom than other classrooms. People tend to form relationships when they are in close proximity. Other studies about friendship networks focused on university students. Those have shown that spatial confinement is not a restriction when forming friendship with students on a different grade. Even could be possible that the probability of forming a friendship link in the same class were lower than the probability of forming a friendship link on another class [38]. A plausible explanation is that students at this stage are not constrained by spatial restrictions: they can use social media to meet other people, join other people with similar academic interest or similar academic achievements, outside their classroom. As expected, positive triad closure effects were found in all schools. Then, our networks are highly clustered. We did not

find support for the hypothesis H4 about preferential attachment. This is in contrast with some friendship networks, since these networks usually have a center-periphery structure [39] with popular students in the center. There are some limitations to this research that suggest the need to be cautious about drawing conclusions from this work: First, not all regions of Mexico were included in our data, just some rural and urban small areas from Yucatan. Therefore we make no claims about the mechanism found as statistically significant in friendship network formation to hold in all elementary school of Mexico. Second, the ERGM framework has some practical difficulties, the most important one is degeneracy. Finally, using empirical data to test hypotheses about friendship formation could improve our understanding of dynamical process, such as disease transmission. In particular, we could investigate for a future research if networks generated by ERGMs that match local structures of our friendship networks, exhibited similar dynamics.

References

1. Huerta-Quintanilla, R., Canto-Lugo, E., Viga-de Alva, D.: Modeling social network topologies in elementary schools. *PLoS ONE* **8**(2), e55371 (2013)
2. Goodreau, S.M., Kitts, J.A., Morris, M.: Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* **46**(1), 103–125 (2009)
3. Flores, M.A.R., Papadopoulos, F.: Similarity forces and recurrent components in human face-to-face interaction networks. *Phys. Rev. Lett.* **121**(25), 258301 (2018)
4. Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.F., Van den Broeck, W.: What's in a crowd? Analysis of face-to-face behavioral networks. *J. Theor. Biol.* **271**(1), 166–180 (2011)
5. Baerveldt, C., Van de Bunt, G.G., Vermande, M.M.: Selection patterns, gender and friendship aim in classroom networks. *Zeitschrift für Erziehungswissenschaft*. **17**(5), 171–188 (2014)
6. Oldenburg, B., Van Duijn, M., Veenstra, R.: Defending one's friends, not one's enemies: a social network analysis of children's defending, friendship, and dislike relationships using XPNet. *PLoS ONE* **13**(5), e0194323 (2018)
7. Jiao, C., Wang, T., Liu, J., Wu, H., Cui, F., Peng, X.: Using Exponential Random Graph Models to analyze the character of peer relationship networks and their effects on the subjective well-being of adolescents. *Front. Psychol.* **8**, 583 (2017)
8. Wax, A., DeChurch, L.A., Contractor, N.S.: Self-organizing into winning teams: understanding the mechanisms that drive successful collaborations. *Small Group Res.* **48**(6), 665–718 (2017)
9. Jeong, H., Néda, Z., Barabási, A.L.: Measuring preferential attachment in evolving networks. *EPL (Europhys. Lett.)* **61**(4), 567 (2003)
10. Newman, M.E.: Clustering and preferential attachment in growing networks. *Phys. Rev. E* **64**(2), 025102 (2001)
11. Bianconi, G., Darst, R.K., Iacovacci, J., Fortunato, S.: Triadic closure as a basic generating mechanism of communities in complex networks. *Phys. Rev. E* **90**(4), 042806 (2014)
12. Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph (p^*) models for social networks. *Soc. Netw.* **29**(2), 173–191 (2007)

13. Conway, D.: Modeling network evolution using graph motifs. arXiv preprint [arXiv:1105.0902](https://arxiv.org/abs/1105.0902) (2011)
14. Desmarais, B.A., Cranmer, S.J.: Statistical mechanics of networks: estimation and uncertainty. *Phys. A* **391**(4), 1865–1876 (2012)
15. Saul, Z.M., Filkov, V.: Exploring biological network structure using exponential random graph models. *Bioinformatics* **23**(19), 2604–2611 (2007)
16. Yletyinen, J., Bodin, Ö., Weigel, B., Nordström, M.C., Bonsdorff, E., Blenckner, T.: Regime shifts in marine communities: a complex systems perspective on food web dynamics. *Proc. Roy. Soc. B Biol. Sci.* **2016**(283), 20152569 (1825)
17. Yang, D.H., Yu, G.: Static analysis and exponential random graph modelling for micro-blog network. *J. Inf. Sci.* **40**(1), 3–14 (2014)
18. Krivitsky, P.N.: Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. *Comput. Stat. Data Anal.* **107**, 149–161 (2017)
19. Alexandridis, G., Siolas, G., Stafllopatis, A.: Enhancing social collaborative filtering through the application of non-negative matrix factorization and exponential random graph models. *Data Min. Knowl. Disc.* **31**(4), 1031–1059 (2017)
20. Kukielka, E.A., Martínez-López, B., Beltrán-Alcrudo, D.: Modeling the live-pig trade network in Georgia: implications for disease prevention and control. *PLoS ONE* **12**(6), e0178904 (2017)
21. Hermans, F., Sartas, M., Van Schagen, B., van Asten, P., Schut, M.: Social network analysis of multi-stakeholder platforms in agricultural research for development: opportunities and constraints for innovation and scaling. *PLoS ONE* **12**(2), e0169634 (2017)
22. Nita, A., Rozylowicz, L., Manolache, S., Ciocănea, C.M., Miu, I.V., Popescu, V.D.: Collaboration networks in applied conservation projects across Europe. *PLoS ONE* **11**(10), e0164503 (2016)
23. De La Haye, K., Dijkstra, J.K., Lubbers, M.J., Van Rijsewijk, L., Stolk, R.: The dual role of friendship and antipathy relations in the marginalization of overweight children in their peer networks: the TRAILS Study. *PLoS ONE* **12**(6), e0178130 (2017)
24. Cherepnalkoski, D., Karpf, A., Mozetič, I., Grčar, M.: Cohesion and coalition formation in the European Parliament: roll-call votes and Twitter activities. *PLoS ONE* **11**(11), e0166586 (2016)
25. Salehi, S., Holmes, N., Wieman, C.: Exploring bias in mechanical engineering students' perceptions of classmates. *PLoS ONE* **14**(3), e0212477 (2019)
26. Campbell, B.W., Marrs, F.W., Böhmler, T., Fosdick, B.K., Cranmer, S.J.: Latent influence networks in global environmental politics. *PLoS ONE* **14**(3), e0213284 (2019)
27. Newman, M.: Networks. Oxford University Press, Oxford (2018)
28. Dietterich, T.G.: Ensemble methods in machine learning. In: International Workshop on Multiple Classifier Systems, pp. 1–15. Springer (2000)
29. Park, J., Newman, M.E.: Solution of the two-star model of a network. *Phys. Rev. E* **70**(6), 066146 (2004)
30. Park, J., Newman, M.: Solution for the properties of a clustered network. *Phys. Rev. E* **72**(2), 026136 (2005)
31. Snijders, T.A., Pattison, P.E., Robins, G.L., Handcock, M.S.: New specifications for exponential random graph models. *Sociol. Methodol.* **36**(1), 99–153 (2006)
32. Hunter, D.R., Handcock, M.S.: Inference in curved exponential family models for networks. *J. Comput. Graph. Stat.* **15**(3), 565–583 (2006)

33. Angst, M., Hirschi, C.: Network dynamics in natural resource governance: a case study of Swiss landscape management. *Policy Stud. J.* **45**(2), 315–336 (2017)
34. Levy, M.A., Lubell, M.N.: Innovation, cooperation, and the structure of three regional sustainable agriculture networks in California. *Reg. Environ. Change* **18**(4), 1235–1246 (2018)
35. Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M.: ergm: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* **24**(3), nihpa54860 (2008)
36. Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., Morris, M.: statnet: Software tools for the representation, visualization, analysis and simulation of network data. *J. Stat. Softw.* **24**(1), 1548 (2008)
37. Stehlé, J., Charbonnier, F., Picard, T., Cattuto, C., Barrat, A.: Gender homophily from spatial behavior in a primary school: a sociometric study. *Soc. Netw.* **35**(4), 604–613 (2013)
38. Hernández-Hernández, A.M., Viga-de Alva, D., Huerta-Quintanilla, R., Canto-Lugo, E., Laviada-Molina, H., Molina-Segui, F.: Friendship concept and community network structure among elementary school and university students. *PLoS ONE* **11**(10), e0164886 (2016)
39. Rombach, M.P., Porter, M.A., Fowler, J.H., Mucha, P.J.: Core-periphery structure in networks. *SIAM J. Appl. Math.* **74**(1), 167–190 (2014)



Impact of Natural and Social Events on Mobile Call Data Records – An Estonian Case Study

Hendrik Hiir¹, Rajesh Sharma¹(✉), Anto Aasa², and Erki Saluveer³

¹ Institute of Computer Science, University of Tartu, Tartu, Estonia
{hendrik.hiir,rajesh.sharma}@ut.ee

² Department of Geography, University of Tartu, Tartu, Estonia
anto.aasa@ut.ee

³ OÜ Positium, Tartu, Estonia
erki.saluveer@positium.com

Abstract. Mobile Call Data Records (CDR) can be used for identifying human behavior. For example, researchers have studied mobile CDR to understand the social fabric of a country or for predicting the human mobility patterns. Additionally, CDR data has been combined with external data, for example, financial data to understand socio-economic patterns. In this paper, we study an anonymised CDR dataset provided by one of the biggest mobile operators in Estonia with two objectives. First, we explore the data to identify and interpret social network patterns. Our study points that mobile calling network is fragmented and sparse in Estonia. Second, we study the impact of natural and social events on mobile call activity. Our results show that these activities do have an impact on the calling activity. To the best of our knowledge, this is the first study, which has analysed the impact of varied types of events on mobile calling activity specifically in Estonian landscape.

Keywords: Mobile Call Data Records · Social network analysis · Sociocultural analysis

1 Introduction

Call Data Records (CDR) are information collected by mobile network operators during the course of service in order to assemble billing data. Call Data Records contain the timestamp, calling party and called party in pseudonymised form, call type, duration and coordinates of corresponding parties for localization. CDR have been analysed from many different perspectives. One of the main technique which has been used for CDR is social network analysis [1]. For example, in [2] authors reported that the square of the distance between two individuals is inversely proportional to their connectedness. In a different study [3], the CDR was analysed from the perspective of strong and weak ties in the mobile network. Authors in [4] analysed the sociocultural aspects of a city.

CDR have been used for identifying the nature of human mobility [5–7] and population distribution [8].

In addition, CDR are often combined with other datasets for research purpose. For example, in [9], authors combined the mobile records with GIS for predicting human mobility and in another study conducted in Norway [10] researchers used mobile phone data in combination with the railway infrastructure and train traffic data to predict the number of train travelers. In a different work, researchers combined the CDR with the bank data to identify different economic status groups [11].

In this work, we analysed anonymised CDR data provided by one of the biggest Estonian mobile operator. Our analysis consists of two dimensions. Firstly, we performed a descriptive analysis using the following activities.

1. We first report results of our analysis of the CDR dataset by modeling it as a social network.
2. Next, we analysed the data at the county level by aggregating the calls between counties. We analyse this dataset to identify self-centered counties and discuss the reasons behind them.
3. We also look into call activity distribution over time by analysing the activity by days, workday hours and weekend hours.

Secondly, we examine the impact of various events on calling activity. To the best of our knowledge, this is the first study, which has studied the impact of such a variety of events on calling activity. We categories these events into the following two categories.

1. **Natural events:** This part of our analysis is somewhat similar to [12] but the authors examined only weather conditions, however, we have also considered the impact of full moon on call activity.
2. **Social event:** We analyse the effects of a football match as an example of social event to understand its impact on calling activity.

The outcome of our analysis shows that mobile social network based on CDR is very sparse and fragmented and spread out in general. Also, time of the day as well of the days of the week and various events do have an impact on the call activity.

The rest of the paper is organized as follows. Next, we discuss related works. Section 3 presents results of our descriptive analysis of the dataset and in Sect. 4 we study the impact of various events on calling activity. We conclude with a discussion of future directions in Sect. 5.

2 Related Work

Mobile Call data records (CDR) has attracted a lot of research in the last decade. CDR data when analysed using social network analysis, can provide valuable information about the social fabric of the society [4,13], human behaviour analysis to find movement patterns by age groups [14], to identify strong or weak ties between individuals [3], population distribution [8] and flu prediction [15].

However, CDR data when combined with other information such as financial and GPS information can produce lateral information. For example, when combined with economic data it can help in identifying spending patterns [16], or socio-economic classes [11]. CDR has also found its application for urban planning when combined with GPS data [17]. In a different application, researchers used mobile phone data with railways traffic for predicting the number of travellers [10].

A good amount of research using CDR data has been done for identifying mobility patterns. For example, [5] and [6] authors demonstrate that humans path is predictable and reproducible. In another work authors proposed a human mobility model and validated using real dataset from New York and Los Angeles metropolitan areas [7].

In addition, CDR data has also been used in tourism as well. In particular, [18] researchers study the travel behaviour in several cities such as Boston, San Francisco, Lisbon. In another study foreign tourists' favourite places in Estonia was investigated in [19]. CDR can also help in narrow down the suspects by detecting suspicious activity [20,21].

In this work, we analyse CDR data provided by one of the largest mobile operators in Estonia to understand the effect of various events on people's calling activity. This work is different from [12] as only the impact of weather was analysed in that. However, in this work, we study the impact of time period and different types of event such as natural and social events on the calling activity.

3 Descriptive Analysis

In this section, we first describe the dataset, next, we discuss the results of our descriptive analysis, which includes modeling the CDR as social networks, and comparing calling activity over different time intervals, that is analysing the activity by days, workday hours and weekend hours.

3.1 Dataset Description

The anonymised dataset of calling activity was provided by a leading mobile operator in Estonia. The dataset includes 10% of caller IDs who made calls in Estonia during the period of 1st March 2015 to 31st March 2015. In total, the dataset included the records of 722,724 calls. Initial analysis pointed to a number ID which received a significant higher number of incoming calls compared to rest of the other numbers. Specifically, this number received 94,008 incoming calls (with no outgoing calls) compared to the second largest receiver which received only 584 calls. Suspecting it to be a call center number of the mobile operator we removed it from the dataset.

3.2 Interpreting Network Measures

We measure various network metrics of CDR network by modeling the call activity among the individuals as a graph, where the nodes represents the set of call IDs in our dataset and edges represent the call activity between two individuals. For simplicity, we considered the network as unweighted and undirected.

Using the above mentioned graph model, we measure various network metrics of CDR network. Table 1 provides information about the results of these metrics. The total number of callers in our dataset are 130,093 (Row 1) which have interacted at least 137,809 times (Row 2). The average degree is 2.11, which can be used for inferring that on an average each person only calls just two other persons. Although the network is highly disconnected, with 10,176 disconnected components (Row 4), however, 73% individuals are present in the biggest connected component. Edge density (Row 6) and clustering coefficient (Row 7) indicates that sharing of common contacts is rare in this network which makes the network very sparse. The sparseness of the network is furthermore confirmed with the high number of communities (Row 10). The value of the diameter (Row 8, value 39) indicates that the network is highly spread out, although the average path length is lower (Row 9) but still indicative of relatively high spreadness in the network. In summary, we found the network to be highly sparse and highly spread out.

Table 1. Properties of used dataset

Row	Metrics	Values
1	Nodes (Callers)	130,093
2	Edges (Interactions)	137,809
3	Average degree	2.11
4	Disconnected components	10,176
5	Nodes in the biggest component	95,182
6	Edge density	1.63e-05
7	Clustering coefficient	0.036
8	Diameter	39
9	Average path length	13.37
10	Communities	10,344

Table 2. Abbreviations and populations of the counties

Abbreviation	County	Population
Ha	Harju	575,601
Ta	Tartu	151,377
Id-V	Ida-Viru	147,597
Pä	Pärnu	82,349
Lä-V	Lääne-Viru	59,039
Vi	Viljandi	47,010
Ra	Rapla	34,436
Võ	Võru	33,172
Sa	Saare	31,706
Jõ	Jõgeva	30,841
Jä	Järva	30,109
Va	Valga	29,944
Põ	Põlva	27,438
Lä	Lääne	24,070
Hi	Hiiu	8,582

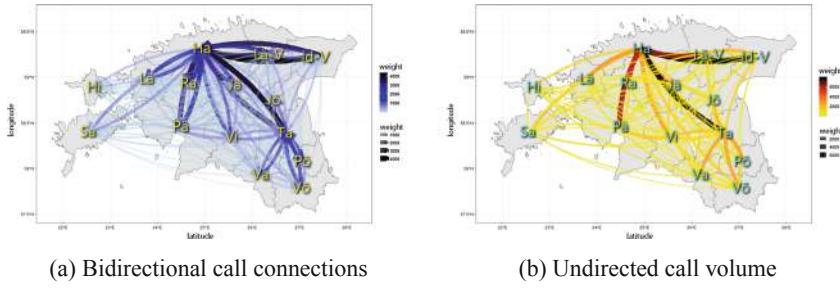
3.3 Call Activity Description over Counties

In Estonia there are 15 counties, with Harju county, which includes capital Tallinn being the most populous and Hiiu county being the least populous. The following subsection analyses connection between the counties. The Table 2 gives an overview of abbreviations for each county used on network maps and their populations as of January 1st 2015 in descending order.

Bidirectional Call Connections: The Fig. 1a shows bidirectional connections between each of Estonian counties. In 10 out of 14 possible cases (as in-area calls were excluded), Harju county is the most popular destination for other counties outside the county's borders. In-area calls are the calls where both nodes are in the same county. For Harju county itself, the most popular destination is Ida-Viru county. It is notable that both counties have the biggest percentage of Russian community in Estonia, which is the likely reason for this bidirectional relation between these two areas. In 4 out of 14 possible cases, the second most populous Tartu county is the most popular destination outside the area's border. This is the case for the areas of Jõgeva, Valga, Põlva and Võru. All of them are either neighboring areas of Tartu (Jõgeva, Valga and Põlva) or having Tartu as the significant nearby center and working area (Võru).

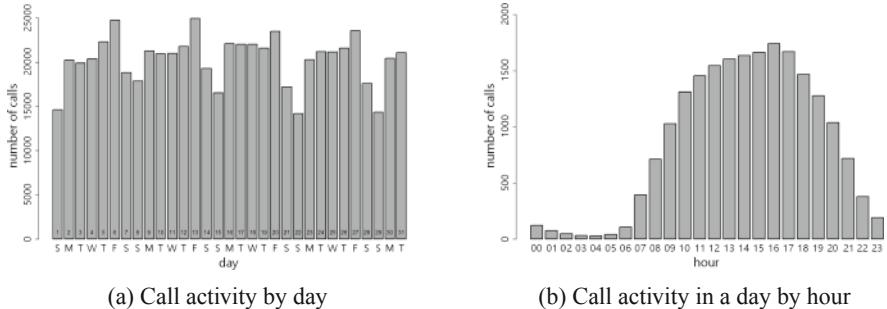
Hiiu county is the least popular destination for 7 areas (all of them being in eastern part of Estonia) while Põlva county is the least popular destination for 4 areas (all of them being on the western part of Estonia). Another area that was the least popular in more than one case is Saare county, which received the least amount of calls from the areas of Jõgeva and Põlva. The islands of Hiumaa and Saaremaa (Hiiu county and Saare county) have unique least popular destinations: Võru county and Jõgeva county correspondingly. Therefore, the islands do not follow the trend of other western counties that have Põlva county as the least popular destination. Although the margins were not big, it is one of the ways to demonstrate how counties' communities differ.

Undirected Call Volume: The Fig. 1b shows the volume between each of Estonian counties. Top 8 most frequent connections include Harju as one of the nodes. The most frequent connection is between Harju and Ida-Viru counties with 7807 calls (8.22% of all calls excluding in-area calls), mostly because of the large Russian community in both of these counties. The second most frequent was between two most populous counties: Tartu and Harju with 7581 calls (7.99% of all calls excluding in-area calls). The least frequent connection with just 7 calls (0.007% of all calls excluding in-area calls) was between the least populous Hiiu county and Russians-dominated Ida-Viru county. As the connections between the counties were bidirectionally similar like the Fig. 1a demonstrates, with some small exceptions such as more calls from Tartu county to Harju county than vice versa, despite Harju having 3.81 times bigger population, undirected call volume (demonstrated in Fig. 1b) could be a good measure to analyse network between the counties.

**Fig. 1.** Call activity

3.4 Calling Activity over Time

Call Activity by Day: The Fig. 2a describes the call activity by days over the period of March 1st 2015 to March 31st 2015. It can be seen that the Friday is always the most active day of the week with the average of 24189.25 calls, being 19.27% above average activity. The reason for this is that for the majority of the people it is the end of the working period of Monday to Friday and therefore a favorable day to make appointments with people. From Monday to Thursday there is generally similar amount of calls made each day. Weekend days are always less active than working days due to significantly smaller amount of work-related calls. Sunday is the least active day of a week with 15522.2 calls on an average, which is 23.46% below the average activity. The reason for this is that traditionally it is the day of rest in most countries, including Estonia.

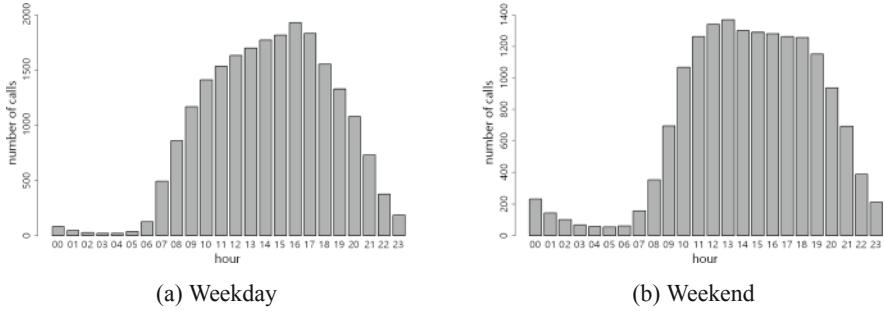
**Fig. 2.** Call activity in a day

Call Activity by Hour: The Fig. 2b shows the average number of calls in a day by per hour. For this figure, working days and weekend days are merged together in order to visualize average hourly activity over all the days. It can be seen that activity between the period of 04:00 to 16:59 is constantly increasing and activity between the period of 16:00 to 04:59 is constantly decreasing.

The most active hour when taking all days into consideration is 16:00 to 16:59 which comes mostly from the period of Monday to Friday when work is commonly finished and people are available to call. The least active period is between 01:00 and 06:59 when on an average 53.94 calls are made in an hour compared to daily hourly average of 845.05 calls per hour (15.67 times below average).

Call Activity by Hour During Working Days: The Fig. 3a shows the average number of calls in an hour during working days (Monday to Friday) and the average number of calls in an hour is 905.69. The activity distribution is similar to the one with all days considered (displayed in Fig. 2b) due 22 out of 31 observed days (70.97%) being working days. Working days are also more active than weekend days. The difference that can be noticed is that the hour of 16:00 to 16:59 is even more active compared to other hours and the preceding period from 10:00 more gradual. The increasing and decreasing call activity periods are the same, that is from 04:00 to 16:59 and 16:00 to 03:59 respectively. Another notable difference is less calls during the night period. While the merged data has 53.94 calls made on an average during the period of 01:00 to 06:59, the average number of calls for the same period is 43.45 (21.53% difference), which is 0.20% of all the calls during the day.

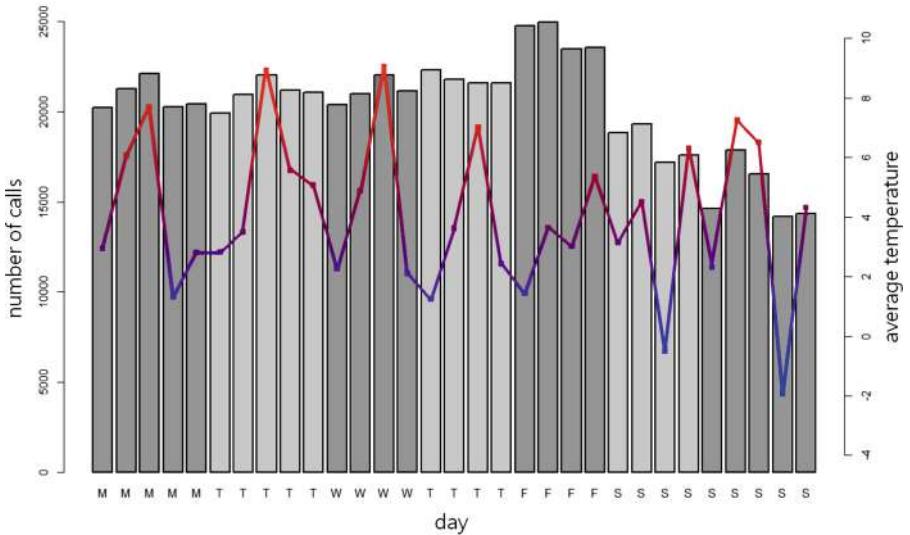
Call Activity by Hour During Weekend Days: The Fig. 3b shows the average number of calls in an hour for only weekends (Saturday and Sunday), where the average number of calls in an hour is 696.81. The main differences compared to working days' hourly calls is that there are more calls during the night, and more even activity distribution between 10:00 and 19:59 and peak time of 12:00 to 13:59 instead of 16:00 to 17:59 compared to weekdays. The main reason for this distribution difference is that most of the people are available during the whole day on weekends and therefore it's possible to make appointments with other people during the day time as well. The average number of calls during the period of 01:00 to 06:59 is 79.57, which is 0.48% of all calls during the days. This means that nightly activity is 2.38 times higher during the weekend than during the working days. This can be explained with the fact that more people are making plans during the night due to fact that next day being free of duties. Another difference compared to working days is in terms of call activity increasing and decreasing periods. In contrast to 04:00 to 16:59 on weekdays, call activity increasing periods during the weekend are 05:00 to 13:59 and 23:00 to 00:59 and decreasing periods are from 00:00 to 05:59 and 13:00 to 23:59.

**Fig. 3.** Call activity in a day by hour

4 Impact of Events on Call Activity

4.1 Impact of Natural Events on Call Activity

Call Activity vs Temperature: The Fig. 4 shows daily call activity in comparison with the average temperature. The days are sorted in the order they appear in the week. For average temperature, we used the average temperature of four Estonian most populous cities (Tallinn, Tartu, Narva, Pärnu) at hours 00:00, 06:00, 12:00 and 18:00 over the period of March 1st 2015 to March 31st 2015, which is used for temperature axis values of the figure. The average temperature was 4.02 °C.

**Fig. 4.** Call activity vs Temperature

It can be seen from the figure that the coldest day of the month (-1.94°C) was also the least active of the month. Although the day was Sunday (14188 calls compared to 15523 in average, 8.60% below average), which is the least active day of the week, it was the least active out of all Sundays. The preceding Saturday, which was the second coldest day of the month (-0.5°C), was also the least active Saturday (17198 calls compared to 18225.25 on an average, 5.64% below average) of the month. It is also notable that both days were one of the six days when it was snowing during that month in at least some of observed locations.

Also, it was noted that the warmest day of the month (9.06°C) was the most active Wednesday of the month (22053 calls compared to 21152.75 in average, 4.36% above average).

Lunar Effect on Call Activity: We also studied the impact of full moon on calling activity. The Fig. 5 shows the number of calls for the observed period during the nights between 22:00 and 06:00. The period of 22:00 to 06:00 is chosen because it was the interval when the moon was visible and people generally go to bed. During the month, full moon occurred on the nights of 5–6 and 6–7 [22]. For the night of 5–6, full moon was in a form of Micromoon and also the furthest and smallest of the year [23]. The following night of 6–7 was a continuation of the full moon. Although the night of 4–5 was also considered to have a full moon in some areas of Earth [24], it was not 100% visible and still growing according to Estonian data. The night of 5–6 when the Micromoon occurred, was between Thursday and Friday and 751 calls were registered. The average number of calls between Thursday and Friday night was 729.25, which means 2.91% increase in the call activity. The following night of 6–7 also had a full moon, and was

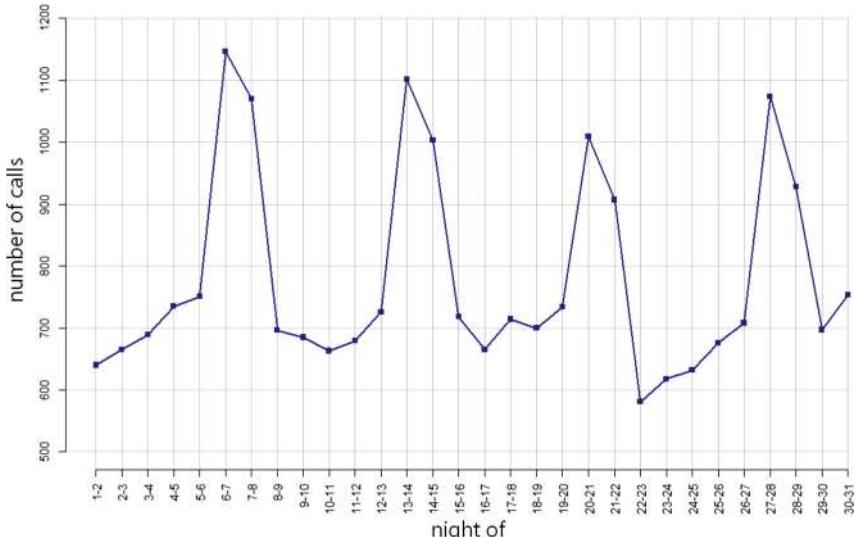


Fig. 5. Lunar effect on Call activity. Number of calls during the nights between 22:00 and 06:00

the most active night of the month with 1146 calls registered in comparison to 1082.75 that was average between the nights of Friday and Saturday. This means 5.84% increase in activity. It is notable that although the night of 4–5 had moon still in the growing phase, it was still the most active night of month between Wednesday and Thursday with 5.38% increase in the call activity. Based on the activity displayed in the figure, we found that full moon does make people relatively more active during the night.

4.2 Impact of a Social Event

During the March 31st there was Estonia-Iceland football match and the Fig. 6 describes the number of calls inside and around the A.Le.Coq arena where the match took place. There were no other matches during the observed period in that location. A.Le.Coq arena is the arena where international matches featuring Estonian national football team are played. The gray bar shows the total number of calls for each day. The blue bar shows how many of these calls were made when only the period of 17:00 to 22:00 is considered: the period from gates opening until 75 min after the match is finished during the match days. There were 5334 spectators during the match. The used dataset included a total of 573 calls from the observed period around the arena giving an average of 18.48 calls per day. During the match day, there were 34 calls, 1.84 times more than the average. For the period of 17:00 to 22:00, the average number calls is 6.65. During the match time, there were 19 calls, which is 2.86 times above the average number of calls for that period. 42.11% of these calls were made before the match, 31.58%

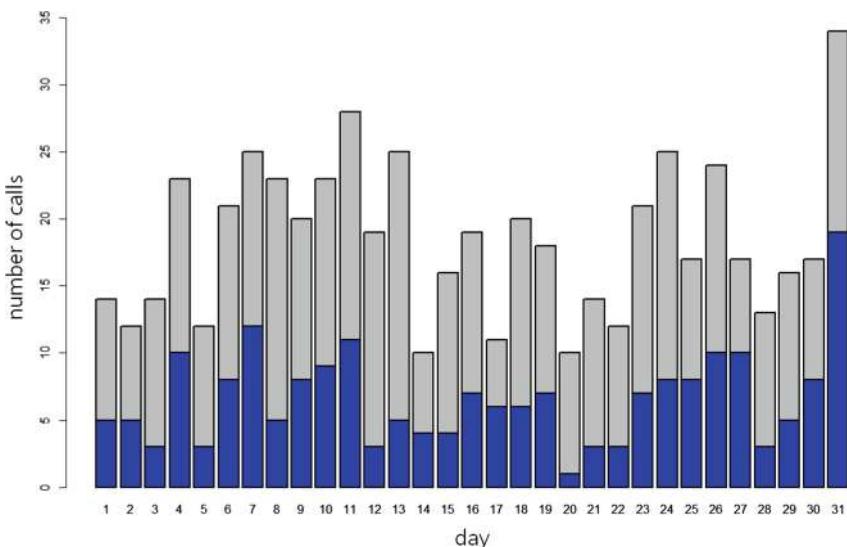


Fig. 6. Number of calls around the main football arena in Estonia during the March 2015. Blue shows the number of calls during the typical match related period (17:00 to 22:00)

of calls were made after the match and the remaining 26.32% of calls were made either during the very beginning, very end of match or during the break. In conclusion, people were focused on the match and avoided making calls during the match time.

5 Conclusion and Future Work

With quest to understand the social fabric of Estonia as well as to investigate if various events can impact the calling activity, we analysed a large call data records provided by one of the biggest mobile operators in Estonia. We analysed the data from two broad dimensions. Firstly, we performed a descriptive analysis using social network analysis and calling patterns in different time periods. Secondly, we studied the impact of external events on calling activity. In particular, we analysed the impact of natural and social events. The results indicate that human calling activity do depends on the time period of calling and it also gets impacted by events happening around it.

We have multiple future directions for this work. We would like to perform our analysis on a larger dataset which spans a longer period of time and consisting of more individuals to understand the impact of other social, cultural events. We would also like to combine the mobile CDR data with other events such as bank related data to understand the socio-economic patterns in Estonia.

Acknowledgments. This work has been supported in part by EU H2020 project SoBigData and Estonian Research Council project *Understanding the Vicious Circles of Segregation. A Geographic Perspective* (PUT PRG306). We are also thankful to Estonian mobile operator for providing us the data.

References

1. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
2. Lambiotte, R., Blondel, V., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., Van Dooren, P.: Geographical dispersal of mobile communication networks. *Phys. A* **387**, 5317–5325 (2008)
3. Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.-L.: Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7332–7336 (2007)
4. Ponieman, N.B., Sarraute, C., Minnoni, M., Travizano, M., Rodriguez Zivic, P., Salles, A.: Mobility and sociocultural events in mobile phone data records. *AI Commun.* **29**, 77–86 (2015)
5. Song, C., Zehui, Q., Blumm, N., Barabási, A.-L.: Limits of predictability in human mobility. *Science* **327**(5968), 1018–1021 (2010)
6. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.-L.: Understanding individual human mobility patterns. *Nature* **453**(7196), 779–782 (2008)
7. Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., Willinger, W.: Human mobility modeling at metropolitan scales. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, MobiSys 2012, pp. 239–252. ACM, New York (2012)

8. Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J.: Dynamic population mapping using mobile phone data. *Proc. Natl. Acad. Sci.* **111**(45), 15888–15893 (2014)
9. Williams, N.E., Thomas, T.A., Dunbar, M., Eagle, N., Dobra, A.: Measures of human mobility using mobile phone records enhanced with GIS data. *CoRR*, abs/1408.5420 (2014)
10. Sørensen, A.Ø., Bjelland, J., Bull-Berg, H., Landmark, A.D., Akhtar, M.M., Olson, N.O.: Use of mobile phone data for analysis of number of train travellers. *J. Rail Transp. Plan. Manage.* **8**(2), 123–144 (2018)
11. Leo, Y., Karsai, M., Sarraute, C., Fleury, E.: Correlations of consumption patterns in social-economic networks. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, pp. 493–500 (2016)
12. Horanont, T., Phithakkitnukoon, S., Leong, T.W., Sekimoto, Y., Shibasaki, R.: Weather effects on the patterns of people's everyday activities: a study using GPS traces of mobile phone users. *PLoS ONE* **8**, e81153 (2013)
13. Eagle, N., Pentland, A.(Sandy), Lazer, D.: Mobile phone data for inferring social network structure. In: Liu, H., Salerno, J.J., Young, M.J. (eds.) *Social Computing, Behavioral Modeling, and Prediction*, pp. 79–88. Springer, Boston (2008)
14. Ahas, R., Mark, Ü.: Location based services—new challenges for planning and public administration? *Futures* **37**(6), 547–561 (2005)
15. Farrahi, K., Emonet, R., Cebrian, M.: Predicting a community's flu dynamics with mobile phone data. In: Computer-Supported Cooperative Work and Social Computing, Vancouver, Canada, March 2015
16. Singh, V.K., Freeman, L., Lepri, B., Pentland, A.: Predicting spending behavior using socio-mobile features. In: International Conference on Social Computing, SocialCom, Washington, DC, USA, pp. 174–179 (2013)
17. Ratti, C., Frenchman, D., Pulselli, R.M., Williams, S.: Mobile landscapes: using location data from cell phones for urban analysis. *Environ. Plan.* **33**(5), 727–748 (2006)
18. Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C.: The path most traveled: travel demand estimation using big data resources. *Transp. Res. Part C Emerg. Technol.* **58**, 162–177 (2015). Big Data in Transportation and Traffic Engineering
19. Ahas, R., Aasa, A., Mark, Ü., Pae, T., Kull, A.: Seasonal tourism spaces in Estonia: case study with mobile positioning data. *Tour. Manag.* **28**, 898–910 (2007)
20. Kumar, M., Hanumanthappa, M., Kumar, T.V.S.: Crime investigation and criminal network analysis using archive call detail records. In: 2016 Eighth International Conference on Advanced Computing (ICoAC), pp. 46–50, January 2017
21. Khan, E.S., Azmi, H., Ansari, F., Dhalvelkar, S.: Simple implementation of criminal investigation using call data records (CDRs) through big data technology. In: 2018 International Conference on Smart City and Emerging Technology (ICS CET), pp. 1–5, January 2018
22. Kuufaaside kalender. marts 2015. <https://ilm.pri.ee/kuufaaside-kalender?month=3&year=2015>
23. Tana öösel näeme taiskuud (2015). <https://ilm.ee/?513460>
24. Moon phases, March 2015. https://www.calendar-12.com/moon_calendar/2015/march



Describing Alt-Right Communities and Their Discourse on Twitter During the 2018 US Mid-term Elections

Ángel Panizo-LLedot^{1(✉)}, Javier Torregrosa², Gema Bello-Orgaz³, Joshua Thorburn⁴, and David Camacho³

¹ Computer Science Department, Universidad Autónoma de Madrid, Madrid, Spain
angel.panizo@uam.es

² Biological and Health Psychology Department, Universidad Autónoma de Madrid, Madrid, Spain
francisco.torregrosa@uam.es

³ Departamento de Sistemas Informáticos, Universidad Politécnica de Madrid, Madrid, Spain

{gema.borgaz,david.camacho}@upm.es

⁴ RMIT University, Melbourne, VIC, Australia
joshthorburn18@gmail.com

Abstract. The alt-right is a far-right movement that has uniquely developed on social media, before becoming prominent in the 2016 United States presidential elections. However, very little research exists about their discourse and organization online. This study aimed to analyze how a sample of alt-right supporters organized themselves in the week before and after the 2018 midterm elections in the US, along with which topics they most frequently discussed. Using community finding and topic extraction algorithms, results indicated that the sample commonly used racist language and anti-immigration themes, criticised mainstream media and advocated for alternative media sources, whilst also engaging in discussion of current news stories. A subsection of alt-right supporters were found to focus heavily on white supremacist themes. Furthermore, small groups of alt-right supporters discussed anime, technology and religion. These results supported previous results from studies investigating the discourse of alt-right supporters.

Keywords: Polarization · Far-right · Community finding · Topic extraction · Political extremism

1 Introduction

In recent years, far-right political parties and candidates have achieved increased electoral success across various Western states. Furthermore, terrorist attacks by far-right extremists, such as in Norway and New Zealand, demonstrate the risk that this extremism poses. Although it has a long history, the far-right has flourished in recent years, with the effective use of social media strongly contributing

to this rise. While the alt-right itself is “highly decentralized”, with no official leaders, ideology or political party, this broad far-right amalgamation has effectively used social media to attract followers and influence political discourse. Indeed, the alt-right played a key supporting role in the election of US President Donald Trump. Alt-right supporters have used a variety of social media networks to spread content and engage in political discussion, including Twitter, Facebook, Reddit, 4Chan and YouTube. Notably, in response to social media companies increasingly removing hate speech and extreme far-right accounts, alt-right supporters have recently congregated on less-restrictive websites, such as Gab, an alternative to Twitter.

While the discourse of other white extremist groups has been previously studied by academics [6], there are few articles concerning how alt-right sympathizers communicate and interact online. However, the alt-right is a difficult group to study due to the movement’s fragmentation and lack of ideological agreement. With these limitations in mind, the present research aims to analyze how a set of alt-right supporters communicated on Twitter during the 2018 US midterm elections, a politically relevant period. To evaluate this, two questions were posed. First, *“Are there different subsections of alt-right followers?”*, and second, *“What are the topics discussed in these groups?”*.

2 Related Work

This section first describes the alt-right discourse and its similarities and dissimilarities with other ultra-conservative groups, before reviewing common techniques for doing online discourse analysis.

2.1 Alt-Right Discourse

Firstly, opposition to immigration, particularly from the Middle East, Africa or Latin American countries, is a central political stance of alt-right supporters and the far-right. For example, the refugee crisis stemming from the Syrian Civil War led to far-right groups using hashtags to reject European countries accepting these immigrants, [10], such as #Ausländernraus (“Foreigners out!”), #EuropeforEuropeans or #refugeesnotwelcome. Furthermore, alt-right followers have been supportive of Trump’s strict immigration policy proposals [7]. Lyons [11] found that original manifestos from prominent ideological thinkers of the alt-right were deeply rooted in a rejection of non-white immigration. Similarly, alt-right supporters commonly use racist language to describe immigrants, much like other far right groups [8]. Indeed, the alt-right has popularised the use of white supremacist hashtags such as “#ItsOkToBeWhite, or #WhiteGenocide [16]. Supporters of the alt-right frequently attack minorities and are especially hostile towards black people, Jews and Muslims. Consequently, while they are often aligned with Trump, alt-right followers have criticized him for being too soft against Saudi Arabia and Israel because they symbolically represent Islam and Judaism [12].

Criticism of mainstream media is also common in alt-right discourse [11], with this often mirroring President Trump's frequent lamentations about "fake news". A study conducted by Forscher and Kteily [5] found that alt-right supporters tend to have high levels of suspicion towards mainstream media sources (i.e., New York Times, CNN, etc.) and strong trust in alternative media sources, such as the conspiratorial InfoWars. Conspiracy theories are also commonly believed alt-right followers. For example, it is commonly believed that certain media companies are actively trying to bring down the Trump presidency, or that a secretive Jewish elite hold enormous power [4].

Alt-right supporters tend to be very politically active on various online social networks. Through the analysis of user descriptions from the online platform Gab (an online platform similar to Twitter) it was found that most of its far right users included references to Trump or his campaign slogans, conservative topics, religion or America [20]. Hashtags against Islam, about the alt-right itself were also found on the website. This analysis of Gab found that there posts by users contained 2.4 times more hate words than on Twitter, but less than half than those on 4chan. 4Chan itself, especially its /pol/ ('Politically Incorrect') message board, has been strongly linked with providing a fertile breeding ground for the alt-right to develop as a movement [13].

Although many of the ideological underpinnings of the alt-right are shared with and stem from earlier and contemporary far-right movements, the alt-right has differed in some respects too. For example, although some alt-right supporters are fervently religious, they often differ strongly from the Evangelical conservative Christian right in the US. Indeed, "cuck"/"cuckservative", the frequently used alt-right pejorative, derives from a sexual act. This slang is also representative of another characteristic of the alt-right: its ability to use humour to attract supporters. By using ironic, dark-humoured satire, the alt-right has succeeded in flippantly spreading white supremacist ideology on various online platforms [9]. Further, the alt-right has used internet memes, as well as creating its own imagery and slogans to attract supporters.

2.2 Techniques for Online Discourse Analysis

A number of technical approaches can be used to analyse online communications. Depending on the focus of the research (e.g. text, interactions, frequency of writing, etc.), certain methods are more appropriate. Natural Language Processing (NLP), which includes sentiment analysis, topic extraction or linguistic inquiry, is a common approach in analysing large textual datasets. Examples of this includes a study by Torregrosa et al. [17] where the tweets of jihadi supporters on Twitter was examined using Linguistic Inquiry Word Count (LIWC), or in a study by Tumasjan et al. [18] where sentiment analysis was used to analyze tweets referring to political parties.

Understanding how people and groups interact on Online Social Networks (OSNs) is important when analyzing discourses. Social Network Analysis (SNA) techniques are especially useful to study these interactions. While some SNAs are focused on extracting common characteristics of a network, such as the density or

its diameter, others focus on characterizing properties of the users of the network. For example, analyzing the Homophily of the network or finding cohesive groups of users (communities) can be an effective approach. ‘Community finding’ is one of the most commonly used techniques to investigate and analyze OSNs.

Several studies have combined community finding with NLP techniques. Among them, three main approaches can be found in the literature. The most frequent approach combines NLP techniques and community finding into a singular process. In this approach, communities are not only consistent in a structural manner, but also in the topics they discuss. Feller et al. [3] used this type of method when applying a network analysis to contextualize the users political preferences during the 2016 Germany elections. The second approach applies both process simultaneously but separated, comparing the outcomes later. An example of this approach can be found in Surian et al. [15], where they analyze the discussions about HPV vaccines on Twitter. Finally, the third approach initially applies the community finding process, and then applies NLP techniques in each community independently. This was used by Yin et al. [19] in a study comparing two datasets, with one containing computer science journals and the other composed of tweets that included the keywords “Obama” or “media”. In the present research, this third approach was applied to examine the tweets and the communities.

3 Methodology

3.1 Data Description

The dataset used in this study contained 52903 tweets published by 123 alt-right Twitter accounts collected between the 30th of October and the 13th of November, 2018. This period corresponds to the weeks before and after the 2018 US midterm elections. In addition to the text of each tweet, information such as publication date, retweets or hashtags was gathered. The selected alt-right Twitter accounts were chosen from a dataset created by Thorburn, Torregrosa and Panizo [16]. The alt-right users published between 2 and 2590 tweets during the data collection period. In fact, $\approx 60\%$ of the accounts tweeted every day of this period and $\approx 80\%$ of the users in the dataset tweeted on at least 10 different days.

As shown in Fig. 1(a), the number of tweets published per day was between 2656 and 4539 tweets, with increased activity surrounding election day (November 6, marked as a red vertical line), with this peaking on November 7, the day immediately after the elections. However, in regards to the number of users posting each day (shown in Fig. 1(b)), it can be noticed that the same spike in activity is not present. Finally, regarding the daily distribution of tweets published by each account (see Fig. 1(c)), the median and the whiskers of the box blot diagram indicate that this also increased surrounding election day.

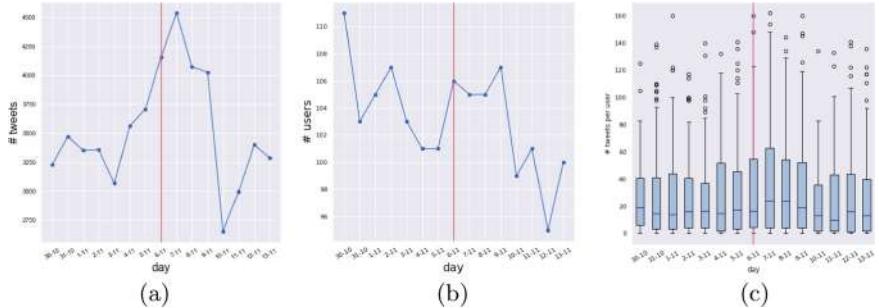


Fig. 1. Subfigure (a) shows the number of tweets published each day. The x-axis contains the days and the y-axis the number of tweets published. Subfigure (b) shows the number of accounts that published something every day of the data collection period. The x-axis contains the days and the y-axis the number of accounts that tweeted on each day. Subfigure (c) shows the distribution of the number of tweets published by an account for each day. The x-axis contains the days and the y-axis the number of tweets published. The red line marks election day (November 6).

3.2 Protocol

Once the data was gathered to create the dataset, this information was divided and analyzed using community finding and topic detection algorithms. The phases of this process is detailed as follows:

- Data Extraction:** The data collected to perform the analysis of the alt-right discourse was extracted from the Twitter API. Every tweet published by the users in the selected two time periods was included for analysis. The extracted tweets are stored in a database (MongoDB) together with their related information, such as creation date, location, source, retweet count, etc.
- Data Preprocessing:** Two re-tweet networks of users were created, representing the interactions between all users before and after the election day. To create each network, users were considered as the network nodes, and their relationships represented the edges. The relationships were established using the re-tweets.
- Community Detection of Alt-Right Users:** The detection of relevant groups of users within the OSN can be addressed by the application of *community detection algorithms* in the re-tweet networks. Before applying the community search algorithm, a filter was used to get the largest connected component of each re-tweet network. Then, the most similar users were grouped together using the Clauset-Newman-Moore greedy modularity maximization method [2, 14].
- Topic Extraction by the Communities:** Once the different communities were obtained for each re-tweet network, a topic modeling technique was applied to identify the topics covered by a collection of texts. With this

Table 1. Categories used to classify the topics covered in the alt-right communities.

Categories	Example keywords	Example hashtags	Code
Racial discourse	White, Racist, Black, Hate, Discrimination, Diversity, Race, Israel, Semitic, Jewish, Jerusalem, Islam, Iran, Pakistan	#Iran, #WhiteGenocide, #ItsOkToBeWhite	RD
Politics	Trump, Obama, Hillary, President, Right, Left, Party, Republican, Democrats, Maga, Antifa	#Maga, #Democrats, #AmericaFirst	P
Immigration	Border, Immigration, Caravan, Migrants, Citizenship, Wall, Asylum, Wave, Illegal, Invasion, Refugees	#BirthrightCitizenship, #WalkAway, #StandUpForEurope	I
Media	Media, News, Fake, Gab, Alternative media, Press, Hivemind, Propagandistic, CNN, Conspiracy, Journalist	#Killstream, #WSJKillsKids, #CNN	M
Elections	Election, Vote, Campaign, Senate, Mid-term, Rally, Ballots	#Midterms2018, #VoteRed, #ElectionDay	E
Policy debate	Money, Taxes, Taxation, Marxism, Socialism, Communist, Freedom, Nationalism, Brexit	#TaxationIsTheft	PD

aim, the Latent Dirichlet Allocation (LDA) model [1] was fitted and applied for each community, using only the text of the tweets published by the users belonging to the particular community. In addition, the most mentioned hashtags for each community were extracted.

5. **Analysis of the Social Discourse and Evolution of the Communities:** The analysis of the discourse of each alt-right community on Twitter was conducted using the evolution of its relevant topics. For this purpose, topics were manually categorized according to the general themes that were found to be relevant in the discourse of the alt-right groups in the related work section. A summary of the categories can be found in Table 1.

4 Results

Table 2 summarizes the communities extracted from the sample before the election, along with information relating to the topic extraction process, the most mentioned hashtags and the categories in which the topics were included. Nearly all the communities contained politically focused discourse, including references to the elections (#ElectionDay), right-wing political slogans (“Make America Great Again”, #MAGA) or other relevant political groups (#Anonymous).

Most of the groups include words associated with far-right ideologies and the alt-right more specifically. Racist discourse was identified in many of the communities. This can be subcategorized into white supremacy (communities 1, 3, 4, 10 and 12), antisemitism (communities 1 and 10) and Islamophobia (community 10). Group 12 was found to be especially extreme in its language, using hashtags such as #WhiteGenocide and #WhiteLivesMatter, which are associated with the alt-right and white supremacism.

The references to the media were divided between pro-comments (comments defending alternative media) and anti-comments (comments against traditional media). In the first subcategory, there are references to media related with far-right movements (#Killstream, #GAB, #RedNationRising). There are also references to Gab (some users presented a link to their Gab profile on their Twitter descriptions). The second subcategory includes criticisms against traditional media, with this including hashtag campaigns against CNN (“#CNNCredibilityCrisis) and the Wall Street Journal (#WSJKillsKids). Criticism of traditional media outlets was especially notable in community 14, where CNN was mentioned several times.

As expected, immigrations was frequently mentioned, and this can be divided between references to illegal immigration from Mexico (e.g. communities 8 and 11) or the refugee crisis in Europe (e.g. communities 5, 10 and 11). The words “Trump” and “Border” appear together in some groups (e.g. community 8), which indicates that users commonly discussed the Trump administration’s immigration policies.

Interestingly, three groups seem to talk about isolated topics. Community 6 is the only one that focused on Brexit, taxes and political ideologies (Communism, Socialism, Neoliberalism, etc.). Furthermore, communities 9 and 15 centred around discussion of anime and religion respectively. These outliers illustrate some of the distinctions and demographics that the alt-right is composed of, which is elaborated on further in the discussion section.

Concerning the week after the election, Table 3 shows changes in some of the communities. Firstly, there was an increase in tweets about the election itself (communities 1, 2, 3, 8, 9, 12 and 13), with the word “election” and the #ElectionDay being used more. Notably, #Broward featured prominently, with this hashtag referring to difficulties in vote counting in the county of Broward, Florida. Various stories from right-wing media and political figures alleged that there had been a conspiracy to rig the election in favour of Democrats in this county. This alleged conspiracy was especially prominent in communities 2, 3 and 12, where the #StopTheSteal was frequently used.

While no community maintained their original structure from the previous week, some of them conserved their topics and their most relevant users, as seen on Fig. 2. This was the case in community 6, which featured political debate in both weeks; community 7, which was dedicated to discussing anime (like community 9 from previous week); and community 10, which was the most extreme community in the second week too, using exactly the same hashtags that community 12 had used previously.

Table 2. Communities detected before the Election Day and the general topics on which their users have posted messages.

Id.	N.U.	Top 3 Topics	Top 3 Hashtags	Labels
1	909	people white new want america got racist better hate nyt anti gop twitter election black semitic	Killstream MAGA WSJKillsKids	RD(we, as) P E
2	745	trump president obama media campaign trump left border caravan people democrat gillum breaking party florida	MAGA Midterms2018 VoteRed	E P M (anti)
3	531	really trump true girl love good man bad today voting women gt hate american twitter	WSJKillsKids Killstream HappyHalloween	E M (anti) RD (w, b)
4	294	vote democrats red voting trump campaign wow maga know migrants trump president obama democrats america	MAGA BLEXIT VoteRed	E RD (we) P
5	274	vote trump president america florida democrats family brownley senate women california shawn company insurance stopped program	Iran RedNationRising MAGA	E M (pro) I
6	205	good time woman high child read sense world better vegans people socialism hate free marxist	TaxationIsTheft DonLemon mises	PD
7	177	people online love media god gab speech free hate left world getting think read gt	VerifiedHate Update Gab	M (pro)
8	160	just think know trump middle trump media job usa hit immigration wall country border asylum	EnemyOfThePeople wages manufacturing	I M
9	119	starting rain evidently mustang sins halloween anime game minmod mini eat bad hey bro crunch	jojo_anime azurlane NoTatsukiNoTanoshii	Anime
10	115	trump obama white caravan migrant trump president caravan migrant midterm democrats iran midterms sanctions kavanaugh	TheTruthCommunity FaroeIslands Anonymous	I E RD(ai)
11	99	stop country home triangle lemon trump illegal beto cnn caravan texas border invasion campaign southern illegal	Halloween2018 HappyHalloween2018 MAGA	I E
12	67	white hatred sick wshis girls diversity itsokaytobewhite defund whitepeople word code migrants refugees europe eu border	WhiteGenocide StandUpForEurope WhiteLivesMatter	RD (we) I
13	64	memes hillary tweet trump black timeline country economy train book votereditosaveamerica hivemind critical marxist theory	MAGA2018 BTFOMobRule memewarmidterms	E P
14	63	trump campaign candidate gillum navy roff voting sundaymorning mondaymotivation thewalkingdead cnn vote obama gillum rally	CNNCredibilityCrisis ElectionDay ElectionDay2018	E M (anti)
15	4	american stones demons boy bad church gt christ right blessed strates protestant fascism authoritarian mary collectivist	CavalierNationalism WhiteHonorKillings	Religion

Most of the topics identified in the week before the election were also found in the second week. White supremacist (communities 2, 3, 8 or 10) and antisemitic (communities 9 and 17) discourse was maintained, but there was no community

Table 3. Communities detected after the Election Day and the general topics on which their users have posted messages.

Id.	N.U.	Top 3 Topics	Top 3 Hashtags	Labels
1	784	election county tucker broward antifa white america black live press democrat gillum breaking party florida	MAGA Editorial Broward	E I
2	758	county ballots election votes florida trump acosta white states jim year boom years women stop	StopTheSteal ElectionNight Broward	E RD (we) P
3	571	trump president just vote got broward ballots county florida fraud snipes white people election democrats steal	StopTheSteal BrowardCounty MAGA	E M (anti) RD (w, b)
4	256	time day men fam long people good don like white right new fuck bad community	WoolseyFire electionnight ff	RD (we)
5	215	streaming company home muslim think white pride thread beagle deadly sargon really youtube just people	Killstream WSJKillsKids BREAKING	M (pro, anti)
6	178	socialism marxist nationalism party british men group make things word western christian asiabibi asylum british	MeToo AsiaBibi JimAcosta	PD
7	157	wow evangelion taught looks whoah real needs energy avi world time undertale aha know kinda	DELТАRUNE	Anime
8	149	white women vote yes god trump election republican ballots vote broward county tucker twitter scott	ElectionDay Florida Broward	E P RD (we)
9	144	followers help million raise communist watch just media trump kid nationalism broward hitler ballots county	StopTheSteal TaxationIsTheft Nationalism	E RD (as)
10	104	white people problem asylum facebook migrants defund discrimination whites sense gab migrant media free right facebook	WhiteGenocide StandUpForEurope WhiteLivesMatter	RD (we) I M(pro)
11	100	border entry ports story italy white fact german ohio language trump european imagine language jewish	2A FoxNews France	I
12	96	trump stop american democrat abortion macron women michelle football saudi florida election county broward fraud	StopTheSteal USMC MarineCorpsBirthday	E Abortion
13	95	sounds fuck shit cultural florida trump great non cruz stop got wonder family black white	VeteransDay2018 Nature MAHwld5	E
14	85	marine love saved children television god bless noticed lady today press jim acosta stophesteal trump	Iran Veteran MAGA	Military M (anti)
15	63	retired proud crypto jeff major president ai stage crypto blockchain tonight broward skating sponsorship paccoin	Blockchain AI WashingtonElite	Technology
16	28	news talking fake weird fbpe british little ffs great sure just white good make ve video	TwoMinuteSilence RemembranceDay2018 ArmisticeDay100	M (pro)
17	10	antifa pay polish fuck jewish wanking media deficit social migrant trump voters punish jobs black jerusalem	MeToo PolishIndependenceDayMarch SouthAfrica	RD (as)

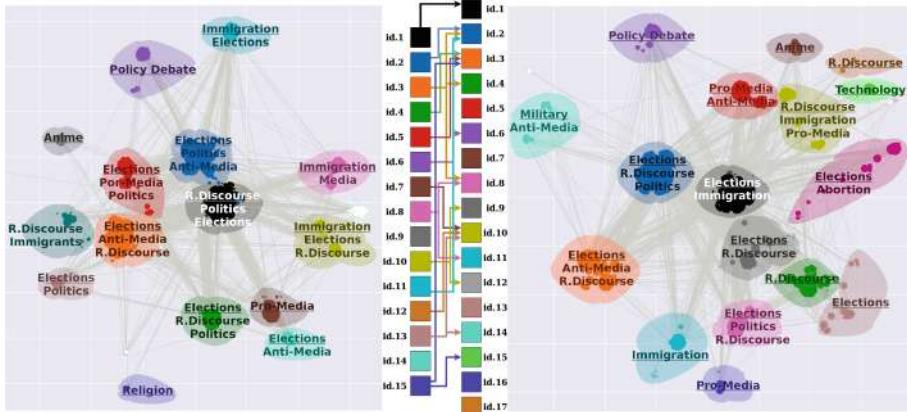


Fig. 2. On the left, communities detected before the election day. On the right, communities detected after the election day. Each community is labeled with their topics category, the underscored text represents main topic categories. An arrow joining two community on the legend indicates the transference of a community between snapshots.

that included a clear reference to Islam in the two weeks that our data covers. Still, discussion of immigration featured in communities 1, 10 and 11. Again, there were comments against traditional media (communities 3, 5, 14 and 16), including with references to “fake news” (Community 16). There were also tweets supporting alternative media (Community 5, 10 and 16). Interestingly, Gab and the hashtag #KillStream (this hashtag is for a far-right podcast of the same name) are mentioned again. Lastly, several communities mention politics in their discussions (e.g. Community 8).

Finally, there were three new categories in the week following the election. First, community 14 contained references to war and military forces, mentioning the hashtags #Veteran and #Iran. The second was community 15, which was focused on debates about technology in general (including references to blockchain, crypto currency, AI, etc.). Lastly, debate about abortion was mentioned by community 12.

5 Discussion and Future Work

This study aimed to analyze the communications of a group of alt-right supporters in the context of an important political event, in this case, the 2018 US midterm elections. Firstly, we demonstrated that there was increased Twitter activity in the days immediately surrounding the election. The outcomes of this study also demonstrates that sub-communities from this database were mostly unstable over the data collection period, but that many topics did maintain over this period.

Regarding the communities, it was found that none maintained the same structure of users over both weeks. Some of the relevant users remained in the

same communities from one week to another (Table 2), but they did not represent more than the 20–40% on average of the previous relevant users. Therefore, it was easier to interpret communities using the topic they discussed, than by the users that were included. If we compare the topics covered by the communities and their categorization, we can find that some of the communities did maintain over time. A perfect example would be the community using white supremacist hashtags (community 12 pre-elections, 10 post-elections), or community 6, which discussed the economy and general politics in both weeks. Therefore, while communities of users did change during the data collection period, similar topics were discussed in many of these groups.

Indeed, the topics that our data identifies were commonly discussed reflected earlier research on the alt-right. For example, anti-immigrant sentiment [7,10,11] and racially derogatory comments [8] were common. Furthermore, hashtags such as #WhiteGenocide and #WhiteLivesMatter were used frequently in some communities, and these hashtags are strongly associated with white supremacy. As to be expected, discussion of current political news stories was frequent, with this frequently mirroring right-wing media talking points. Similarly, criticism of traditional news media (such as CNN and The Wall Street Journal) was detected, with this often mirroring President Trump's denunciations of the press. In contrast, "alternative" media sources were frequently shared or discussed, such as from Kill Stream and Red Nation Rising, with this reflecting other research on the alt-right [11].

Interestingly, our results indicate that anime and technology were discussed by many alt-right supporters. These findings provide support to other literature which has linked internet subcultures to the growth of the alt-right. For example, in Kill All Normies, Angela Nagle navigates the link between #GamerGate, 4Chan and even sections of anime fans with the rise of the alt-right. Similarly, in Mike Wendling's book on the alt-right, he agrees with this assessment and also connects proponents of technologies like cryptocurrencies with the movement. The results of this study provides further evidence of this link between the alt-right and certain sections of anime fans and cryptocurrency proponents.

The results of this study also indicate that there is a small section of alt-right supporters who commonly discuss religion. While alt-right supporters are typically hostile towards Judaism and Islam, the movement cannot be broadly characterised as Christian, as a large contingent appear to be irreligious [8]. Still, results here identified a small community of users where religion was discussed before the election and another where abortion was a central topic after the midterm elections. This is perhaps evidence of the small, paleoconservative element which is aligned with the alt-right, where opposition to abortion is a strongly held belief [8].

Considering the similarity on the topics, the increased relevance of elections on the second week and the increasing on the tweeting rate until the day of the elections, we can suggest that the Alt-Right communicational activity grows when a political event is near. However, a study picking longer dates or making a deeper assessment on the daily evolution of the communities could help to

confirm this hypothesis. Also, it could be useful to determine why communities are quite different, how and when do they break into different groups, and why the topics covered by this communities are maintained.

Overall, this research demonstrated that alt-right supporters were more active in the days immediately surrounding the 2018 US midterm elections. Furthermore, this study identified a number of sub-communities within the broader alt-right, that gathered around certain topics. Similar to previous literature on the alt-right, anti-immigrant sentiment, racism towards Jews and Muslims, as well as support for white supremacy, were all frequently reflected in the results. Also supporting earlier literature, our results identified smaller, yet still notable segments of the alt-right sample used in this study discussing anime, technology and religion.

Acknowledgements. This work has been supported by several research grants: Spanish Ministry of Science and Education under TIN2014-56494-C4-4-P grant (DeepBio) and Comunidad Autónoma de Madrid under S2013/ICE-3095 grant (CYNAMON).

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003)
2. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. *Phys. Rev. E* **70**(6), 066111 (2004)
3. Feller, A., Kuhnert, M., Sprenger, T.O., Welpe, I.M.: Divided they tweet: the network structure of political microbloggers and discussion topics. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
4. Finkelstein, J., Zannettou, S., Bradlyn, B., Blackburn, J.: A quantitative approach to understanding online antisemitism. arXiv preprint [arXiv:1809.01644](https://arxiv.org/abs/1809.01644) (2018)
5. Forscher, P.S., Kteily, N.: A psychological profile of the alt-right (2017)
6. Graham, R.: Inter-ideological mingling: white extremist ideology entering the mainstream on Twitter. *Sociol. Spectr.* **36**(1), 24–36 (2016)
7. Greven, T.: The rise of right-wing populism in Europe and the United States. A Comparative Perspective. Friedrich Ebert Foundation, Washington DC Office (2016)
8. Hawley, G.: Making Sense of the Alt-Right. Columbia University Press, New York (2017)
9. Hine, G.E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., Stringhini, G., Blackburn, J.: Kek, cucks, and god emperor trump: a measurement study of 4chan’s politically incorrect forum and its effects on the web. In: Eleventh International AAAI Conference on Web and Social Media (2017)
10. Kreis, R.: # refugeesnotwelcome: Anti-refugee discourse on Twitter. *Discourse Commun.* **11**(5), 498–514 (2017)
11. Lyons, M.N.: Ctrl-alt-delete: the origins and ideology of the alternative right. Political Research Associates, Somerville, MA, 20 January 2017
12. Mirrlees, T.: The alt-right’s discourse on “cultural marxism”: a political instrument of intersectional hate. *Atlantis Crit. Stud. Gend. Cult. Soc. Justice* **39**(1), 49–69 (2018)

13. Nagle, A.: Kill All Normies: Online Culture Wars from 4chan and Tumblr to Trump and the Alt-Right. John Hunt Publishing, Alresford (2017)
14. Newman, M.E.: Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* **103**(23), 8577–8582 (2006)
15. Surian, D., Nguyen, D.Q., Kennedy, G., Johnson, M., Coiera, E., Dunn, A.G.: Characterizing Twitter discussions about HPV vaccines using topic modeling and community detection. *J. Med. Internet Res.* **18**(8), e232 (2016)
16. Thorburn, J., Torregrosa, J., Panizo, Á.: Measuring extremism: validating an alt-right Twitter accounts dataset. In: International Conference on Intelligent Data Engineering and Automated Learning, pp. 9–14. Springer (2018)
17. Torregrosa, J., Thorburn, J., Lara-Cabrera, R., Camacho, D., Trujillo, H.M.: Linguistic analysis of pro-ISIS users on Twitter. *Behav. Sci. Terr. Polit. Aggress.*, 1–15 (2019)
18. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: Fourth International AAAI Conference on Weblogs and Social Media (2010)
19. Yin, Z., Cao, L., Gu, Q., Han, J.: Latent community topic analysis: integration of community discovery with topic modeling. *ACM Trans. Intell. Syst. Technol. (TIST)* **3**(4), 63 (2012)
20. Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., Blackburn, J.: What is gab: a bastion of free speech or an alt-right echo chamber. In: Companion Proceedings of the The Web Conference 2018, pp. 1007–1014. International World Wide Web Conferences Steering Committee (2018)



Social Network Analysis of Sicilian Mafia Interconnections

Annamaria Ficara^{1(✉)}, Lucia Cavallaro², Pasquale De Meo³, Giacomo Fiumara⁴, Salvatore Catanese⁴, Ovidiu Bagdasar², and Antonio Liotta⁵

¹ University of Palermo, via Archirafi 34, 90123 Palermo, Italy
aficara@unime.it

² University of Derby, Kedleston Road, Derby DE22 1GB, UK

³ University of Messina, Polo Universitario Annunziata, 98122 Messina, Italy
⁴ MIFT Department, University of Messina, 98166 Messina, Italy

⁵ Edinburgh Napier University, 10 Colinton Road, Edinburgh EH10 5DT, UK

Abstract. In this paper, we focus on the study of Sicilian Mafia organizations through Social Network Analysis. We analyse datasets reflecting two different Mafia Families, based on examinations of digital trails and judicial documents, respectively. The first dataset includes the phone calls logs among suspected individuals. The second one is based on police traces of meeting that have taken place among different types of criminals. Our breakthrough is twofold. First in the method followed to generate these new datasets. Second, in the method used to carry out a quantitative phenomena investigation that are hard to evaluate. Our networks are weighted ones, with each weight catching the frequency of interactions between criminals. Therefore, our analysis focuses on weight and shortest paths distributions in both networks. We identify new types of unusual interactions in the Mafia networks, leading to substantial differences between Mafia networks and other types of criminal or terrorist networks.

Keywords: Criminal networks · Complex networks · Social Network Analysis · Graph theory

1 Introduction

In recent years there has been a growing interest in the application of methods from Statistical Physics and Social Network Analysis (SNA) to the study of different kinds of crimes and, particularly, terrorism. We focus on a specific criminal organization, the Sicilian Mafia, which originated in Sicily and has now spread worldwide [14, 15, 19]. Due to its global spread, Mafia controls entire economic sectors, influencing the social and political life of a country (e.g. by interfering in the results of electoral competitions).

There is a vast number of studies of criminal organizations and, terrorist acts; yet Mafia has characteristics that make it unique. If we analyze terrorism, it is

possible to see how the organizations are formed by individuals who collaborate to pursue an objective and are even willing to sacrifice their lives to achieve that goal (e.g. the Twin Towers attack). After reaching their goal, the terrorist organization generally dissolves (e.g. the IRA in Ireland).

The Mafia's *modus operandi* is different. A Mafia clan lasts for several generations, and is characterized by two key elements: the close links among affiliates and the ability to lead illegal activities to pursue specific objectives. In the first element, the links between mobsters are very close and are marked by reciprocal altruism. In the second element, the clan identifies objectives that it considers to be profitable and almost safe (e.g., human trafficking), and focuses its resources on those objectives. As the objectives change over time (e.g. the members of the clan risk overexposure and capture, or if the deal is no longer profitable), the organization defines new objectives and redesigns its structure to achieve them.

In complex networks terms, this is a rather unusual behaviour, which is almost never detected in other criminal networks. For this reason, many studies have been done on the structure and evolution of a Mafia syndicate (called "cosca", "clan" or "Family"). Yet, existing studies are mainly qualitative since very few (and restricted) datasets exist. We overcome this limitation by creating datasets that allow network analysis.

The main challenge was in the definition of a new dataset directly derived from judicial documents. The information in these documents were verified by the police and the magistrates during a trial, making this a reliable dataset. An additional challenge is that such datasets are bound to be incomplete, for instance due to gaps in the investigation process. A common case is when the police is not authorized to intercept a specific group of individuals during a certain period of time.

The reference scenario under scrutiny is particularly interesting, since it considers a "cosca" operating in the North of Sicily that exercises almost total control over the procurement and execution of public works. This "cosca" acts also as a link between the most important "criminal Families" operating in Palermo and Catania (the biggest cities in Sicily). Thus, the case study addressed in the paper is derived from real-world data and has great significance. Our dataset describes the "cosca" from two different perspectives, including phone calls logs and meetings. The first one has been obtained from eavesdropping, while the second one has been derived from police surveillance data. The availability of two disjoint datasets allow us to evaluate network efficiency and entirety from two different angles. Furthermore, the use of two datasets has allowed us to compensate for the missing data (gaps) incurred in both the interceptions and in the surveillance one.

From those datasets we built two weighted undirected graphs: the *Phone Calls* and *Meetings* networks, which capture interceptions and eavesdropping, respectively. The availability of weighted real criminal networks is an innovative tool that allowed us to conduct several analysis. The weights represent the number of interactions among individuals. This parameter is also very significant to understand how much these criminal networks are different from other existing ones.

Indeed, in these networks, the affiliates avoid talking to each other, while there are frequent contacts both in-between family members and outside the family (e.g. when a fugitive makes contact with his sister). Interestingly, the leaders of the organization (i.e., the bosses) have only minimal interactions, and these are directed to a fairly restricted set of trusted collaborators. In this case, the edge weight acts more like “noise” than actual information.

The plan of the paper is as follows. Section 2, presents the state of the art. In Sect. 3 there is a brief description of the important theoretical definitions required to replicate the experiments here presented. The dataset description extracted from legal acts is addressed in Sect. 4. Next, in Sect. 5, the experiments and results are discussed. In Sect. 6, our conclusions and future developments of the work are showed.

2 Related Work

Social Network Analysis (SNA) is increasingly used by law enforcement agencies (LEAs) to analyze criminal networks as well as to investigate the relations among criminals based on calls, meetings and other events derived from investigations [20–23].

Given the social embeddedness of organized crime and, in particular, of Mafia-like organizations, the analysis of the social structure of Sicilian Mafia syndicates generated a great scientific interest [10, 11].

In 1876, the Italian deputy Leopoldo Franchetti [19] depicted the Mafia as a criminal organization so deeply rooted in Sicilian society to be impossible to destroy except through a real change in Sicilian social institutions.

Sarnecki [18] applied Social Network Analysis to study co-offending behaviors among Swedish teenagers.

Then, Morselli [16] studied the connections within the Gambino family (i.e., a New York-based family), focusing on the career of one of its members, Saul Gravano. Gravano’s ability to build and extend his personal network of contacts over time was a key factor in climbing the Gambino’s family organization.

McGloin [15] analyzed the network structure of street gangs in Newark, New Jersey.

Natarajan [17], built a network of phone calls starting from a dataset consisting of 2,408 wiretap conversations (gathered during the prosecution of a heroin-dealing Mafia syndicate in New York). This network revealed the core of the criminal organization and showed that most of the members had very limited contacts with others in the group.

Calderoni [4] showed that high-status Mafia members were able to indirectly manage illicit drug traffics, by keeping the middle-level criminals in more central and visible positions.

SNA is not only a tool to describe the structure and functioning of a criminal organization, but it is largely employed in the construction of crime prevention systems [8]. For instance, Xu and Chen [23] jointly applied SNA, using hierarchical clustering algorithms. Their approach worked in two stages: firstly,

a criminal network was partitioned into subgroups using a clustering algorithm; secondly, block modelling techniques have been used to extract interaction patterns between these subgroups.

Agreste *et al.* [1] applied percolation theory to efficiently dismantle mafia syndicates.

Calderoni and Superchi [5] showed that the node's betweenness centrality in a meetings network is evidence of Mafia leadership, suggesting that this variable could be exploited by LEAs in selecting the most suitable targets for additional investigations and disruption.

Social Network Analysis tools were also used to identify leaders within a criminal organization. For instance, Mastrobuoni and Patacchini [14] investigated the structure of criminal ties between mobsters using a dataset of 800 Mafia members' criminal profiles. These criminals were active in the United States from 1950s to 1960s. Authors considered various features (such as family relationships, legal and illegal activities) to predict the criminal rank of a mobster.

While the studies above provided insight into the social organization of and possible countermeasures against criminal organizations, the application of SNA to criminal groups nearly inevitably faces problems of *noisy or incomplete information*. Information on a criminal network is often likely to be missing or hidden, due to the covert and stealthy nature of criminal actions [12, 23]. Consequently, the derived networks are incomplete, incorrect, and inconsistent, either due to deliberate deception on the part of criminals, or to limited resources or unintentional errors by LEAs [1, 3, 6, 7, 9]. These limitations may bias the analysis and cause problems of uncertain information, potentially jeopardizing the effectiveness of the investigations [21].

This paper is advancing the state-of-the-art through a creation of novel datasets from real-world data (two weighted undirected graphs), and an analysis that shows how the peculiarities (and internal dynamics) of Mafia families connections may be unveiled via SNA methods.

3 Background

In this section we introduce basic definitions which will be largely used throughout the paper. For more details, see [2, 13].

Undirected Graph. An *undirected graph* is a graph $G = \langle N, E \rangle$, where all the edges E between nodes N are bidirectional. An undirected graph is sometimes called an undirected network. In contrast, a graph where the edges point in a direction is called a *directed graph*.

In this work we dealt with *undirected graphs*, i.e., if $\langle i, j \rangle \in E$ (where $\langle i, j \rangle$ represents the link from the node n_i to n_j), then $\langle j, i \rangle$ belongs to E too.

Weighted Graph. A *weighted graph*, denoted as G with a little abuse of notation, is a triplet $G = \langle N, E, W \rangle$ in which N is the set of nodes, E is the set of edges and $W : E \leftarrow \mathbb{R}^+$ is a function that maps an edge $\langle i, j \rangle$ onto a non negative real number w_{ij} . If a graph is made by weights equals only to zero or one, (i.e., $w_{ij} = 1$, $w_{ij} = 0$), than it is called *unweighted graph*.

Each unweighted and undirected graph G is associated with a symmetric matrix \mathbf{A} such that \mathbf{A}_{ij} equals 1 if and only if there is an edge from i to j , 0 otherwise. Analogously, Each weighted and undirected graph G is associated with a symmetric matrix \mathbf{W} such that $\mathbf{W}_{ij} = w_{ij}$ if and only if there is an edge from i to j with weight w_{ij} and 0 otherwise.

In Sect. 5 we compared the results obtained in weighted scenario with the unweighted one.

Path. A *path* is a sequence of nodes such that each node is connected to the next node along the path by a link. Each path consists of $n + 1$ nodes and n links. The length of a path is the number of its links, counting multiple links multiple times. It is a route that runs along the links of the network. The number of links the path contains is called *path length*. In weighted networks, the path's length is given by the sum of the weighted edges of the path.

Shortest Path. The *shortest path* from non-adjacent node n_i to n_j is the path with the fewest number of links. The shortest path is often called the distance d_{ij} . Multiple shortest paths of the same length d_{ij} can exist. The shortest path never contains loops or intersects itself. In an undirected network $d_{ij} = d_{ji}$. In directed networks, often $d_{ij} \neq d_{ji}$; indeed, in those kind of networks the existence of a path from n_i to n_j does not guarantee the existence of a path from n_j to n_i .

We investigate on the shortest path lengths on both the available datasets to better show the communication behaviour inside a “cosca”. This is an important tool to demonstrate how they act to avoid to overexpose their bosses using a balanced number of intermediates to safely communicate between mobster, as deeply described in Sect. 5.

4 Dataset Description

In this section we explain the structure of the two networks taken into account, shown in Fig. 1 as undirected and weighted graphs. The first dataset, called *Meetings*, describes meetings among suspected criminals, while the second one, *Phone Calls*, refers to phone calls among other criminals.

Table 1 shows the structure of two criminal networks which are extracted from these two datasets. Both the *Meetings* and *Phone Calls* networks can be viewed as *unweighted* and *weighted networks*: for each pair of individual, in fact, we store a coefficient (i.e., the weight w) which represents the number of times the two individuals had a meeting (as reported by the police) and the number of times two individuals called each other. These coefficients are, therefore, understood as the strength of the tie binding two individuals, in each of the two networks.

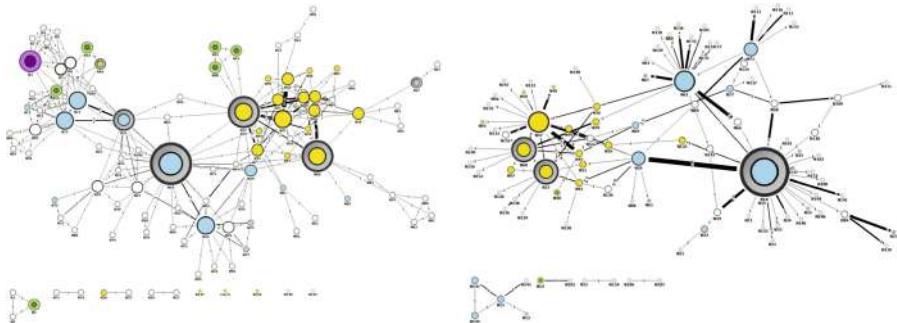


Fig. 1. Left Panel The *Meetings* graph. **Right Panel** The *Phone Calls* graph. The colors represent the different clans. In particular, turquoise nodes represent the members of the “Mistretta” family, while the “Batanesi” family is drawn with yellow nodes. Circled nodes correspond to leaders (i.e., bosses) investigated for having promoted, organized, and directed the Mafia association. The green and purple circled nodes refer to bosses of Mafia families of other mandates. Finally, the white nodes represent other subjects who are close to a family, but are not classifiable in any of the previous categories. In both graphs, the edges width is proportional to the number of meetings or phone calls, and the size of the nodes is proportional to their degree.

Table 1. Statistics of *Meetings* and *Phone Calls* networks.

Parameter	Meetings	Phone Calls
No. nodes	101	100
No. edges	256	124
Max. weight	10	8
Max. frequency	200	100
Max. shortest path	7	14

Nodes may belong to different categories. They may be “bosses” (i.e., the leaders of the criminal organization), or “picciotti” (i.e., the soldiers of the organization). Nonetheless, there are also roles that are not necessarily pointing to criminals. For instance the fruit seller or the baker may somehow be connected to members of the organization for a range of reasons.

The *Meetings* network contains a set of 101 nodes (individuals) that different law enforcement agencies found to have participated to secret meetings. The network includes a total of 256 weighted edges, connecting pairs of nodes. On the other hand, the *Phone Calls* network is composed by a set of 100 individuals who have been legally tapped, or were listed in phone logs. This network includes a total of 124 weighted edges. A subset of nodes (47 in total) are in common between the two networks.

Table 1 provides further useful information, including the maximum weight, the maximum frequency in the affiliates’ interconnections, and the highest number

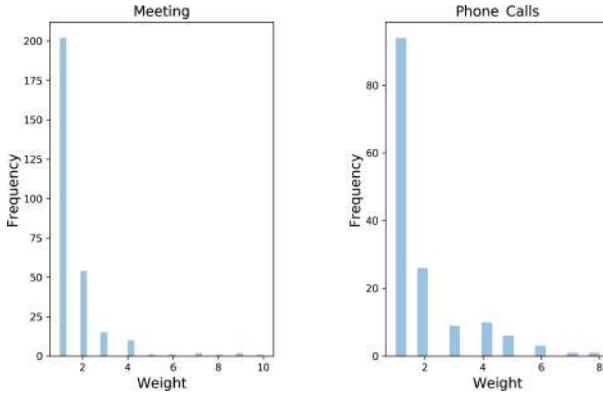


Fig. 2. Weights distribution. **Left Panel** *Meetings* network. **Right Panel** *Phone Calls* network.

of individuals required to connect mobsters, considering all the shortest paths on both datasets. The importance of these data is further discussed in Sect. 5.

5 Dataset Analysis

We begin our study by discussing the edges weights distribution in both *Meetings* and *Phone Calls* networks. In Fig. 2, we have plotted the weight distribution which specifies, respectively, the amount of meeting and phone calls exchanged between pairs of individuals in the networks. On the horizontal axis we report edge weights, while the vertical axis shows frequencies.

Noticeably, both networks exhibit similar characteristics and include several low-weight links. A possible explanation is that the affiliates want to reduce the risk of being intercepted by law enforcement, and even by other people outside the clan. In the *Meetings* network this trend is even more accentuated (the maximum frequency in low-weight links is almost double, as shown in Table 1). Moreover, the maximum weight of interactions among affiliates in the *Meetings* network (i.e., $w = 10$) is greater than the one in the *Phone Calls* network (i.e., $w = 8$). A possible explanation is that mobsters prefer to communicate by physical meeting rather than calling each other, to reduce the risk of being intercepted by the police. Mobsters will find it easier to crypt their conversations in face-by-face meeting, for instance by using body language, or generating background noise. Furthermore, bosses often have to participate to Mafia events to pursue their power inside a clan. For instance, bosses have to participate to funerals of other affiliates, and other solemn religious demonstrations (masses, processions, etc.). During those kinds of events, they also have the opportunity to pass messages to their closest subordinate affiliates. Moreover, it is harder for criminals to notice that they are going to be intercepted rather than to be eavesdropped by the police.

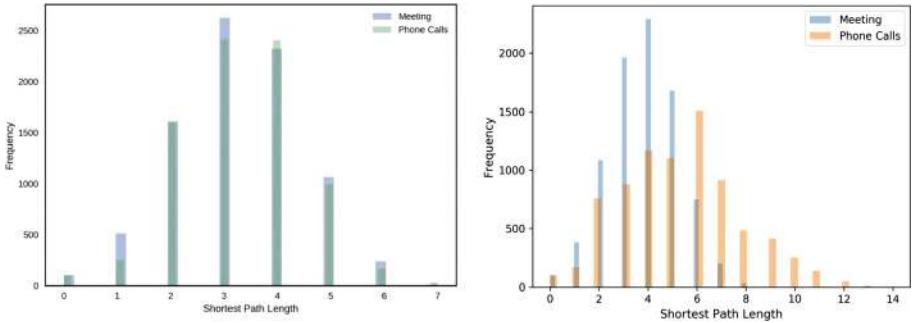


Fig. 3. Distribution of shortest path lengths in *Meetings* e *Phone Calls* networks. **Left Panel** The *unweighted* graph. **Right Panel** The *weighted* graph.

The histograms of shortest path length distributions of Fig. 3 provide useful statistical characterizations of the two networks under scrutiny. Path length statistics are closely related to dynamic properties such as velocities of network spreading processes. Usually, criminal organizations are structured in a way as to optimize the number of communications among members, and to efficiently disseminate information. These members can be discovered by following short paths of communications. Moreover, we can discover relationships among individuals belonging to distant groups in the graph because, even when two nodes seem to be distant, there may exist a relatively short path that connects them.

There are similarities between the weighted and the unweighted shortest path length analysis. In both scenarios, indeed, there is a higher interaction frequency among affiliates having a balanced number of intermediates. This behaviour confirms the hypothesis that inside a “cosca” it is better to avoid the borderline cases. On one hand, if the shortest path is composed of a lower number of affiliates, the bosses are overexposed to police investigations. On the other hand, the longest the number of intermediates, the higher the chances to be intercepted by people outside the Family.

Furthermore, in the weighted simulations emerges a lower frequency of interactions in the *Phone Calls* dataset compared with the same shortest path length of the *Meetings* network. This behaviour, emerged also in Fig. 2, proves that the clan tries to minimize the risk of interceptions, specially to avoid exposing those mobsters who are hierarchically in a higher rank.

The availability of a real weighted graph is a valuable asset in order to conduct a more thorough network analysis. Indeed, in the unweighted scenario this behaviour is not highlighted because both datasets seem to act in the same way.

6 Conclusions

In this paper, important improvements compared to the state-of-the-art have been achieved. The main challenge was the generation of two real-world datasets,

capturing the interaction among Sicilian Mafia members, which have been validated with law and law-enforcement experts. What makes this paper unique is also a quantitative study on the unusual interactions among the clan affiliates. Indeed, a “cosca” network acts differently from other criminal organisations (e.g. terrorist affiliations). Individuals tend to periodically re-aggregate in pursuit of changing goals and to survive over time. There are cases in which the same goal persists for several generations. Whereas in other occasions goals will change rapidly, depending on external socio/economic/legal changes.

From the weight distribution of the two datasets we could figure out which individuals called or met more often. While each connected pair provides evidence of at least one interaction, the edge weights have proved invaluable in unveiling the most significant connections, both within and outside Mafia families.

The comparative analysis of shortest path length between weighted and unweighted graphs shows how Mafia members favor indirect communications through a well-trusted set of intermediaries. Also this form of communication effectively spreads information among key affiliates.

Moreover, the weighted analysis gives us more accurate results than in the unweighted analysis. Indeed, the frequency in the *Phone Calls* dataset in the weighted scenario is lower than the *Meetings* one. This highlights how the clans succeed in reducing the risk of being intercepted by the police. This behaviour is masked in the unweighted analysis.

This paper opens the doors to a vast range of further analyses. What emerges is that conventional analysis based on node centrality are insufficient on their own. New metrics have to be considered to gain better insights into Mafia clans interconnections, which communicate differently from other social networks. One possibility would be to combine popular centrality metrics with new ones that better capture these anomalous types of communications. The clan bosses are indeed the most powerful individuals. Yet they appear to generate the least frequent interactions. Instead, the soldiers (or “picciotti”) emerge as the most important nodes. Thus, conventional network analysis will fail in identifying the bosses.

Under this prospective, it may also be possible to analyse the role that small traders (e.g. greengrocers or bakers) have in facilitating the interactions within and outside the Families mobsters. Observing how often a member of a clan meets people outside the organization, could make it possible to discover communication patterns and, in turn, differentiate between two types of interactions: (1) unrelated to the Mafia context (e.g. mobsters who occasionally buy something); (2) requests for protection (i.e. “pizzo”/racket) by the traders, which is typically periodical. This could be achieved using temporal networks analysis.

References

1. Agreste, S., Catanese, S., De Meo, P., Ferrara, E., Fiumara, G.: Network structure and resilience of Mafia syndicates. *Inf. Sci.* **351**, 30–47 (2016). <https://doi.org/10.1016/j.ins.2016.02.027>
2. Barabási, A.-L., Pósfai, M.: *Network Science*. Cambridge University Press, Cambridge (2016)
3. Calderoni, F.: Morselli, Carlo: inside criminal networks. *Eur. J. Crim. Policy Res.* **16**(1), 69–70 (2010). <https://doi.org/10.1007/s10610-010-9118-7>
4. Calderoni, F.: The structure of drug trafficking mafias: the ‘Ndrangheta and cocaine. *Crime Law Soc. Change* **58**(3), 321–349 (2012). <https://doi.org/10.1007/s10611-012-9387-9>
5. Calderoni, F., Superchi, E.: The nature of organized crime leadership: criminal leaders in meeting and wiretap networks. *E. Crime Law Soc. Change* **72**(4), 419–444 (2019). <https://doi.org/10.1007/s10611-019-09829-6>
6. Campana, P., Varese, F.: Listening to the wire: criteria and techniques for the quantitative analysis of phone intercepts. *Trends Organ. Crime* **15**(1), 13–30 (2012). <https://doi.org/10.1007/s12117-011-9131-3>
7. Catanese, S., De Meo, P., Ferrara, E., Fiumara, G.: Detecting criminal organizations in mobile phone networks. *Expert Syst. Appl.* **41**(13), 5733–5750 (2014). <https://doi.org/10.1016/j.eswa.2014.03.024>
8. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: Crime data mining: a general framework and some examples. *IEEE Comput.* **37**(4), 50–56 (2004). <https://doi.org/10.1109/MC.2004.1297301>
9. Ferrara, E., De Meo, P., Catanese, S., Fiumara, G.: Visualizing criminal networks reconstructed from mobile phone records. In: *CEUR Workshop Proceedings*, vol. 1210 (2014)
10. Kleemans, E.R., Van de Bunt, H.G.: The social embeddedness of organized crime. *Transnatl. Organ. Crime* **5**, 19–36 (1999)
11. Kleemans, E.R., De Poot, C.J.: Criminal careers in organized crime and social opportunity structure. *Eur. J. Criminol.* **5**(1), 69–98 (2008). <https://doi.org/10.1177/1477370807084225>
12. Krebs, V.E.: Mapping networks of terrorist cells. *Connections* **24**(3), 43–52 (2002)
13. Latora, V., Nicosia, V., Russo, G.: *Complex Networks: Principles, Methods and Applications*. Cambridge University Press, Cambridge (2017)
14. Mastrobuoni, G., Patacchini, E.: Organized crime networks: an application of network analysis techniques to the American Mafia. *Rev. Netw. Econ.* **11**(3) (2012). <https://doi.org/10.1515/1446-9022.1324>
15. McGloin, J.M.: Policy and intervention considerations of a network analysis of street gangs. *Criminol. Public Policy* **4**(3), 607–635 (2005). <https://doi.org/10.1111/j.1745-9133.2005.00306.x>
16. Morselli, C.: Career opportunities and network-based privileges in the Cosa Nostra. *Crime Law Soc. Change* **39**(4), 383–418 (2003). <https://doi.org/10.1023/A:1024020609694>
17. Natarajan, M.: Understanding the structure of a large heroin distribution network: a quantitative analysis of qualitative data. *J. Quant. Criminol.* **22**(2), 171–192 (2006). <https://doi.org/10.1007/s10940-006-9007-x>
18. Sarnecki, J.: *Delinquent Networks: Youth Co-offending in Stockholm*. Cambridge University Press, Cambridge (2001). <https://doi.org/10.1017/CBO9780511489310>
19. Sonnino, S., Franchetti, L.: *La Sicilia nel 1876*. G. Barbèra (1877)

20. Sparrow, M.K.: The application of network analysis to criminal intelligence: an assessment of the prospects. *Soc. Netw.* **13**(3), 251–274 (1991). [https://doi.org/10.1016/0378-8733\(91\)90008-H](https://doi.org/10.1016/0378-8733(91)90008-H)
21. Strang, S.J.: Network analysis in criminal intelligence. In: Masys, A. (ed.) Networks and Network Analysis for Defence and Security. LNSN. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-04147-6_1
22. Van der Hulst, R.C.: Introduction to Social Network Analysis (SNA) as an investigative tool. *Trends Organ. Crime* **12**(2), 101–121 (2009). <https://doi.org/10.1007/s12117-008-9057-6>
23. Xu, J., Chen, H.: Criminal network analysis and visualization. *Commun. ACM* **48**(6), 100–107 (2005). <https://doi.org/10.1145/1064830.1064834>



Who Ties the World Together? Evidence from a Large Online Social Network

Guanghua Chi^{1(✉)}, Bogdan State², Joshua E. Blumenstock¹,
and Lada Adamic²

¹ School of Information, U.C. Berkeley, Berkeley, CA 94720, USA

{guanghua,jblumenstock}@berkeley.edu

² Facebook, Menlo Park, CA 94025, USA

bogdan@instagram.com, ladamic@fb.com

Abstract. Social ties form the bedrock of the global economy and international political order. Understanding the nature of these ties is thus a focus of social science research in fields including economics, sociology, political science, geography, and demography. Yet prior empirical studies have been constrained by a lack of granular data on the interconnections between individuals; most existing work instead uses indirect proxies for international ties such as levels of international trade or air passenger data. In this study, using several billion domestic and international Facebook friendships, we explore in detail the relationship between international social ties and human mobility. Our findings suggest that long-term migration accounts for roughly 83% of international ties on Facebook. Migrants play a critical role in bridging international social networks.

Keywords: Migration · Social networks · Big data

1 Introduction

Social connections between individuals in different countries provide a foundation for international trade and commerce, and for global peace and cooperation [23, 40]. A rich literature documents *how* the world is connected, examining the nature, determinants and consequences of social connections between countries. While early studies relied heavily on customs data, foreign direct investment accounts, and international trade data [18], more recent research has integrated data from online sources such as messaging applications and social media sites [19, 30, 45]. Much less is known about *who* connects the world, and how micro connections affect macro network structure. Understanding how the world is connected has practical value, as it can provide a starting point for scholars and policy makers who seek to understand international relations from a network perspective [20], including, for instance, work on the importance of network brokerage (see [10]). More generally, a better understanding of this transition from

the individual to the transnational comes to address the *micro-to-macro* problem identified by Coleman [13] as the fundamental challenge on the path to a science of society.

This study uses Facebook data to provide a disaggregated understanding of the network connections of migrants and non-migrants on one of the world's largest social networks. The Facebook dataset allows for a high-level view of the demographic characteristics and network structures of the world's "international brokers," i.e., the people whose social ties quite literally connect the world. This allows us to ask the central question of our study: who ties the world together?

We present three main results. First, we provide empirical evidence that migrants are a central binding force in the global social network. The act of migration reshapes the network by transforming domestic ties to international ones. The friends they made prior to their move now all know someone who lives in a different country. At the same time, the friends they make in the new country now potentially have a new international tie. These friends now know someone who is *from* another country. With such potential to convert or generate new international ties, it is perhaps unsurprising that over 83% of all international ties involve migrants. These results are consistent with macro-level analyses performed by Perkins and Neumayer [38], who found migrants to play an important role in international communication networks.

Second, we find that migrants act as a bridging force that shrinks the network distance between other people in the Facebook social graph. This is evident in simple descriptive statistics: migrants have higher betweenness in the Facebook graph, particularly when considering connections across countries. We also run simulations that compare the approximate average shortest path length in two graphs: one containing only ties between non-migrants, and one both locals and migrants. Despite our increasing the number of nodes in the graph, we find that the average shortest path length decreases when migrants are included. Both results emphasize the bridging role of networks in connecting distant sub-networks.

Finally, we expand our analysis to the characteristics of migrants and their *local* social networks, to better understand the role that migrants play in their immediate network neighborhood. We establish that migrants' ego networks have fewer dense cores, and that migrants tend to occupy a less redundant position in their ego network, leading us to the conclusion that migrants are also more likely to act as local network bridges. Taken together, these results emphasize the important role that international migrants play in binding together global communities.

2 Related Work

A varied literature has examined social connections between countries. We distinguish between three main areas of research: urban networks, online social networks, and research on international migration.

Traditional international network analysis has focused on understanding urban networks using aggregated datasets such as flight passenger flows, telecommunication volume, and corporate organization [15, 42]. Airline passenger flows have been used to proxy international human flows across urban networks, under the assumption that important cities receive more airline passengers. Common inter-airport passenger flow datasets have been extracted from the International Civil Aviation Organization (ICAO) [27, 42] and Marketing Information Data Transfer (MIDT) [14, 17], which have been used to rank key cities in Western Europe and North America [14, 25, 42], find global hierarchical structures [44, 55], and detect temporal changes of a city's importance in the global city network [35, 44] by adopting network analysis methods. Derudder and Witlox [16] pointed out several limitations posed by the use of airline passenger flow data, including the lack of origin and destination information because of stopovers, missing inter-state flow, and possible flows to tourist destinations. In spite of these issues, airline passenger flows remain the most commonly used data source to analyze international urban networks.

Internet backbone networks can also reflect the role of cities and the connections between countries, under the assumption that important cities would have more high-speed internet connections and more connections to other cities [4, 5, 34, 49]. This assumption is often untenable, however. A small city may act as a gateway between core cities and its centrality in the internet backbone network may exaggerate its importance in the worldwide social system [41]. Another traditional dataset comes from the realm of multinational corporate organization. International business companies create new offices globally to distribute their service for their corporate benefits. The transnational network formed by international offices captures the information flow and products flow [6]. The use of this dataset comes with its own limitations, given that transnational flows are inferred instead of directly obtained like airline passenger flows [16].

In recent years, the growing availability of large social datasets has enabled a new, fine-grained level for the understanding transnational social networks, thanks to increases in Internet penetration and the development of global social networking platforms, such as Microsoft Messenger instant-messaging system [30], Twitter [19, 28, 47], Flickr [11], and Facebook [3, 52]. Network structures are analyzed to understand the properties of social networks, including degree distribution, clustering, the small-world effect, and homophily [2, 37, 50]. For example, Backstrom et al. [2] found that the degree of separation is 3.74 based on 721 million people at Facebook in 2011. The most recent result is 3.6 degrees of separation in 2016, showing that people have grown more interconnected [7].

There has been growing interest in combining spatial and social network analyses to understand the relationship between social networks and migration [1, 8, 12, 32]. International and internal migration patterns have been explored using different sources of new datasets, such as geo-tagged tweets [21, 45], IP geolocation [46, 53], and social network profile fields [22]. This research has focused on the factors related to international social networks and migration, including distance and trade, community structure, and interactions across countries. In

this line of work, three recent papers are most relevant to this study. Kikas et al. [26] found that social network features can explain international migration in terms of net migration per country and migration flow between a pair of countries. Herdagdelen et al. [22] analyzed the social networks of migrants in the United States by leveraging profile self-reports of home countries. Zagheni et al. [54] showed the viability of conducting demographic research related to international migration through the public Facebook advertising API.

Our research comes to extend the study of international social networks using online data, shifting the focus from the country-to-country to the individuals whose social connections span the boundaries of countries and who quite literally connect the world. We develop a vocabulary to describe social ties in terms of both parties' home and current countries, which we use to provide an examination of both triads and ego networks. Our analysis concludes with a foray into the role of migrants with regard to the connectivity of the global Facebook social graph.

3 Data and Methods

Our analysis makes use of de-identified profile and social connection data available on Facebook, presently the world's largest social networking platform, which as of the time of writing numbered more than 2.25 billion monthly active users. These data have several key limitations: the population of Facebook users is not representative, particularly outside of the U.S. and Western Europe; the connections observed on Facebook are a biased sample of actual social connections; and the data are not broadly accessible to the research community [9, 36]. Yet the ability to observe the social connections between such a substantial fraction of the world's population also provides unique advantages for social and demographic research.

We use the Facebook data to simultaneously observe social network structure and migration status for the full population of Facebook users (where available through profile self-reports) in 2018. Each active user represents a node in the network; two nodes are connected by an edge if they have mutually agreed to be 'friends' on the online platform. Example subnetworks are depicted later, in Fig. 3.

Separately, we use de-identified Facebook profile information to determine the current and origin country of each user. The country of origin is determined by the self-reported "home town" that users enter on their profile pages. The current country assignment is determined by Facebook for growth accounting purposes, and is based on typical country-level geolocation signals, such as recent IP addresses. There is a considerable amount of measurement error in this approach to inferring migration, as how people report their "home" town is the result of subjective interpretation. While we do not think this measurement error entirely undermines the high-level analysis that we present in this paper, such data may not be well-suited to more disaggregated analysis, or seen as a substitute for official statistics.

By aggregating home and current country of users we were able to generate a migrant stock dataset, showing the current numbers of individuals “from” one country who currently live in another country. We validated the country-to-country dataset we generated against data on international migrant stocks provided by the World Bank [39]. Here we chose those countries with more than 1 million monthly active users, and those country pairs with more than 0.001% of migrants. The magnitude of migrant stocks quantified using Facebook data is highly (though not perfectly) correlated to migrant stock estimates produced by the World Bank (Pearson’s ρ : 0.87), which is similar to the findings of Zagheni et al. [54]. Because migration events may be short-lived (e.g. study abroad or volunteer programs) for young adults, we focus our analysis on users aged over 30 at the time of our study.

4 Results

4.1 Migrants Tie the World Together

Our first set of results highlight the substantial fraction of international ties on Facebook that are comprised by migrants. Formally, we denote the home and current country of a person i by H_i and C_i , and say that i is a migrant if $H_i \neq C_i$. A social tie exists between i and j if they are friends on Facebook. International ties exist if i and j have different current countries ($C_i \neq C_j$) or different home countries ($H_i \neq H_j$).

A striking result is evident when we look at the fraction of international and domestic ties that involve migrants. While only 17.1% of all ties on Facebook involve a migrant, a staggering 82.91% of international ties involve at least one migrant. These results are presented and disaggregated in Table 1.

Table 1. Domestic and international ties (univariate statistics)

	International ties (%)	Domestic ties (%)	All ties (%)
Non-migrants	17.09	99.14	82.90
Migrants	82.91	0.86	17.10
... Two migrants	7.66	0.86	2.21
... Migrant to a resident in the destination country	39.40	0	7.79
... Migrant to a resident in the origin country	27.88	0	5.52
... Migrant to a resident in other countries	7.97	0	1.58

Of the international ties we observe, 39.4% exist between migrants and locals in destination countries, and 27.88% of international ties connect migrants with people in the country of origin.

Only 17.09% of all international ties in our sample are between non-migrants – individuals in different countries whose own current countries are the same as their

stated home countries. This leads to the staggering conclusion that international migration is responsible for over 83% of social ties between countries. Even this statistic may underestimate the percentage of international ties due to migration, given that our analysis does not account for return migration – i.e., the situation in which an individual has returned to their country of origin but maintains ties in their former migrant destination.

Further strengthening the conclusion regarding the crucial role migrants play in providing international ties is Fig. 1, which shows that the distribution of the per-individual proportion of international ties is bimodal, comprised of a mixture of migrants, who have a high concentration of international ties (the average migrant’s network contains 90.5% international ties), and non-migrants, whose social networks are dominated by domestic ties (only 10% of their ties are international).

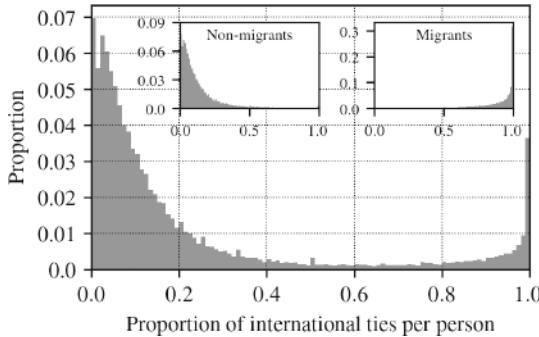
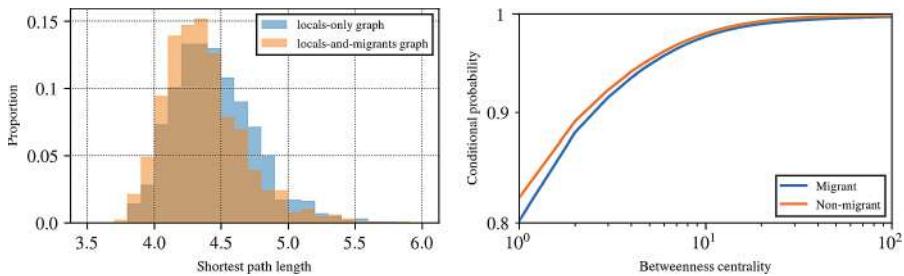


Fig. 1. Proportion of ties that are international.

4.2 Migrants and Measures of Global Cohesiveness

Our second set of results investigate the extent to which migrants play a binding role in the global social network. Here we reproduce the approximation of the average shortest-path computed by Bhagat et al. [7] and Backstrom et al. [2], using two graphs as input. The *locals-only* graph, only contains those users for whom the home country is the same as the current country. The *locals-and-migrants* graph results from adding migrants (users with known different home and current countries) to the *locals-only* graph. We sample 1000 seed nodes in each graph to compute the approximate average shortest path using the methodology described in Bhagat et al. [7]. It should be noted that the approximate average shortest path length from these two graphs is not directly comparable to previous results about the entire Facebook social graph, since home-country self-reports are only available for a fraction of Facebook users. We found that the average shortest path length is 4.45 for the *locals-only* graph, and 4.37 for *locals-and-migrants* graph (Fig. 2a). In other words, the degree of



(a) Shortest path length in the *locals-only* graph vs. the *locals-and-migrants* graph. (b) Betweenness centrality distribution of migrants vs. non-migrants.

Fig. 2. Bridging role of migrants in international social networks

separation is 3.45 in the *locals-only* graph, and 3.37 in the *locals-and-migrants* graph. A two sample t-test confirms that this difference is statistically significant ($p < 0.001$). Even though there are more nodes in the *locals-and-migrants* graph than the *locals-only* graph, the average shortest path in the *locals-and-migrants* graph is smaller, meaning that the migrants serve as a bridge to bring the world together.

Table 2. Betweenness centrality statistics for migrants (M) and locals (L).

Statistic		Mean	S.D.	Median
Betweenness	M	8.12	25302.26	1.07
	L	7.66	69286.75	1.04
... same	M	45.95	90612.70	1.26
... country	L	79.99	305134.88	1.08
... different	M	6.25	16219.46	1.07
... country	L	3.79	8400.1	1.04
Degree	M	372	513	214
	L	395	544	244

In addition to measuring the shrinkage in the global Facebook graph when migrants are added, it is also possible to compute the number of shortest paths which would be routed through migrants and non-migrants when a social search is performed. To this end, we compute weighted approximate betweenness centrality: starting from 24 randomly-selected seeds we compute shortest paths to all nodes in the Facebook social graph (friendships of monthly active users). We then count the number of shortest paths passing through each vertex in the graph, weighted so that the weights of multiple shortest paths connecting any two vertices all sum to 1. Betweenness statistics for migrants and non-migrants

are shown in Table 2, suggesting that migrants have higher betweenness despite having lower degree. To better understand what drives this dynamic we plot cumulative distribution function for migrants' and locals' betweenness centrality in Fig. 2b. The figure shows that migrants are over-represented among individuals with very high betweenness compared to locals.

While the majority of both migrants and locals have relatively low betweenness, there are more migrants among those who act as conduits for many of the shortest paths in the Facebook social graph. To better understand the role that migrants play in brokering international ties we can also distinguish between situations where ego and the seed are in the same country or in different countries. When making this distinction we can see in Table 2 that, among users in a different country than the seed, migrants help route almost twice as many (6.25) shortest paths as locals (3.79), whereas migrants only route about half as many shortest paths (45.95) as locals (79.99) to a seed in the same current country. This further seems to suggest that migrants have a particularly important role in providing inter-country connectivity: they not only participate in a great number of international ties but their ties are also more likely to function as international network bridges.

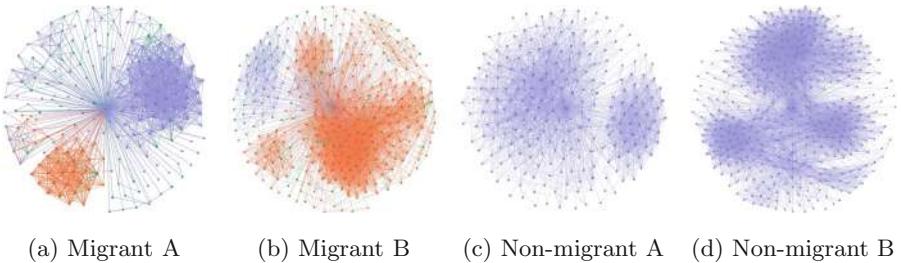


Fig. 3. Ego networks of two migrants and two non-migrants. *Note:* The center node is the ego. All the other nodes are his or her friends. The node color refers to different countries: orange nodes are living in the ego's home country; violet nodes are living in the ego's current country; green nodes are living in other countries.

4.3 Ego-Networks

We have seen so far that migrants have more international ties, and that they play an oversize role in improving connectivity in the global social graph. A natural question arises as to whether migrants' *local* networks differ in other structurally meaningful ways from those of non-migrants. The analysis of ego-networks can help establish the extent to which individuals help connect disjoint collections of alters, providing important measures of network brokerage. Figure 3 shows four example ego networks, two of migrants and two of non-migrants, with violet nodes and edges indicating connections in the current country and orange nodes and edges representing connections in the home country.

We can see that the two migrants' home and current country networks are disjoint, with no direct connection between alters in the home and current country. In this case the migrant ego provides a shortest path between each pair of alters in the home and current country, respectively.

To measure the ego-networks of users we measure multiple statistics:

- size of ego network, i.e. a user's number of Facebook friends (alters).
- ego's clustering coefficient, or the proportion of triads ego participates in that are closed.
- k -cores, or the maximal subgraph of the ego graph, in which nodes have degree of at least k . We compute k -cores for all possible k 's in the ego-network.

Table 3. Ego-network statistics for migrants (M) and locals (L). Note: G_H is the graph of all users who share their home country. G_C is the graph of all users who share their current country.

Statistic		Whole		G_H		G_C	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
Degree	M	373	517	129	228	160	312
	L	388	533	255	358	352	491
	p -val.	0.04		<0.01		<0.01	
Density	M	0.120	0.134	0.247	0.248	0.209	0.206
	L	0.118	0.119	0.139	0.136	0.126	0.127
	p -val.	0.19		<0.01		<0.01	
8-core	M	0.865	0.553	0.462	0.557	0.498	0.548
	L	0.871	0.512	0.732	0.561	0.839	0.515
	p -val.	0.38		<0.01		<0.01	
64-core	M	0.070	0.256	0.014	0.116	0.025	0.157
	L	0.077	0.267	0.041	0.198	0.067	0.251
	p -val.	0.07		<0.01		<0.01	

Given the computational requirements of the analysis, running it for all users would be prohibitively expensive. Because we are interested in the structural differences between migrants and non-migrants, we chose to run an analysis on a balanced sample of users. We analyzed a sample of 20,000 users (10,000 migrants and 10,000 non-migrants) drawn at random from among monthly active Facebook users aged between 30 and 80. Ego-network statistics were computed for the entire ego-graph, as well as for two subgraphs: the graph of all users who share their current country (G_C), and the graph of all users who share their home country (G_H). As Table 3 reveals, migrants appear to have slightly lower degree than locals. On average, a migrant in our sample had 373 Facebook friends, whereas a local had 388 Facebook friends, this difference being statistically significant at the 0.05 level ($p = 0.04$ using a two-sample t-test).

Migrants were also comparatively less connected to their home and current countries than locals. On average, the home ego-network G_{Hi} of a migrant i – composed of people with the same stated home country as the ego – had 129 nodes, whereas the home ego-network G_{Hj} of a local j had 255 nodes. Similarly, the ego-network in the current country G_{Ci} of a migrant i had a mean of 160 nodes, whereas the ego-network in the current country G_{Cj} of a local j had 352 nodes. Given that their ego networks are split between home and current country, it is not surprising that migrants have fewer alters to draw on in each country. These alters are more likely to be connected to one another however: migrants' home-country ego networks have a density of .247, compared to .139 for locals. The same numbers are reflected when G_{Ci} are considered: .209 for migrants and .126 for locals. This result would seem to suggest that migrants' home and current countries are more cohesive than non-migrants, but one has to consider the fact that degree and clustering coefficient have been found to be inversely correlated [24, 29, 31]. That is, it is possible that migrants have different network foci split between home and current country, whereas all of a local's foci will be in their current country. For instance, a migrant who leaves after high school to attend university in a different country may have one high school friendship group in the home country and another college friendship group in the current country, whereas a local will have both groups in the same country. Even if the two friendship groups have the same density, the migrants' home and current countries will appear to be denser because they only contain their high school and college friendship groups, respectively.

Table 3 also reports the average number of 8- and 64-cores in migrants' and locals' ego-networks. A k -core is defined as a subset of nodes in the ego-network network which have a degree of at least k when connected to one another. These results reveal that migrants have fewer 8- and 64-cores in their home and current country ego networks, while the difference between the number of k -cores in their overall ego networks is much smaller (.865 for migrants vs. .871 for locals for 8-cores, $p = 0.38$ and .070 for migrants vs. .077 for locals for 64-cores, $p = 0.07$). This suggests that migrants ties' are about as clustered as non-migrants', but the cores in their ego-networks are divided between multiple countries. The k-core structure reinforces the multiple country-foci explanation advanced above.

4.4 Triadic Closure

Beyond the direct connections between two individuals, larger graph structures can provide insight into the role that migrants play in the broader social network. In particular, network *triads* – which indicate whether two friends of an individual are themselves friends – have long been recognized as fundamental elements of social networks irreducible to their parts [43].

The triadic view poses a more complex challenge due to the exponential increase in complexity resulting from the various combinations possible between the home and current countries of the three actors who participate in a triad. We therefore downsample the Facebook graph to 10% of all monthly active users for

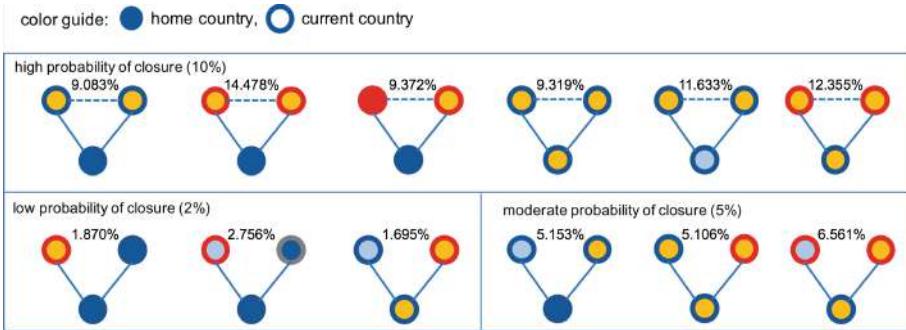


Fig. 4. Triadic closure probabilities for a sample of triads, illustrating that closure is most likely for migrants sharing home and current country. Each node is an individual, with fill color designating a home country, and the border color designating their current country.

whom both home and current country were available. We counted 15bn triads connecting this subset of users.

Figure 4 shows a sample of possible triads. The figure suggests that when two people share a friend in common as well as the same home and current country, they are most likely to be friends themselves. People who share neither home nor current country are unlikely to be friends, even if they share a common friend, while friends-of-friends who share either home or current country are moderately likely to be acquainted themselves. Given that triads – and the extent to which they are closed or not – form the building blocks of social networks, we hope that these closure probabilities can be useful to future research efforts into the topology and dynamics of large-scale social networks.

5 Conclusion

Both mundane and essential, social ties underpin the global political and economic system. The connection between social networks and globalization has long elicited a great deal of interest among social scientists. Studies of the global social network have only become possible recently, thanks to increases in Internet penetration and the development of global social networking platforms. Increasingly, we can understand international interactions not just through proxies of international flows such as air passenger data and internet bandwidth between countries, but also through the records of connections between people. In this study, to our knowledge the first of its kind at this scale, we focus on the people who connect the world's social network.

We use an de-identified, aggregated dataset from the Facebook platform to examine the relationship between human mobility and the development of international ties. Our findings suggest that long-term migrations likely account for about 83% of the world's international ties. Our ego network analysis revealed

that migrants' networks have higher density, but lower degree, in both home and current countries than non-migrants'.

We also confirmed the "bridging" role of migrants in connecting the world's social network. By computing the average shortest path length in a social graph with and without migrants, we showed that migrants effectively decrease the length of the average shortest path. We also learned that migrants tend to act as conduits for more shortest paths than non-migrants. From these results we can conclude that migrants play an important role in the global economy and society [33,48], effectively bringing the world closer together.

We acknowledge the particularly strong tension in network datasets between data privacy and research reproducibility, and hope that both academia and industry will continue working together to find effective ways for sharing large datasets for social science research purposes. To help future researchers with understanding the complex interactions between friendship and international mobility, we have also computed exhaustive triadic closure probabilities between all combinations of migrants and locals. We found that, generally speaking, triads tend to be closed when migrants are present, but only if a current or home country is shared between alters. We hope these aggregations will likewise help advance future social network analysis research, for instance by providing the baseline for simulations.

While this paper has focused on the structure of the network formed by friendship ties between people, there are other types of connections which span the globe. One could ask, for example, what fraction of newspapers' international readership stems from migrants? For local newspaper readership, do migrants read more international news? Do they share international news with their friends? What role do migrants play in helping artists become globally popular? Since migrants help to make the world just a bit smaller, by stretching their own ties across the globe, it would also be interesting to examine the role of social media in helping to sustain such long-range ties. We leave these and other questions for future work.

Even though much remains to be done until the mechanisms of social networks will be fully understood, the analyses presented in this paper would have been hard to conceive of 50 years ago when Travers and Milgram [51] performed the first social search experiments. A half century later, it is possible not only to measure the world's connectivity but to ask novel questions of it. We hope that our work will advance scientists' grasp of the social web that envelops the Earth, and of the people who effectively connect the world.

References

1. Adams, J., Faust, K., Lovasi, G.S.: Capturing context: integrating spatial and social network analyses. *Soc. Netw.* **34**(1), 1–5 (2012)
2. Backstrom, L., Boldi, P., Rosa, M., Ugander, J., Vigna, S.: Four degrees of separation. In: Proceedings of WebSci 2012, pp. 33–42 (2012)
3. Bailey, M., Cao, R., Kuchler, T., Stroebel, J., Wong, A.: Social connectedness: measurement, determinants, and effects. *J. Econ. Perspect.* **32**(3), 259–280 (2018)

4. Barnett, G.A.: A longitudinal analysis of the international telecommunication network, 1978–1996. *Am. Behav. Sci.* **44**(10), 1638–1655 (2016)
5. Barnett, G.A., Jacobson, T., Choi, Y., Sun-Miller, S.: An examination of the international telecommunication network. *J. Int. Commun.* **3**(2), 19–43 (1996)
6. Beaverstock, J.V., Smith, R.G., Taylor, P.J.: World-city network: a new metageography? *Ann. Assoc. Am. Geogr.* **90**(1), 123–134 (2000)
7. Bhagat, S., Burke, M., Diuk, C., Filiz, I.O., Edunov, S.: Three and a half degrees of separation. *Facebook Research* (2016)
8. Blumenstock, J.E., Chi, G., Tan, X.: Migration and the value of social networks. *CEPR Discussion Papers*, No. 13611 (2019)
9. Boyd, D., Crawford, K.: Critical questions for big data. *Inf. Commun. Soc.* **15**(5), 662–679 (2012)
10. Burt, R.: Structural holes and good ideas. *Am. J. Sociol.* **110**(2), 349–399 (2004)
11. Cha, M., Mislove, A., Gummadi, K.P.: A measurement-driven analysis of information propagation in the flickr social network. In: *Proceedings WWW 2009*, p. 721 (2009)
12. Cho, E., Myers, S.A., Leskovec, J.: Friendship and mobility: user movement in location-based social networks. In: *Proceedings of KDD 2011*, pp. 1082–1090 (2011)
13. Coleman, J.S.: *Foundations of Social Theory*. Harvard University Press, Cambridge (1994)
14. Derudder, B., Witlox, F.: An appraisal of the use of airline data in assessing the world city network: a research note on data. *Urban Stud.* **42**(13), 2371–2388 (2005)
15. Derudder, B.: On conceptual confusion in empirical analyses of a transnational urban network. *Urban Stud.* **43**(11), 2027–2046 (2006)
16. Derudder, B., Witlox, F.: Mapping world city networks through airline flows: context, relevance, and problems. *J. Transp. Geogr.* **16**(5), 305–312 (2008)
17. Derudder, B., Witlox, F., Taylor, P.J.: U.S. cities in the world city network. *Urban Geogr.* **28**(1), 74–91 (2007)
18. Feenstra, R.C.: *Advanced International Trade: Theory and Evidence*. Princeton University Press, Princeton (2015)
19. García-Gavilanes, R., Mejova, Y., Quercia, D.: Twitter ain't without Frontiers: economic, social, and cultural boundaries in international communication. In: *Proceedings of ICWSM 2014*, pp. 1511–1522 (2014)
20. Hafner-Burton, E.M., Kahler, M., Montgomery, A.H.: Network analysis for international relations. *Int. Organ.* **63**(3), 559–592 (2009)
21. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C.: Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **41**(3), 260–271 (2014)
22. Herdağdelen, A., State, B., Adamic, L., Mason, W.: The social ties of immigrant communities in the United States. In: *Proceedings of WebSci 2016*, pp. 78–84 (2016)
23. Hollis, M., Smith, S.: *Explaining and Understanding International Relations*. Clarendon Press, Oxford (1990)
24. Jacobs, A.Z., Way, S.F., Ugander, J., Clauset, A.: Assembling the facebook: using heterogeneity to understand online social network assembly. In: *Proceedings of WebSci 2015* (2015). Article 18
25. Keeling, D.J.: Transport and the world city paradigm. In: *World Cities in a World-System*, pp. 115–131 (1995)
26. Kikas, R., Dumas, M., Saabas, A.: Explaining international migration in the skype network: the role of social network features. In: *Proceedings of the 1st ACM Workshop on Social Media World Sensors*, pp. 17–22 (2015)

27. Kyoung-Ho, S., Timberlake, M.: World cities in Asia: cliques, centrality and connectedness. *Urban Stud.* **37**(12), 2257–2285 (2000)
28. Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E.: Mapping the global Twitter heartbeat: the geography of Twitter. *First Monday* **18**(5) (2013)
29. Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: KDD, pp. 462–470. ACM (2008)
30. Leskovec, J., Horvitz, E.: Planetary-scale views on a large instant-messaging network. In: Proceedings of the 17th International Conference on World Wide Web - WWW 2008, pp. 915–924 (2008)
31. Leskovec, J., Horvitz, E.: Planetary-scale views on a large instant-messaging network. In: WWW, pp. 915–924 (2008)
32. Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L.: Social sensing: a new approach to understanding our socioeconomic environments. *Ann. Assoc. Am. Geogr.* **105**(3), 512–530 (2015)
33. Lucas, R.E.B.: African migration. In: Chiswick, B.R., Miller, P.W. (eds.) *Handbook of the Economics of International Migration*, chap. 26, vol. 1, pp. 1445–1596 (2015)
34. Malecki, E.J.: The economic geography of the internet's infrastructure. *Econ. Geogr.* **78**(4), 399–424 (2002)
35. Matsumoto, H.: International urban systems and air passenger and cargo flows: some calculations. *J. Air Transp. Manag.* **10**(4), 239–247 (2004)
36. Mellon, J., Prosser, C.: Twitter and Facebook are not representative of the general population: political attitudes and demographics of British social media users. *Res. Polit.* **4**(3) (2017)
37. Onnela, J.P., et al.: Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci.* **104**(18), 7332–7336 (2007)
38. Perkins, R., Neumayer, E.: The ties that bind: the role of migrants in the uneven geography of international telephone traffic. *Global Netw.* **13**(1), 79–100 (2013)
39. Ratha, D.: Migration and remittances Factbook 2016. The World Bank (2016)
40. Rauch, J.E.: Business and social networks in international trade. *J. Econ. Lit.* **39**(4), 1177–1203 (2001)
41. Rutherford, J., Gillespie, A., Richardson, R.: The territoriality of Pan-European telecommunications backbone networks. *J. Urban Technol.* **11**(3), 1–34 (2004)
42. Short, J.R., Kim, Y., Kuus, M., Wells, H.: The dirty little secret of world cities research: data problems in comparative analysis. *Int. J. Urban Reg. Res.* **20**(4), 697–717 (1996)
43. Simmel, G.: *The Sociology of Georg Aimmel*. The Free Press, Glencoe (1950). Trans. by K.H. Wolff. (Original work published 1908)
44. Smith, D.A., Timberlake, M.F.: World city networks and hierarchies, 1977–1997: an empirical analysis of global air travel links. *Am. Behav. Sci.* **44**(10), 1656–1678 (2001)
45. State, B., Park, P., Weber, I., Macy, M.: The mesh of civilizations in the global network of digital communication. *PLoS ONE* **10**(5), e0122543 (2015)
46. State, B., Weber, I., Zagheni, E.: Studying international mobility through IP geolocation. In: Proceedings of WSDM 2014, pp. 265–274 (2014)
47. Takhteyev, Y., Gruzd, A., Wellman, B.: Geography of Twitter networks. *Soc. Netw.* **34**(1), 73–81 (2012)
48. Todaro, M.: Internal migration in developing countries: a survey. In: *Population and Economic Change in Developing Countries*, pp. 361–402. University of Chicago Press (1980)
49. Townsend, A.M.: Network cities and the global structure of the internet. *Am. Behav. Sci.* **44**(10), 1697–1716 (2001)

50. Travers, J., Milgram, S.: The small world problem. *PSY Today* **1**(1), 61–67 (1967)
51. Travers, J., Milgram, S.: An experimental study of the small world problem. *Sociometry* **32**, 425–443 (1969)
52. Ugander, J., Karrer, B., Backstrom, L., Marlow, C.: The anatomy of the Facebook social graph. [arXiv:1111.4503](https://arxiv.org/abs/1111.4503) (2011)
53. Zagheni, E., Weber, I.: You are where you e-mail: using e-mail data to estimate international migration rates. In: Proceedings of WebSci 2012, pp. 348–351 (2012)
54. Zagheni, E., Weber, I., Gummadi, K.: Leveraging Facebook’s advertising platform to monitor stocks of migrants. *Popul. Dev. Rev.* **43**(4), 721–734 (2017)
55. Zook, M.A., Brunn, S.D.: Hierarchies, regions and legacies: European cities and global commercial passenger air travel. *J. Contemp. Eur. Stud.* **13**(2), 203–220 (2005)

Temporal Networks



Comparing Temporal Graphs Using Dynamic Time Warping

Vincent Froese¹, Brijnesh Jain², Rolf Niedermeier¹, and Malte Renken¹⁽⁾

¹ Faculty IV, Algorithmics and Computational Complexity,
TU Berlin, Berlin, Germany

{vincent.froese,rolf.niedermeier,m.renken}@tu-berlin.de

² Faculty IV, Distributed Artificial Intelligence Laboratory,
TU Berlin, Berlin, Germany
brijnesh.jain@dai-labor.de

Abstract. The links between vertices within many real-world networks change over time. Correspondingly, there has been a recent boom in studying temporal graphs. Proximity-based pattern recognition in temporal graphs requires a (dis)similarity measure to compare different temporal graphs. To this end, we propose to employ dynamic time warping on temporal graphs. We define the dynamic temporal graph warping distance (dtgw) to determine the (dis)similarity of two temporal graphs. Our novel measure is flexible and can be applied in various application domains. We show that computing the dtgw-distance is a challenging (in general NP-hard) optimization problem and we identify some polynomial-time solvable special cases. Moreover, we develop an efficient heuristic which performs well in empirical studies. In experiments on real-word data we show that our dtgw-distance performs favorably in de-anonymizing networks compared to other approaches.

Keywords: Temporal graph alignment · Graph matching · Vertex signatures · NP-hardness · Majorize-minimize algorithm

1 Introduction

A fundamental concept for pattern recognition is the concept of (dis)similarity between objects. For objects that are represented by numerical feature vectors, there exist a lot of well-known (dis)similarity measures such as p -norms or positive semi-definite kernels. In structural pattern recognition, objects are often more naturally represented by more complex (discrete) data structures such as graphs, strings or time series. For these representations, one can often not simply use vector-based (dis)similarity measures. Instead, one needs to define suitable

Full version available on arXiv (<http://arxiv.org/abs/1810.06240>).

B. Jain—Supported by the DFG project JA 2109/4-1.

M. Renken—Supported by the DFG project NI 369/17-1.

domain-specific (dis)similarity measures such as the *edit distances* on graphs or strings or the *dynamic time warping* distance on time series.

The majority of graph (dis)similarity measures focuses on static graphs. This includes the graph edit distance [15], graph kernels [8], and geometric graph distances [11]. However, many complex systems are not static as the links between entities dynamically change over time. Thus, there is a steadily growing research interest in analyzing *temporal graphs* (we also use the term *temporal network* interchangeably). Such temporal graphs can be represented by a series of temporal edges between a fixed set of vertices. Examples are social contact networks, traffic networks, attack networks in computer security, or protein-protein-interaction networks in biology [4, 9, 13, 14, 17]. Many processes described by temporal graphs naturally vary in duration and temporal dynamics (e.g., chemical reactions might proceed with different speed), which makes data mining tasks such as classification challenging. Hence, in order to perform classification or clustering on temporal networks, one needs to find suitable (dis)similarity measures; seemingly for the first time, we attempt to fill this gap in the literature.

Our paper proposes a (dis)similarity measure for temporal graphs based on vertex signature graph distance and dynamic time warping, referred to as *dynamic temporal graph warping* (dtgw). Dynamic time warping is a standard tool in time series mining which can cope with temporal variations. Thus, by combining established methods from graph-based pattern recognition and time series mining in a nontrivial way, we obtain a suitable tool to analyze temporal network data. We study the computational complexity of our method, develop efficient algorithms, and study their behavior on real-world data. Our experimental results confirm the usefulness in practical applications.

Related Work. There are numerous approaches to define (dis)similarity measures on static graphs. A well-known example is the (NP-hard) *graph edit distance* [15]. *Graph kernels* (many of which are polynomial-time computable) are another well-studied class of graph distance measures [8]. A graph distance based on vertex mappings using local vertex signatures was introduced by Jouili and Tabbone [12]. The idea of using vertex mappings can also be found in *optimal assignment kernels* in a machine learning context [6].

Regarding (dis)similarity measures on temporal graphs, we are not aware of any direct approaches. However, other techniques could be used to compare temporal graphs. For example, one approach is based on network embeddings where vertices are mapped into a feature space [3, 18]. Another approach is based on network alignments [5, 17] where a vertex mapping that optimizes some criteria is computed. However, dynamic time warping has not been used in this context so far. Dynamic time warping [16] is an established measure for mining time series data [1] which is specifically designed to cope with temporal variation in the data via nonlinear alignment of observations. We lift this standard concept to the domain of temporal graphs.

Our Contributions. We define the dynamic temporal graph warping distance (dtgw) as a twofold discrete minimization problem involving computation of

an optimal vertex mapping and an optimal warping between time layers (see Sect. 3). As a byproduct, our approach does not only yield a distance measure but also yields an interpretable mapping between vertices of two temporal input graphs which can, for example, be used for de-anonymization in the context of social networks.

We show that the dtgw-distance is NP-hard to compute in general (Theorem 2). In contrast, we point out several polynomial-time solvable special cases. This includes the case when either a vertex mapping or a warping path is fixed (Observation 1), the case of deciding whether the dtgw-distance is zero (Theorem 4), and the case when the lifetimes of the two temporal graphs differ only by a constant and the warping path length is restricted (Theorem 5). Facing the general computational hardness results, in Sect. 5, we propose an efficient and experimentally validated heuristic approach. In Sect. 6, we demonstrate that our heuristic can successfully be used for de-anonymization of real-world temporal social networks and is faster than other existing methods.

Due to the lack of space, several proofs are omitted and can be found in the full version. The full version additionally contains a quadratic programming formulation and additional, more extensive experimental investigations.

2 Preliminaries

For $T \in \mathbb{N}$, we define $[T] := \{1, \dots, T\}$. A *temporal graph* $\mathcal{G} = (V, E_1, \dots, E_T)$ consists of a vertex set V and a sequence of $T \geq 1$ edge sets $E_i \subseteq \binom{V}{2}$. By $G_i = (V, E_i)$, we denote the i^{th} layer of \mathcal{G} and we call T the *lifetime* of \mathcal{G} . We remark that all definitions and results in this work can easily be extended to labeled temporal graphs (with vertex and/or edge labels).

A *vertex mapping* between two vertex sets V and W is a set $M \subseteq V \times W$ containing $\min(|V|, |W|)$ tuples such that each $x \in V \cup W$ is contained in at most one tuple of M . We denote the set of all vertex mappings between V and W by $\mathcal{M}(V, W)$. Let $V_M \subseteq V$ be the subset of vertices in V that are contained in some tuple of M ($W_M \subseteq W$ is defined analogously). Note that $V_M = V$ or $W_M = W$ holds since $|M| = \min(|V|, |W|)$. Computing optimal vertex mappings between two temporal graphs can be solved via the **ASSIGNMENT PROBLEM** which is a fundamental problem in combinatorial optimization: Given two sets A and B of equal size and a cost function $c: A \times B \rightarrow \mathbb{Q}$, the goal is to find a bijection $\pi: A \rightarrow B$ such that $\sum_{a \in A} c(a, \pi(a))$ is minimized. The **ASSIGNMENT PROBLEM** is solvable in $O(|A|^3)$ time [2, Theorem 12.2].

The *dynamic time warping* distance [16] is a distance between time series. It is based on the concept of a warping path. A *warping path* of order $n \times m$ is a set $p = \{p_1, \dots, p_L\}$ of $L \geq 1$ pairs $p_\ell = (i_\ell, j_\ell)$ such that $p_1 = (1, 1)$, $p_L = (n, m)$, and $p_{\ell+1} \in \{(i_\ell + 1, j_\ell + 1), (i_\ell, j_\ell + 1), (i_\ell + 1, j_\ell)\}$ for all $1 \leq \ell < L$. We denote the set of all warping paths of order $n \times m$ by $\mathcal{P}_{n,m}$. For two temporal graphs $\mathcal{G} = (V, E_1, \dots, E_T)$, $\mathcal{H} = (W, F_1, \dots, F_U)$, every warping path p of order $T \times U$ defines a *warping* between \mathcal{G} and \mathcal{H} , that is, a pair $(i, j) \in p$ *warps* the layer G_i to layer H_j .

3 Dynamic Temporal Graph Warping (DTGW)

In this section, we define a temporal graph distance based on dynamic time warping using a vertex-signature-based graph distance as cost function. We choose this graph distance for the following reasons: First, in contrast to the NP-hard edit distance, it is polynomial-time computable. Second, it is based on a mapping between the two vertex sets, which allows to enforce a consistency over time. This consistency is naturally required in many temporal network applications where the vertices in both networks correspond to the same set of objects over time (one can also easily drop this assumption if it is not desired). Note also that it implicitly allows to identify vertices within the two networks. Third, vertex signatures allow for a high flexibility since they can be chosen arbitrarily (as can the metric) in order to incorporate essential information (local or global) for the application at hand (e.g., one might use feature vectors obtained via network embeddings).

Graph Distance Based on Vertex Signatures. The following approach is due to Jouili and Tabbone [12]. For a (static) graph $G = (V, E)$, a *vertex signature function* $f_G: V \rightarrow \mathbb{Q}^k$ encodes arbitrary information about a vertex. Let $d: \mathbb{R}^k \times \mathbb{R}^k \rightarrow \mathbb{R}$ be a metric.

For two (static) graphs $G = (V, E)$ and $H = (W, F)$ with vertex signatures $f_G: V \rightarrow \mathbb{Q}^k$ and $f_H: W \rightarrow \mathbb{Q}^k$ and a given vertex mapping M between V and W , we define the *cost* of M as

$$C(G, H, M) := \sum_{(u, v) \in M} d(f_G(u), f_H(v)) + \sum_{v \in V \setminus V_M} \Delta_G(v) + \sum_{v \in W \setminus W_M} \Delta_H(v),$$

where $\Delta_G(v) \in \mathbb{Q}$ is the (predefined) cost of “deleting” vertex v from G since it is not mapped by M to any vertex in the other vertex set. The value $\Delta_G(v)$ might for example depend on the vertex signature of v . Note that “deleting” a vertex does not affect the signatures of other vertices. By definition of a vertex mapping, at least one of the last two sums on the right-hand side above is zero.

The vertex-signature-based distance between G and H is then defined as

$$D(G, H) := \min_{M \in \mathcal{M}(V, W)} C(G, H, M).$$

Depending on the application, one might normalize the distance D by some appropriate factor (typically depending on $|V|$ and $|W|$; for example, Jouili and Tabbone [12] normalize by $\min(|V|, |W|)^{-1}$).

Throughout this work, we assume that vertex signature functions f_G are computable in polynomial time in the size of G and we assume all metrics d to be polynomial-time computable. We neglect the running times for computing the values of f_G and d because we assume that all vertex signatures are precomputed once in polynomial time.

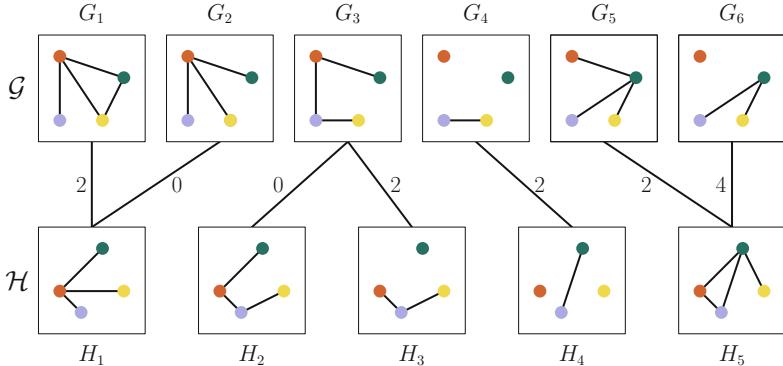


Fig. 1. Example of the dtgw-distance between two temporal graphs \mathcal{G} and \mathcal{H} on four vertices with lifetimes six and five. The vertex coloring indicates an optimal vertex mapping M . The connections between the boxes indicate an optimal warping path $p = \{(1, 1), (2, 1), (3, 2), (3, 3), (4, 4), (5, 5), (6, 5)\}$. The labels correspond to the costs $C(G_i, H_j, M)$ where the vertex signatures are their degrees and the metric is the absolute value of the difference. For example the costs of warping G_1 to H_1 is 2, as the green (darkest) and yellow (lightest) vertex each have degree two in G_1 but only degree one in H_1 . The resulting dtgw-distance is $\text{dtgw}(\mathcal{G}, \mathcal{H}) = 2 + 0 + 0 + 2 + 2 + 2 + 4 = 12$.

Dynamic Time Warping Distance for Temporal Graphs. We transfer the concept of dynamic time warping to temporal graphs in the following way. Let $\mathcal{G} = (V, E_1, \dots, E_T)$ and $\mathcal{H} = (W, F_1, \dots, F_U)$ be two temporal graphs and let $f_{G_1}, \dots, f_{G_T}: V \rightarrow \mathbb{Q}^k$ and $f_{H_1}, \dots, f_{H_U}: W \rightarrow \mathbb{Q}^k$ be corresponding vertex signature functions.

We define the vertex-signature-based *dynamic temporal graph warping distance* (dtgw-distance) between \mathcal{G} and \mathcal{H} as

$$\text{dtgw}(\mathcal{G}, \mathcal{H}) := \min_{M \in \mathcal{M}(V, W)} \min_{p \in \mathcal{P}_{T, U}} \sum_{(i, j) \in p} C(G_i, H_j, M).$$

Figure 1 depicts an example illustrating the dtgw-distance of two temporal graphs. Intuitively, the vertex mapping identifies vertices with similar behavior over time and the warping path identifies the time layers with similar vertex behavior. Note that (for $T = U$) if one fixes $p = \{(1, 1), (2, 2), \dots, (T, T)\}$, then we get a temporal graph distance without time warping (which is similar to the Euclidean distance).

The following facts are easily observed and play a central role for our subsequent algorithms.

Observation 1. Let $\mathcal{G} = (V, E_1, \dots, E_T)$ and $\mathcal{H} = (W, F_1, \dots, F_U)$ be two temporal graphs with $|V| \leq |W| =: n$.

- (i) For a fixed vertex mapping $M \in \mathcal{M}(V, W)$, a warping path $p \in \mathcal{P}_{T,U}$ minimizing $\sum_{(i,j) \in p} C(G_i, H_j, M)$ can be computed in $\mathcal{O}(T \cdot U \cdot n)$ time.
- (ii) For a fixed warping path $p \in \mathcal{P}_{T,U}$, a vertex mapping $M \in \mathcal{M}(V, W)$ minimizing $\sum_{(i,j) \in p} C(G_i, H_j, M)$ can be computed in $\mathcal{O}(n^2 \cdot |p| + n^3)$ time.

Proof. (i) For a given vertex mapping M , an optimal warping path can be computed by a well-known dynamic program for dynamic time warping in $\mathcal{O}(T \cdot U \cdot n)$ time [16]. Here, $\mathcal{O}(n)$ is the time required to compute the costs $C(G_i, H_j, M)$.

- (ii) Let $V' := V \cup Q$, where Q is a set of $|W| - |V|$ dummy vertices (that is, $|V'| = n$) with $Q \cap V = \emptyset$. For every $(u, v) \in V' \times W$, define

$$\sigma(u, v) := \begin{cases} \sum_{(i,j) \in p} d(f_{G_i}(u), f_{H_j}(v)), & u \in V \\ \sum_{(i,j) \in p} \Delta_{H_j}(v), & u \in Q \end{cases}.$$

Then, we need to find $M \in \mathcal{M}(V', W)$ that minimizes $\sum_{(u,v) \in M} \sigma(u, v)$. Note that the vertex mapping M defines a bijection between V' and W . Hence, computing M is an ASSIGNMENT PROBLEM instance solvable in $O(n^3)$ time [2, Theorem 12.2]. Computing all $\sigma(u, v)$ values can be done in $O(n^2 \cdot |p|)$ time. \square

Note that Observation 1 (i) implies that if we already know the vertex mapping up to a constant number of vertices, then dtgw can be computed in polynomial time (since we can try out all polynomially many possible vertex mappings). Furthermore, Observation 1 (ii) implies that dtgw is polynomial-time computable if the optimal temporal alignment between \mathcal{G} and \mathcal{H} is known beforehand.

For given vertex signature function and metric, we refer to the decision problem of testing whether two temporal graphs have dynamic temporal graph warping distance at most some given c by DTGW.

DYNAMIC TEMPORAL GRAPH WARPING (DTGW)

Input: Two temporal graphs \mathcal{G} and \mathcal{H} , $c \in \mathbb{Q}$.

Question: Is $\text{dtgw}(\mathcal{G}, \mathcal{H}) \leq c$?

By Observation 1, DTGW is polynomial-time solvable if one temporal graph has a constant lifetime or a constant number of vertices since there are only polynomially many possible warping paths or polynomially many vertex mappings.

4 Computational Hardness

Even though the dynamic time warping distance and the vertex-signature-based graph distance are both computable in polynomial time, their combined application to temporal graphs yields a distance measure that is generally NP-hard to compute. Intuitively, this is due to the requested consistency of the vertex mapping over all layers. This introduces non-trivial dependencies between the time warping and the vertex mapping which render the problem computationally hard.

Theorem 2. DTGW is NP-complete for every metric when the vertex signatures are vertex degrees.

The extensive and technical proof (see full version (Theorem 4.1)) is by a polynomial-time reduction from the NP-complete 3-SAT problem and allows to conclude even stronger hardness results summarized in the following corollary.

Corollary 3. DTGW is NP-complete for every metric and vertex degrees as vertex signatures even when the maximum degree of each layer is one and the warping path is restricted to lie within a band of width one around the diagonal.

Moreover, this case cannot be solved in $2^{o(|V|+|W|+T+U+c)} \cdot \text{poly}(|\mathcal{G}| + |\mathcal{H}|)$ time unless the Exponential Time Hypothesis¹ fails.

Due to this worst-case hardness of DTGW, there is little hope to solve the general problem efficiently. In the next section, however, we point out two polynomial-time solvable special cases (in addition to Observation 1) as well as a heuristic approach to approximate the dtgw-distance efficiently.

5 Algorithms

Our first algorithmic result is a polynomial-time algorithm deciding whether two temporal graphs with the same number of vertices have dtgw-distance zero. This basic task occurs when checking for duplicates within a data set. In contrast, determining whether two (static) graphs have graph edit distance zero is not known to be polynomial-time solvable (as this is equivalent to the famous GRAPH ISOMORPHISM problem).

Theorem 4. Let $\mathcal{G} = (V, E_1, \dots, E_T)$ and $\mathcal{H} = (W, F_1, \dots, F_U)$ be two temporal graphs with $|V| = |W| = n$. For all vertex signatures and all metrics, one can decide whether $\text{dtgw}(\mathcal{G}, \mathcal{H}) = 0$ in $\mathcal{O}(n^2 \cdot (T + U) + n^3)$ time.

Proof. We will show that for distance zero, an optimal warping path can easily be determined. Polynomial-time solvability then follows from Observation 1.

Let $\mathcal{G} = (V, E_1, \dots, E_T)$ and $\mathcal{H} = (W, F_1, \dots, F_U)$ be two temporal graphs with $V =: \{v_1, \dots, v_n\}$ and $W =: \{w_1, \dots, w_n\}$. For each $i \in [T]$, we define the i^{th} layer signature of \mathcal{G} as $f(G_i) := (f_{G_i}(v_1), \dots, f_{G_i}(v_n))$ (analogously, $f(H_j) := (f_{H_j}(w_1), \dots, f_{H_j}(w_n))$ for $j \in [U]$). Assuming $\text{dtgw}(\mathcal{G}, \mathcal{H}) = 0$, it follows that there exists a vertex mapping $M \subseteq V \times W$ and a warping path $p \in \mathcal{P}_{T,U}$ such that

$$\sum_{(u,v) \in M} d(f_{G_i}(u), f_{H_j}(v)) = 0$$

holds for every $(i, j) \in p$. Since d is a metric, this implies that $f_{G_i}(u) = f_{H_j}(v)$ holds for every $(u, v) \in M$. That is, $f(H_j)$ is a permutation (determined by M)

¹ The Exponential Time Hypothesis, an established concept in computational complexity theory, asserts that there is a constant $c > 0$ such that 3-SAT cannot be solved in $O(2^{cn})$ time, where n is the number of variables in the input formula [10].

of $f(G_i)$. Let $1 \leq i_1 < i_2 \dots < i_q < T$ and $1 \leq j_1 < j_2 < \dots < j_r < U$ be the indices such that

$$\begin{aligned} f(G_i) \neq f(G_{i+1}) &\iff i \in \{i_k : k \in [q]\} \text{ and} \\ f(H_j) \neq f(H_{j+1}) &\iff j \in \{j_k : k \in [r]\}. \end{aligned}$$

Clearly, if $f(G_i) \neq f(G_{i'})$ and layer i is warped to layer j and layer i' is warped to layer j' , then $f(H_j) \neq f(H_{j'})$ since otherwise the cost will not be zero. By the definition of a warping path, it follows that the layers $1, \dots, i_1$ of \mathcal{G} can only be warped to layers $1, \dots, j_1$ of \mathcal{H} and the layers $i_1 + 1, \dots, i_2$ of \mathcal{G} can only be warped to layers $j_1 + 1, \dots, j_2$ of \mathcal{H} and so on. Note that this is only possible if $q = r$. If this is the case, then we can assume that the warping path p has the following form:

$$\begin{aligned} p = \{(1, 1), (1, 2), \dots, (1, j_1), (2, j_1), \dots, (i_1, j_1), \\ (i_1 + 1, j_1 + 1), \dots, (i_1 + 1, j_2), \dots, (i_2, j_2), \\ \dots, \\ (i_q + 1, j_q + 1), \dots, (i_q + 1, U), \dots, (T, U)\}. \end{aligned}$$

By Observation 1 (ii), we can now check whether there exists a vertex mapping that yields distance zero for the warping path p in $\mathcal{O}(n^2 \cdot (T + U) + n^3)$ time. Computing p can be done in $\mathcal{O}(n(T + U))$ time. \square

We remark that if the vertex signatures and the metric satisfy the property that every pair of different vertex signatures has distance at least δ for some constant $\delta > 0$, then DTGW is polynomial-time solvable for any constant cost bound c . For example, this is the case when the vertex signatures contain only integers and d is any p -norm (for $p \geq 1$). Then, every pair of different signatures has distance at least $\delta = 1$. The idea of the algorithm is to “guess” the tuples of a warping path which cause non-zero cost (at most c/δ many) and to check whether it is possible to complete the warping path without further costs. The latter can be done in polynomial time using similar arguments as for the case $c = 0$ (Theorem 4).

In contrast, if the dtgw-distance is normalized (e.g. divided by the number of vertices), then the differences between vertex signatures can be arbitrarily small. In that case, DTGW is NP-complete even for a constant value of c .

To overcome this hardness, we consider special cases based on parameters regarding the warping path length. We assume that the lifetimes of the inputs differ by at most a constant, that is, $T = U + t$ for some $t \geq 0$ (in practice, $t = 0$ might often be the case). Note that, by definition, every warping path of order $T \times U$ has length at least T . We define the parameter λ to be the difference between the warping path length and the lower bound T , that is, we consider only warping paths of order $T \times U$ and length at most $T + \lambda$ (in practice, long warping paths are often considered unnatural). The following theorem implies polynomial-time solvability of DTGW if t and λ are constants.

Theorem 5. For all vertex signatures and all metrics, DTGW is solvable in

$$O((T + \lambda)^\lambda \cdot T^{\lambda+t} (n^2 \cdot (T + \lambda) + n^3))$$

time if $n = \max(|V|, |W|)$, $T = U + t$, and the warping paths have length at most $T + \lambda$.

For unbounded t , we conjecture that DTGW is NP-hard even if $\lambda = 0$.

Finally, we present a heuristic to compute the dtgw-distance. The idea is to start with some initial warping path (or vertex mapping) and to compute an optimal vertex mapping (warping path) based on Observation 1 in polynomial time. This process is then repeated by alternating between optimal warping path and optimal vertex mapping computation until the solution converges to a (local) minimum (or some other criterion is reached). We call this majorize-minimize approach *alternating minimization* (AM). Convergence is guaranteed since we decrease the objective in each alternation and the search space is finite. We propose to initialize with a shortest warping path (that is, of length $\max(T, U)$).²

The running time of one iteration (that is, computing a vertex mapping and an optimal warping path) is $O((T + U) \cdot n^2 + n^3 + T \cdot U \cdot n)$. While the number of iterations might depend on the choice of initialization, in our experiments the heuristic always converged after very few iterations. Regarding the solution quality, it is possible to construct adversarial examples where the heuristic performs arbitrarily bad. In practice, however, it performs well (see full paper).

6 Experiments

In the full paper, we evaluated the performance of the AM heuristic we described in Sect. 5. The experiments indicate that the AM heuristic is very fast and usually finds close to optimal solutions. Moreover, we conducted a clustering experiment where the dtgw-distance successfully recovered original data from noise.

We demonstrate that the dtgw-distance can be used for de-anonymization of temporal social networks. Recall that besides measuring a distance between temporal graphs, the dtgw-distance additionally provides a mapping between the vertex sets which implicitly allows to identify vertices.

Data. We used three data sets from the SocioPatterns collaboration [7]. Each of these consists of two temporal social networks, both recorded simultaneously with the same individuals as vertices. The first network is a face-to-face contact network whereas the second one is a co-presence network where edges represent spatial proximity. All six networks have a temporal resolution of 20s. For our experiments, only the first day of each network was used and isolated vertices were discarded. The three different data sets were recorded at a primary school (“LyonSchool”, 237 vertices, 1700 layers), a scientific conference (“SFHH”, 403 vertices, 2300 layers) and a workplace (“InVS15”, 180 vertices, 2100 layers).

² In the full arXiv version we discuss several alternative initializations, all of which performed comparably well in experiments. Notably, initializing with a shortest warping path is the fastest initialization.

Setup. We used the AM heuristic to compute the dtgw-distance (with degrees as vertex signatures³) on all three data sets. For computations we used a 4.0 GHz i7-6700K processor (single-threaded). We implemented the AM heuristic in Python,⁴ using an existing C++ implementation solving the ASSIGNMENT PROBLEM.⁵ We counted how many vertices were correctly re-identified (that is, mapped to their copies) in the resulting vertex mapping. We compared our results to the following alternatives:

- DynaMAGNA++ [17]: A search-based evolutionary algorithm computing a vertex mapping that maximizes edge conservation and vertex conservation over time.
- Temporal Network Embedding [18]: Hawkes process-based Temporal Network Embedding (HTNE) computes a low-dimensional embedding of the vertices of a temporal network. We compute a vertex mapping minimizing the Euclidean distances between these vertex feature vectors.
- Fixed dtgw: Note that our dtgw-distance allows to fix the warping path beforehand (Observation 1 (ii)). Since in our case each pair of temporal graphs was recorded using synchronized clocks, it is natural to use a fixed warping path that aligns layer i with layer i .

To simulate a situation in which the temporal graphs represent processes which do not run synchronously in time, we created two modified versions of each of the data sets. In the first one, called “shifted”, all events of the first graph were delayed by 3 min. In the second version, called “randomized”, each layer of each of the graphs was randomly and independently replaced by X layers where $X \in \{1, 2, \dots\}$ is a random variable with $\mathbb{P}(X \geq x) = x^{-3}$. Since we pretend that the nature of these modifications is unknown to the tested algorithms, dtgw with fixed warping path is not applicable to these variants.

For DynaMAGNA++, we used a population of size 15 000 and a maximum of 10 000 generations. With HTNE, we computed 128-dimensional vertex embeddings using a batch of size 10 000, a learning rate of 0.1, a history length of 2, and five negative samples. Unlike dtgw, both methods utilized all four processor cores (GPU-based computation was not available).

Results. The results and running times are listed in Table 1. Most notably, the re-identification rate of HTNE was poor on all data sets, suggesting that these embeddings are ill-suited for comparing vertices taken from different networks. Furthermore, all methods failed to re-identify any significant number of vertices on the primary school data set. This might be explained by the fact that the co-presence network is very different from the face-to-face contacts due to a low spatial resolution (as was also noted by Génois and Barrat [7]).

³ We also tested other signatures such as size of the connected component or betweenness centrality. However, the performance was (slightly) worse.

⁴ Source code available at www.akt.tu-berlin.de/menue/software.

⁵ Available as a Python module at [www.https://github.com/src-d/lapjv](https://github.com/src-d/lapjv).

Table 1. Percentages (rounded) of vertices that were re-identified by the tested methods. Also average running times (in seconds) over the three versions of each data set are given.

	Data set	dtgw	Fixed dtgw	DynaMAGNA++	HTNE
School	Original	2	1	1	1
	Shifted	1	–	0	1
	Randomized	1	–	1	0
	Average running time	95 s	65 s	15 070 s	250 s
Conference	Original	86	90	80	0
	Shifted	86	–	27	0
	Randomized	65	–	30	1
	Average running time	200 s	125 s	20 320 s	90 s
Workplace	Original	38	43	51	1
	Shifted	38	–	19	0
	Randomized	10	–	8	1
	Average running time	45 s	20 s	1 600 s	50 s

The overall performance was much better on the other two data sets, especially on the conference data where up to 90% of participants could be re-identified whereas on the workplace data set the best result was 51%. Unsurprisingly, fixing the correct layer alignment on the unmodified graphs sped up the dtgw computation significantly while also yielding slightly better results. On these instances, dtgw performed comparably to DynaMAGNA++, being better in one case and worse in the other, although requiring much less computational effort. In contrast, on the shifted and randomized data sets dtgw always achieved the best results (notably the performance of dtgw did not decrease on shifted data). In all cases the AM heuristic converged after at most six iterations and DynaMAGNA++ converged within 2 000 generations.

7 Conclusion

We introduced a new similarity measure for comparing temporal graphs by transferring dynamic time warping to temporal graphs. This yields a challenging computational problem for which we proposed exact algorithms and a heuristic approach which runs fast in practice and yields good approximations of optimal solutions. As a showcase, we demonstrated that our method is capable of de-anonymizing social networks.

Our work opens several directions for future research. We believe that the dtgw-distance is a promising tool for example in biology and chemistry. Processes like epidemic disease spread or chemical reactions can naturally be viewed as temporal graphs where the vertices represent individuals or (macro-)molecules (unfortunately we could not test this, as there is still a lack of openly available

temporal molecular data [17]). Since the exact time scales of these processes often vary, the ability of dynamic time warping to compensate for such differences could be especially helpful in this context.

References

1. Abanda, A., Mori, U., Lozano, J.A.: A review on distance based time series classification. *Data Min. Knowl. Disc.* **33**(2), 378–412 (2019)
2. Ahuja, R.K., Magnanti, T.L., Orlin, J.B.: *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Upper Saddle River (1993)
3. Bagavathi, A., Krishnan, S.: Multi-Net: a scalable multiplex network embedding framework. In: *Complex Networks and Their Applications VII*. SCI, vol. 813, pp. 119–131. Springer (2019)
4. Braha, D., Bar-Yam, Y.: From centrality to temporary fame: dynamic centrality in complex networks. *Complexity* **12**(2), 59–63 (2006)
5. Elhesha, R., Sarkar, A., Boucher, C., Kahveci, T.: Identification of co-evolving temporal networks. In: *Proceedings of BCB 2018*, pp. 591–592. ACM (2018)
6. Fröhlich, H., Wegner, J.K., Sieker, F., Zell, A.: Optimal assignment kernels for attributed molecular graphs. In: *Proceedings of ICML 2005*, pp. 225–232. ACM (2005)
7. Génois, M., Barrat, A.: Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Sci.* **7**(1), 11 (2018)
8. Ghosh, S., Das, N., Gonçalves, T., Quaresma, P., Kundu, M.: The journey of graph kernels through two decades. *Comput. Sci. Rev.* **27**, 88–111 (2018)
9. Holme, P., Saramäki, J.: Temporal networks. *Phys. Rep.* **519**(3), 97–125 (2012)
10. Impagliazzo, R., Paturi, R.: On the complexity of k -SAT. *J. Comput. Syst. Sci.* **62**(2), 367–375 (2001)
11. Jain, B.J.: On the geometry of graph spaces. *Discrete Appl. Math.* **214**, 126–144 (2016)
12. Jouili, S., Tabbone, S.: Graph matching based on node signatures. In: *Proceedings of GbRPR 2009*. LNCS, vol. 5534, pp. 154–163. Springer (2009)
13. Kostakos, V.: Temporal graphs. *Phys. A* **388**(6), 1007–1023 (2009)
14. Li, A., Cornelius, S.P., Liu, Y.Y., Wang, L., Barabási, A.L.: The fundamental advantages of temporal networks. *Science* **358**(6366), 1042–1046 (2017)
15. Riesen, K.: Structural pattern recognition with graph edit distance. Springer (2015)
16. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust. Speech* **26**(1), 43–49 (1978)
17. Vijayan, V., Critchlow, D., Milenković, T.: Alignment of dynamic networks. *Bioinformatics* **33**(14), i180–i189 (2017)
18. Zuo, Y., Liu, G., Lin, H., Guo, J., Hu, X., Wu, J.: Embedding temporal network via neighborhood formation. In: *Proceedings of KDD 2018*, pp. 2857–2866. ACM (2018)



Maximizing the Likelihood of Detecting Outbreaks in Temporal Networks

Martin Sterchi^{1,2,3(✉)}, Cristina Sarasua¹, Rolf Grüter²,
and Abraham Bernstein¹

¹ University of Zurich, Zurich, Switzerland
sterchi@ifi.uzh.ch

² Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
³ University of Applied Sciences and Arts Northwestern Switzerland FHNW,
Olten, Switzerland

Abstract. Epidemic spreading occurs among animals, humans, or computers and causes substantial societal, personal, or economic losses if left undetected. Based on known temporal contact networks, we propose an outbreak detection method that identifies a small set of nodes such that the likelihood of detecting recent outbreaks is maximal. The two-step procedure involves (i) simulating spreading scenarios from all possible seed configurations and (ii) greedily selecting nodes for monitoring in order to maximize the detection likelihood. We find that the detection likelihood is a submodular set function for which it has been proven that greedy optimization attains at least 63% of the optimal (intractable) solution. The results show that the proposed method detects more outbreaks than benchmark methods suggested recently and is robust against badly chosen parameters. In addition, our method can be used for outbreak source detection. A limitation of this method is its heavy use of computational resources. However, for large graphs the method could be easily parallelized.

Keywords: Temporal networks · Epidemic spreading · Outbreak detection · Source detection · Submodular set functions · Greedy optimization

1 Introduction

Spreading processes on networks can occur in various contexts, such as infectious disease spreading among humans or animals [3, 14], computer virus spreading over the internet [13], or misinformation spreading in online social networks [5]. In those examples, the spreading process can have disastrous consequences and stopping it effectively is of great importance. For example, if the spread of a new infectious disease remains undetected, it can grow into a global pandemic. Likewise, the undetected spread of misinformation, can have significant financial or political consequences. In some cases, however, the spreading phenomenon

can be seen as desirable, such as in the case of viral marketing [8]. Whether desired or harmful, the mentioned examples are conceptually similar and share the idea that a network of physical or virtual contacts acts as the substrate for the spreading process. While much of the past research considers static contact networks, a recent surge in research focusing on temporal contact networks sheds light on the importance of temporal structure for spreading processes (e.g., [14]).

Several attempts have been made to select nodes for optimal outbreak detection. Most notably, in their seminal work, Leskovec et al. [10] propose a near-optimal outbreak detection strategy that is based on selecting a small set of nodes for monitoring using greedy optimization. Leskovec et al. focus on three optimization objectives: detection likelihood, time until detection, and the population that is affected by an outbreak. While their approach works well in a static water distribution network, where edges do not change over time, it presents a critical shortcoming when applied to a temporal blog network where bloggers post and repost stories. In that case, Leskovec et al. identify nodes for sensor placement based on past (observed) data which are then used to detect future outbreaks. However, the generalizability to future data can be unsatisfactory, especially if the topology of the underlying temporal network is changing rapidly [3, 10]. Another study suggests that central nodes tend to be infected sooner and, therefore, monitoring them may be an effective strategy [6]. Finally, Bajardi et al. [3] suggest monitoring so called sentinel nodes, i.e., nodes exhibiting both a large probability of being infected and little uncertainty about the seed node of the outbreak. The last two methods may lead to unsatisfactory results as they do not involve optimizing the set of nodes selected but are instead based on heuristics.

In this paper, we investigate the problem of outbreak detection for epidemic spreading in *temporal* contact networks. The central premise of this work is that an outbreak can start at any time and from any node within a certain time period. On a given day, we aim to identify a (small) set of k nodes for monitoring such that the probability of detecting an outbreak that started anywhere within the past b days is maximal. Since monitoring resources are typically scarce, the optimal solution of monitoring all nodes in the network is not feasible. Hence, k is set such that it matches the available monitoring resources. Our approach can be divided into two steps: (i) extensive Monte-Carlo simulations of outbreak scenarios for every possible seed configuration in the window over the past b days using a propagation model and (ii) greedy optimization of the detection likelihood. An optional third step extends our method to the problem of outbreak source detection (e.g. [2]). Here, we use the stochastic version of the well-known *susceptible-infected-susceptible* (SIS) model [4] as the propagation model but our approach can be used with any propagation model. Our contributions can then be summarized as follows:

- We introduce a novel method for outbreak detection in temporal networks that combines extensive outbreak simulations and greedy optimization in order to maximize the likelihood of detecting recent outbreaks (Sect. 4). We show that the method extends to the problem of source detection (Sect. 4.3).

- We show that the detection likelihood of a set of nodes is a submodular set function for which greedy optimization is guaranteed to achieve at least 63% of the optimal solution (Sect. 4.2).
- Finally, we evaluate our method on two temporal networks: an undirected network describing sexual contacts in Brazil [14] and a directed network describing pig movements in Switzerland. We show that our method outperforms previously suggested heuristics [3, 6]. Moreover, we provide evidence that it is robust against badly chosen parameters (Sect. 5).

2 Related Work

One of the early works suggesting greedy optimization of submodular set functions for optimal node selection is Kempe et al. [8], who consider the problem of maximizing the spread of influence in social networks. This work has been extended to dynamic networks by Aggarwal et al. [1] and more recently, the use of real diffusion cascades instead of simulated ones has been suggested [12]. The problem of influence maximization is conceptually similar to outbreak detection. In the case of influence maximization, the goal is to select nodes such that the size of the spreading cascades originating from those nodes is maximal, whereas for outbreak detection, the goal is to select nodes such that we catch as many spreading cascades as possible. Much of the research in the field of outbreak detection goes back to the seminal work by Leskovec et al. [10], who not only optimize *detection likelihood*, but also *detection time* and the *share of the population that is affected by an outbreak*. Our approach differs from Leskovec et al.’s approach [10] in that we optimize the set of nodes that we can monitor on a given day, whereas Leskovec et al. optimize the set of nodes without reference to a given day. As a consequence, our approach does not apply to other optimization goals, such as the detection time. The problem of outbreak detection has also been addressed in more heuristic ways. In [6], the authors make use of the observation that central individuals are topologically closer to the average individual in a network and are thus more likely to be infected early. This provides the rationale for monitoring high (in-)degree individuals. Bajardi et al. [3] propose sentinel nodes, i.e., nodes that are reached by outbreaks from many different initial conditions. However, in contrast to our work, this approach assumes that we know when an outbreak started and thus fails to address the uncertainty about the temporal origin of an outbreak.

3 Problem Definition

Imagine that we organize information about contacts between individuals as a directed or undirected time-stamped network $G = (V, E)$ where V denotes the set of nodes (individuals) and E the set of edges. An edge triple $(v_i, v_j, t) \in E$ consists of the two nodes $v_i, v_j \in V$ and a time-stamp t indicating when the contact happened. Note that in a directed network, the edge is directed from v_i to v_j . We suspect that one single node introduced a disease that now spreads

along edges in the network. Furthermore, we assume that we know the underlying propagation model. Here, we use the stochastic version of the SIS model [4] with a probability p that a susceptible node gets infected upon contact with an infected node and the time until recovery μ^{-1} that indicates the number of time steps until an infected individual recovers and becomes susceptible again. Our goal is to identify an optimal set of nodes $\mathcal{S}_t \subseteq V$ at the time of monitoring t such that the likelihood of detecting an outbreak that originated within a window over the past b days is maximal. Since monitoring resources are typically restricted, we introduce a maximal number of nodes k that can be monitored at a time. Therefore, the optimization problem becomes:

$$\max_{\mathcal{S}_t \subseteq V} DL(\mathcal{S}_t) \quad \text{subject to } |\mathcal{S}_t| \leq k, \quad (1)$$

where $DL(\mathcal{S}_t)$ denotes the detection likelihood associated with the set of nodes \mathcal{S}_t , i.e., $DL: \mathcal{S}_t \rightarrow [0, 1]$. If we define $A(s)$ to be the event that the node $s \in \mathcal{S}_t$ detects the outbreak, we can write the detection likelihood as $DL(\mathcal{S}_t) = P(\cup_{s \in \mathcal{S}_t} A(s))$. In other words, the detection likelihood corresponds to the probability that at least one of the nodes $s \in \mathcal{S}_t$ detects an outbreak (cf. Fig. 1 for an example). We use the simplifying assumption that we detect an outbreak if at least one of the nodes $s \in \mathcal{S}_t$ is infected at the date of monitoring.

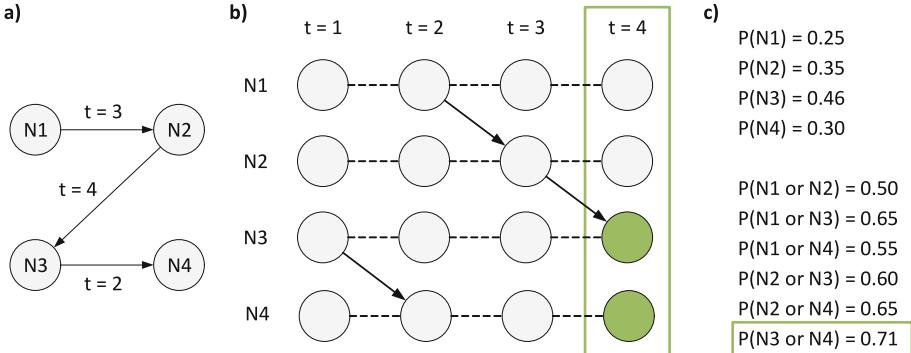


Fig. 1. (a) Simple directed network with 4 nodes and 3 time-stamped edges. (b) The same network but now time-unrolled. The goal is to identify the optimal 2 nodes for monitoring at $t = 4$. We assume a disease spreads as a *susceptible-infected* (SI) process (individuals do not recover) with $p = 0.6$. The outbreak can happen anywhere within the time period $t = 1, 2, 3$ ($b = 3$). Assuming that all seed configurations are equally likely to be the source, we get the detection probabilities in (c). The node with the highest probability to “see” an infection at $t = 4$ is N3. The optimal 2 nodes for outbreak detection are N3 and N4.

4 Proposed Solution

Our solution first simulates outbreaks from every possible seed configuration (cf. Sect. 4.1) and then uses the simulation results to optimally allocate monitoring resources to nodes (cf. Sect. 4.2). Optionally, a last step identifies the source of the outbreak (cf. Sect. 4.3). Here, we discuss these steps in turn.

4.1 Step 1: Spreading Simulations

The first step of the solution consists of simulating outbreaks from every possible seed node v_0 at every possible starting time t_0 within the window of length b . For the simulation, we use the known parameters of the underlying propagation process. As mentioned before, here, we focus on the SIS model with infection probability p and recovery time μ^{-1} . As a result, we get a large number n of outbreak outcomes for every possible pair $\{v_0, t_0\}$, which can be used to approximate the probability that a certain seed configuration $\{v_0, t_0\}$ infects at least one of the nodes in \mathcal{S}_t , i.e., $P(\cup_{s \in \mathcal{S}_t} A(s) \mid \{v_0, t_0\})$. In particular, we simply count the number of simulation runs that infect at least one $s \in \mathcal{S}_t$ and divide it by n . Note that for the sake of notation, we do not condition on the given parameters of the SIS process. By marginalizing out the seed configurations, we can compute the unconditional probability of detecting an outbreak, i.e.,

$$\begin{aligned} DL(\mathcal{S}_t) &= P(\cup_{s \in \mathcal{S}_t} A(s)) \\ &= \sum_{\{v_0, t_0\}} P(\cup_{s \in \mathcal{S}_t} A(s) \mid \{v_0, t_0\}) P(\{v_0, t_0\}). \end{aligned} \quad (2)$$

The running time for the simulation step is $\mathcal{O}(n b |V|)$ assuming the individual simulation has cost 1. With $|V|$ given by the network and the length of the window b determined by epidemic characteristics, we need to choose the number of simulations n such that the procedure is computationally feasible and results in satisfactory approximations of the conditional probabilities. In order to ease the greedy optimization step (Sect. 4.2), we transform the simulation results into an inverted index, a data structure that facilitates fast lookups [10]. Also note that if certain outbreak configurations $\{v_0, t_0\}$ are more likely than others, we can adjust the prior $P(\{v_0, t_0\})$ accordingly. By default, we use a uniform prior.

4.2 Step 2: Optimal Selection of Nodes for Monitoring

As stated in Sect. 3, we aim to identify a set of k nodes such that the detection likelihood is maximal. Solving such an optimization is NP-hard since the number of possible sets \mathcal{S}_t is prohibitively large [9]. For example, finding the optimal 10 nodes in a small network of 50 nodes, would require evaluating over 10 billion different sets of nodes. However, we can make use of a famous result by Nemhauser and Wolsey [9, 11], which states that for non-negative monotone and submodular set functions, a set selected by greedy optimization is within 63% of the optimal but intractable solution.

Monotonicity of a set function F is defined as follows. For any two sets of nodes \mathcal{A} and \mathcal{B} with $\mathcal{A} \subseteq \mathcal{B} \subseteq V$, it holds that $F(\mathcal{A}) \leq F(\mathcal{B})$. Moreover, F is submodular if it satisfies

$$F(\mathcal{A} \cup \{v\}) - F(\mathcal{A}) \geq F(\mathcal{B} \cup \{v\}) - F(\mathcal{B}). \quad (3)$$

for $v \in V \setminus \mathcal{B}$. Intuitively, this means that adding a new node v to a smaller set \mathcal{A} results in a marginal gain at least as large as the one resulting from adding v to \mathcal{B} .

Theorem 1. *The set function $DL(\mathcal{S}_t)$ is monotone and submodular.*

Proof. As a probability, $DL(\mathcal{S}_t)$ is also a measure and thus monotone [15]. For submodularity, we first note that a non-negative linear combination of submodular set functions is submodular [9]. From (2), we can see that $DL(\mathcal{S}_t)$ corresponds to a linear combination of probabilities $P(\cup_{s \in \mathcal{S}_t} A(s) \mid \{v_0, t_0\})$ with weights $P(\{v_0, t_0\}) \geq 0$. It thus remains to show that the probability $P(\cup_{s \in \mathcal{S}_t} A(s) \mid \{v_0, t_0\})$ is submodular. As mentioned in Sect. 4.1, this probability is estimated by counting how many of the n simulation runs (or cascades) infect at least one $s \in \mathcal{S}_t$. For the sake of notation, we will denote $P(\cup_{s \in \mathcal{S}_t} A(s) \mid \{v_0, t_0\})$ as F . Consider two sets of nodes $\mathcal{A} \subseteq \mathcal{B} \subseteq V$ and a node $v \in V \setminus \mathcal{B}$. We can distinguish three cases: (i) the new node v either detects no new cascades or only detects cascades that are already detected by \mathcal{A} and \mathcal{B} . In that case, $F(\mathcal{A} \cup \{v\}) - F(\mathcal{A}) = 0 = F(\mathcal{B} \cup \{v\}) - F(\mathcal{B})$. (ii) the new node v detects cascades that \mathcal{B} detects, but not \mathcal{A} . Then, $F(\mathcal{A} \cup \{v\}) - F(\mathcal{A}) = F(\{v\}) \geq 0 = F(\mathcal{B} \cup \{v\}) - F(\mathcal{B})$. (iii) the new node v detects cascades that neither \mathcal{B} nor \mathcal{A} detects. In that case, $F(\mathcal{A} \cup \{v\}) - F(\mathcal{A}) = F(\{v\}) = F(\mathcal{B} \cup \{v\}) - F(\mathcal{B})$. Hence, in all possible cases the inequality in (3) is satisfied. This concludes the proof. \square

Based on Theorem 1, we can use the greedy optimization algorithm that simply consists of iteratively adding the node that maximizes the marginal increase in detection likelihood to the initially empty set \mathcal{S}_t . The optimal set is found when $|\mathcal{S}_t| = k$. The complexity of the greedy optimization procedure is $\mathcal{O}(k|V|)$ if the evaluation of the set function is considered a constant operation. Note that we can improve the scalability of the algorithm with lazy forward evaluations that have been suggested in [10].

4.3 Step 3 (Optional): Outbreak Source Detection

Based on the simulation results from Sect. 4.1 and a set of observed infected nodes, we can compute the source probability for every seed configuration. If we denote the set of observed infected nodes at time t as \mathcal{I}_t and we want to compute the probability that $\{v_0^*, t_0^*\}$ was the source of this outbreak, we can use Bayes' rule as follows,

$$P(\{v_0^*, t_0^*\} \mid \mathcal{I}_t) = \frac{P(\mathcal{I}_t \mid \{v_0^*, t_0^*\}) P(\{v_0^*, t_0^*\})}{\sum_{\{v_0, t_0\}} P(\mathcal{I}_t \mid \{v_0, t_0\}) P(\{v_0, t_0\})}. \quad (4)$$

$P(\mathcal{I}_t \mid \{v_0, t_0\})$ can be easily approximated by the relative number of simulation runs that infect all nodes in \mathcal{I}_t for a given seed configuration $\{v_0, t_0\}$. Typically, we would assume a uniform prior $P(\{v_0, t_0\})$, i.e., all seed configurations are equally likely a priori. However, it is possible to have different priors for different seed configurations. For example, we may know that more recent outbreaks are more likely and, therefore, we set higher prior probabilities for more recent seed configurations. Or, in case of an animal transport network, we surmise that bigger farms are more likely to start an outbreak than smaller ones.

5 Evaluation

The primary goal of our evaluation is to show that our method is more effective than previously suggested methods and that it can be applied in practical settings where, for example, wrong parameters are chosen for the propagation model. We operationalize these claims with the following hypotheses:

- **H1.** Our method detects a higher fraction of outbreaks than state-of-the-art methods. We specifically compare with methods that are based on central nodes [6] or sentinel nodes [3], or that use past contact data to detect future outbreaks [10].
- **H2.** Our method is robust against badly chosen parameters for the spreading model.

In the remainder of this section, we introduce the datasets used in the evaluation and then discuss the empirical results for each hypothesis.

5.1 Datasets

We evaluate our method with two different datasets. The first dataset—widely used in work on temporal contact networks—describes time-stamped sexual contacts between male sex-buyers and female sex-sellers [14] and thus corresponds to an undirected bipartite network. We assume that this network acts as the substrate for a spread of sexually transmitted diseases (STDs). The full dataset covers a period of 6 years between September 2002 and October 2008. However, in what follows, we only show the results for the last 30 days of the dataset (October 2008) such that $b = 30$. The results for other 30-day periods are similar.¹ The 30-day network we analyze consists of 1,573 nodes and 1,463 edges. Ignoring the temporal dimension, we find that the network consists of 257 connected components and is thus highly fragmented. The density is 0.1131%. We refer to this network as the *escort network*.

The second dataset describes pig movements in Switzerland [16].² Movements are directed and time-stamped with the day of transport. Movements to slaughterhouses are removed and we only consider the farm-to-farm network. Moreover,

¹ We provide the dataset, the code, and additional results and figures under the following link: <https://github.com/martinSter/Outbreak-Detection>.

² The pig movement data contain private information and cannot be shared publicly. For research purposes, a data request can be sent to Identitas AG, Stauffacherstrasse 130A, 3014 Bern, Switzerland.

as in [3], we ignore within-farm dynamics and simply consider a movement as a directed contact between two farms. As with the first dataset, we consider a window of $b = 30$ days (October 2017) since this typically corresponds to the silent spread phase, i.e., the time between introduction of a disease and its first detection [7]. This 30-day network consists of 3,055 nodes 4,048 edges. Compared to the first dataset, the density is 0.0360% and thus lower which is partly due to the directionality of the second network. We refer to this network as the *pig network*.

5.2 Hypothesis H1 (Comparison with Benchmarks)

Benchmark Methods. In order to test $H1$, we use two benchmark methods that have been suggested in the related work [3, 6] and we additionally use a random set of nodes as a baseline. The first benchmark method corresponds to selecting nodes based on their *(in-)degree*. As stated in [6], we expect more central nodes, i.e., nodes with a high (in-)degree, to be good indicators for detecting outbreaks. The second benchmark method [3] identifies so called *sentinels*, i.e., nodes that are reached by many different infection paths and that exhibit a low degree of uncertainty with respect to the seed cluster. The method consists of two key steps. First, a deterministic SIS process with infection probability $p = 1$ and recovery time μ^{-1} yields all possible infection paths starting from any node at some time t_0 . In a second step, all seed nodes with similar infection paths, measured by the Jaccard index (threshold of 0.8 as in [3]), are grouped into seed clusters. We adjust their approach to our setting by using a SIS instead of a SIR model. We expect the method of Bajardi et al. [3] to perform worse than our method for three reasons: (i) their method only considers outbreaks starting at the beginning of the considered period and thus fails to address the fact that an outbreak can start at any time, (ii) their method trades some of the outbreak detection power for lower uncertainty about the seed of the outbreak, and (iii) their method is heuristic and does not maximize an objective function. Since their method does not uniquely focus on outbreak detection, we adjust it slightly and define sentinels as the nodes that are reached most often by infection paths, hereby neglecting the uncertainty criteria.

Methodology. In order to test our method, we stochastically simulate 10,000 outbreak scenarios with a random seed node and starting time within the 30 days considered. We assume that the disease propagates according to a SIS model with known infection probability $p = 0.6$ and recovery time $\mu^{-1} = 15$ days. The same parameter values are used in Step 1 of our method to simulate $n = 1,000$ spreading simulations from every seed configuration $\{v_0, t_0\}$. Since both networks have a temporal resolution of one day, the order of contacts of an individual on a given day is not known and, therefore, we assume that there is no same-day-spread from individuals with multiple contacts on a given day. With the greedy optimization procedure outlined in Step 2 of our method, we find k optimal nodes for monitoring. To compare the different methods, we count the number of detected outbreaks for varying k . Our method is implemented in Python

3.6 and is executed on a Intel Core i7 machine (1.80 GHz) with 16 GB RAM. The execution time is about 22 min for both the escort and pig network. Note however, that this may vary depending on the parametrization of the propagation model.

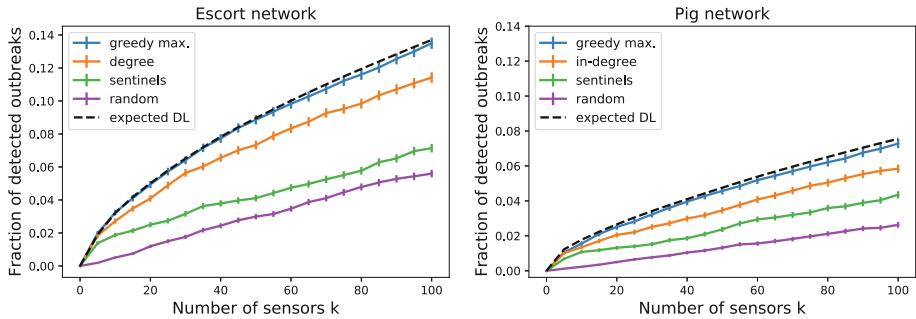


Fig. 2. Fraction of detected outbreaks for the escort network (left) and the pig network (right). The error bars represent \pm one standard deviation. The disease propagates according to a SIS model with $p = 0.6$ and $\mu^{-1} = 15$ days.

Results. Figure 2 presents the fraction of detected outbreaks for both networks. It is apparent that our method (*greedy max.*) performs better compared to the different benchmark methods in both data sets (of different sizes) and coincides with the expected detection likelihood (*dashed line*), which can be computed by plugging the k optimal nodes into (2). If k is small, high (in-)degree nodes achieve a similar detection quality because our approach initially selects mostly high-degree nodes. For larger k , the curves diverge as the greedy selection approach allocates monitoring resources to more remote parts of the network in order to maximize detection likelihood. As expected, sentinel nodes detect substantially fewer outbreaks but still perform better than the random node set. However, note that even with $k = 100$ our approach results in a detection likelihood of only roughly 14% and 7% for the escort and the pig network, respectively. Two factors contribute to these results. First, as seen in Sect. 5.1, both networks are highly fragmented, making outbreak detection difficult. Second, depending on the concrete parametrization, there is a considerable fraction of outbreak scenarios that die out before they could even be detected at the monitoring date. If we skew our analysis to outbreaks that are still active at the monitoring date, the fraction of detected outbreaks increases to 20% in the escort network and to 13% in the pig network, both for $k = 100$. We can go even further and only consider outbreaks with a certain minimal size. To this end, we again simulate outbreak scenarios for the escort network based on a SIS model with $p = 0.6$ and $\mu^{-1} = 15$. For every minimal outbreak size, we simulate 3,000 scenarios and count the number of outbreaks that the different methods (with $k = 10$ nodes) detect. Figure 4 (left) shows that all methods detect larger outbreaks

more effectively in the escort network. Our method detects more than half of all outbreaks that involve at least 5 nodes at the time of monitoring. This makes sense for two reasons: First, the more nodes that are affected, the more likely it is that one of the nodes is in our set of monitored nodes. Second, larger outbreaks tend to happen in more connected parts of the network where our method is better at detecting outbreaks.

Next, we compare the optimal node sets from Fig. 2 with node sets that were selected based on 5 previous 30-day periods. If the contact patterns in the networks were roughly constant, we would expect a similar performance for node sets chosen based on previous data. However, Fig. 3 shows that the performance is generally worse if we use nodes selected based on previous 30-day periods. Interestingly, for the pig network the node set that is closest to the reference period (October 2017) is the one selected based on June 2017. Additionally considering the node set based on February 2017 reveals that the pig network may exhibit some 4-month seasonalities that are associated with the production cycle in the pig industry. Overall, however, those findings imply that contact patterns vary significantly and node sets chosen based on past data do not generalize well for the two networks considered here.

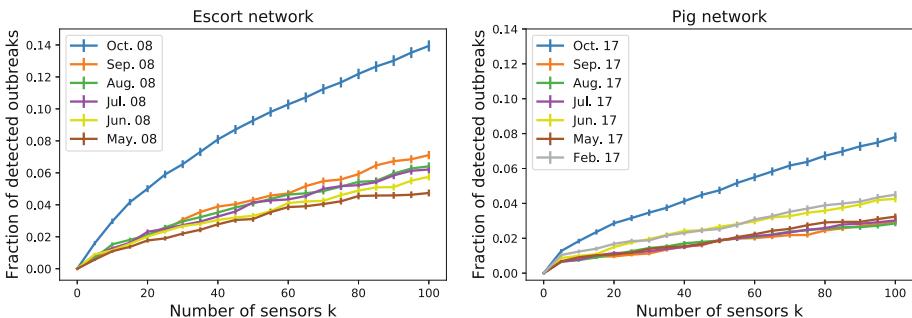


Fig. 3. Fraction of detected outbreaks with corresponding error bars for the escort network (left) and the pig network (right). The performance of the reference period (October) is compared to optimal node sets based on previous 30-day periods.

Finally, our method not only serves as an outbreak detection mechanism, it also detects the source of an outbreak given a sample of infected nodes at the time of monitoring. In order to test the source detection extension, we simulate 1,000 large outbreak scenarios (at least 8 infected nodes at the time of monitoring) according to the SIS model parametrized as above. We then randomly pick 1, 2, and 3 nodes from the simulated infection cascades as observed infected nodes and compute the posterior distribution according to (4), using a uniform prior. Note that for every possible source node, we sum the posterior probabilities over the different starting times in order to get an aggregated source probability per node. Figure 4 (right) shows the results. If we only observe one infected node, the accuracy of the prediction is unsatisfactory because in more than half of

all simulated outbreaks the true source is not ranked among the five top nodes. However, for two or three observed infected nodes, the accuracy of the source detection improves substantially. For example, we detect the source perfectly in more than 30% of all cases if we observe 3 infected nodes.

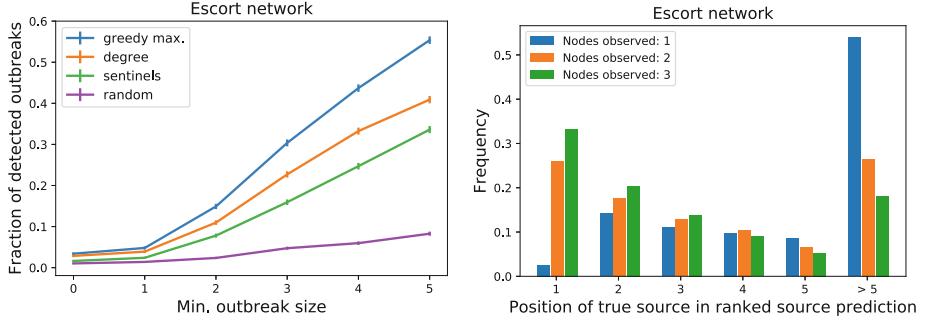


Fig. 4. Fraction of detected outbreaks with corresponding error bars for 10 nodes and varying minimal outbreak size (left). Source detection quality with varying numbers of nodes observed (right).

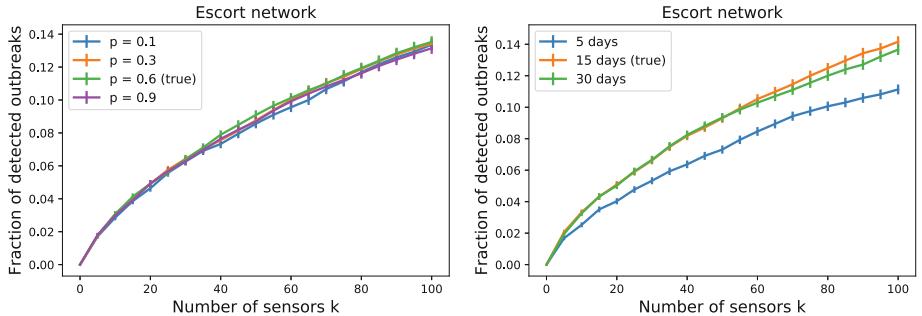


Fig. 5. Fraction of detected outbreaks with corresponding error bars for the escort network when the infection probability p (left) or the recovery period μ^{-1} (right) is misspecified.

5.3 Hypothesis H2 (Robustness)

Methodology. In order to test $H2$, we use our method (Step 1 and 2) to identify nodes with the true underlying infection probability $p = 0.6$ as well as with misspecified probabilities $p = \{0.1, 0.3, 0.9\}$, keeping $\mu^{-1} = 15$ fixed. Likewise, we identify monitoring nodes with the true underlying recovery time $\mu^{-1} = 15$ as well as with misspecified $\mu^{-1} = \{5, 30\}$, keeping $p = 0.6$ fixed. We simulate 10,000 outbreak scenarios based on the SIS model with the true underlying values of $p = 0.6$ and $\mu^{-1} = 15$ and compare the fraction of detected outbreaks.

Results. Our method seems to be robust against misspecified infection probabilities p . The fraction of detected outbreaks in the escort network is only slightly smaller if we use nodes found based on misspecified values of p (Fig. 5, left). However, the detection performance can deteriorate considerably if μ^{-1} is misspecified (Fig. 5, right). Note that the decrease in detection likelihood is more significant if μ^{-1} is smaller than the true value. This makes sense intuitively as assuming a too small μ^{-1} will neglect many possible infection paths.

6 Conclusions and Future Work

The goal of this work is to identify a small set of nodes for monitoring such that the likelihood of detecting recent outbreaks in temporal networks is maximized. Based on an evaluation with two different datasets, we find that our method outperforms other methods and it is especially effective for detecting large outbreaks. Moreover, it is robust with respect to the choice of parameters in the SIS model, as long as a large enough recovery time is chosen. Additionally, our method naturally extends to the problem of infection source detection. Although we restrict our examples to the SIS model, our method can be used with any propagation model that can be simulated. The findings reported here are crucial for the development of new outbreak detection strategies because our approach performs well in different contexts and applies to the realistic scenario where, on a given day, we need to find the optimal individuals to monitor. Moreover, our method is derived from basic concepts in probability theory rather than based on heuristics. The major limitation of our method compared to heuristics suggested previously is its computational intensity that can be prohibitive for large networks. However, the expensive simulation procedure could be easily parallelized. Another limitation of our method is that it only applies to the maximization of the detection likelihood. Other maximization objectives, such as the detection time, cannot be transferred easily to our problem. Further experiments, using different datasets and more complex propagation models, are an essential next step in confirming the generalizability of this method. Overall, we are convinced that our method can be an invaluable tool in a practical disease surveillance context.

Acknowledgement. This work was supported by the Swiss National Science Foundation (SNSF) NRP75, Project number 407540_167303. M. Sterchi was partially supported by the Hasler foundation. We would like to thank Identitas AG for providing the pig movement data and Emily E. Raubach, Heiko Nathues, Beat Hulliger, and the anonymous reviewers for helpful comments.

References

1. Aggarwal, C.C., Lin, S., Yu, P.S.: On influential node discovery in dynamic social networks. In: Proceedings of the 2012 SIAM International Conference on Data Mining, pp. 636–647 (2012)
2. Antulov-Fantulin, N., Lančić, A., Šmuc, T., Štefančić, H., Šikić, M.: Identification of patient zero in static and temporal networks: robustness and limitations. *Phys. Rev. Lett.* **114**, 248701 (2015)
3. Bajardi, P., Barrat, A., Savini, L., Colizza, V.: Optimizing surveillance for livestock disease spreading through animal movements. *J. R. Soc. Interface* **9**(76), 2814–2825 (2012)
4. Barrat, A., Barthélémy, M., Vespignani, A.: *Dynamical Processes on Complex Networks*, 1st edn. Cambridge University Press, New York (2008)
5. Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: Proceedings of the 20th International Conference on World Wide Web, pp. 665–674. ACM, New York (2011)
6. Christakis, N.A., Fowler, J.H.: Social network sensors for early detection of contagious outbreaks. *PLoS ONE* **5**(9), 1–8 (2010)
7. Dubé, C., Ribble, C., Kelton, D., McNab, B.: Comparing network analysis measures to determine potential epidemic size of highly contagious exotic diseases in fragmented monthly networks of dairy cattle movements in Ontario, Canada. *Transbound. Emerg. Dis.* **55**(9–10), 382–392 (2008)
8. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2003, pp. 137–146. ACM, New York (2003)
9. Krause, A., Golovin, D.: Submodular function maximization. In: *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press (2014)
10. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 420–429. ACM, New York (2007)
11. Nemhauser, G.L., Wolsey, L.A.: Best algorithms for approximating the maximum of a submodular set function. *Math. Oper. Res.* **3**(3), 177–188 (1978)
12. Panagopoulos, G., Malliaros, F.D., Vazirgiannis, M.: DiffuGreedy: an influence maximization algorithm based on diffusion cascades. In: Aiello, L.M., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., Rocha, L.M. (eds.) *Complex Networks and Their Applications VII*, pp. 392–404. Springer, Cham (2019)
13. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* **86**, 3200–3203 (2001)
14. Rocha, L.E.C., Liljeros, F., Holme, P.: Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput. Biol.* **7**(3), 1–9 (2011)
15. Schilling, R.L.: *Measures, Integrals and Martingales*. Cambridge University Press, Cambridge (2005)
16. Sterchi, M., Faverjon, C., Sarasua, C., Vargas, M.E., Berezowski, J., Bernstein, A., Grüttner, R., Nathues, H.: The pig transport network in Switzerland: structure, patterns, and implications for the transmission of infectious diseases between animal holdings. *PLoS ONE* **14**(5), 1–20 (2019)



Efficient Computation of Optimal Temporal Walks Under Waiting-Time Constraints

Anne-Sophie Himmel^(✉), Matthias Bentert, André Nichterlein,
and Rolf Niedermeier

Faculty IV, Algorithmics and Computational Complexity, TU Berlin,
Berlin, Germany

{anne-sophie.himmel,matthias.bentert,andre.nichterlein,
rolf.niedermeier}@tu-berlin.de

Abstract. Node connectivity plays a central role in temporal network analysis. We provide a comprehensive study of various concepts of walks in temporal graphs, that is, graphs with fixed vertex sets but edge sets changing over time. Importantly, the temporal aspect results in a rich set of optimization criteria for “shortest” walks. Extending and significantly broadening state-of-the-art work of Wu et al. [IEEE TKDE 2016], we provide an algorithm for computing shortest walks that is capable to deal with various optimization criteria and any linear combination of these. It runs in $O(|V| + |E| \log |E|)$ time where $|V|$ is the number of vertices and $|E|$ is the number of time edges. A central distinguishing factor to Wu et al.’s work is that our model allows to, motivated by real-world applications, respect waiting-time constraints for vertices, that is, the minimum and maximum waiting time allowed in intermediate vertices of a walk. Moreover, other than Wu et al. our algorithm also allows to search for walks that pass multiple edges in one time step, and it can optimize a richer set of optimization criteria. Our experimental studies indicate that our richer modeling can be achieved without significantly worsening the running time when compared to Wu et al.’s algorithms.

Keywords: Temporal networks · Temporal paths · Shortest path computation · Waiting policies · Infectious disease spreading

1 Introduction

Computing shortest paths in networks is arguably among the most important graph algorithms, relevant in numerous application contexts and being used as a subroutine in a highly diverse set of applications. While the case has been studied in static graphs for decades, over the last years there has been an intensified interest in studying shortest path computations in *temporal graphs*—graphs

Full version available on arXiv (<https://arxiv.org/abs/1909.01152>).

A.-S. Himmel—Supported by the DFG, project FPTinP (NI 369/16).

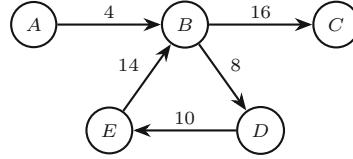


Fig. 1. A temporal graph (with time-labeled edges) with maximum waiting time four for each vertex in which the only temporal walk from A to C visits B twice.

where the vertex set remains static, but the edge set may change over (discrete) time.

Two natural motivating examples for the relevance of path (walk) computations in temporal graphs are as follows. First, Wu et al. [15] discuss applications in flight networks where every node represents an airport and each edge is labeled with a flight’s departure time. Clearly, a “shortest” path may then relate to a most convenient flight connection between two cities. Second, understanding the spread of infectious diseases is a major challenge to global health. Herein, nodes represent persons and time-labeled edges represent contacts between persons where say a virus can be transmitted. “Shortest” path (walk) analysis here may help us (among other concepts of connectivity) to find measures against disease spreading [12, Chapter 17]. Notably, in both examples one might need to also take into account issues such as different concepts of “shortest”—also called optimal—paths (walks) or waiting times in nodes; this will be an important aspect of our modeling.

Our main reference point is the work of Wu et al. [15] on efficient algorithms for temporal path computation. These are also implemented in the temporal graph library of Apache Flink [9]. We extend their model with respect to two aspects. First, we additionally consider waiting-time constraints¹ for the network nodes; importantly, this implies that we need to take into account cycles in any walk from one node to another (in Wu et al.’s model without waiting times there is always an (optimal) walk that is a path because no cycles are necessary); refer to Fig. 1 for a simple example. Actually, if one insists on paths (without repeated nodes) instead of walks, then the optimization becomes NP-hard [3]. The second extension to Wu et al.’s work lies in an increased number of optimality criteria (different notions of optimal walks) and the fact that we do not only deal with optimizing one criterion but a linear combination of any of these, thus addressing richer modeling needs in real-world applications. Interestingly, while we still provide efficient worst-case algorithms, trying to optimize under multiple constraints resp. optimization criteria (and not just a linear combination as we do) leads to NP-hard computational problems [17].

¹ Waiting-time constraints are particularly important in the context of studying social networks and the spread of infectious diseases [12, Chapter 17].

Related Work. One of the first algorithms for computing optimal temporal walks is due to Xuan et al. [16]. They computed temporal walks under different optimization criteria, namely *foremost*, *fastest*, and *minimum hop-count*² for a restricted variant of our temporal graph model. Wu et al. [15] followed up by introducing algorithms for computing optimal walks for the optimization criteria *foremost*, *reverse-foremost*, *fastest*, and *shortest* on temporal graphs with no waiting-time constraints. Their algorithms run in linear and quasi-linear time with respect to the number of time-arcs, provided that transmission times are greater than zero on every time-arc. Research on multi-objective optimal path computation can be found in the related field of route planning [2].

The study of minimum- and maximum-waiting-time constraints in vertices has not received much attention in the context of temporal walks even though they are considered as important extensions to the temporal walk model [6, 13]. Dean [4] investigated waiting-time policies for finding optimal walks on a restricted temporal graph model. Modiri et al. [11] and Kivelä et al. [7] studied reachability in temporal graphs under maximum-waiting-time constraints using event graphs.

Our Contributions. We analyze the running time complexity of computing optimal temporal walks under waiting-time constraints. We develop and (theoretically and empirically) analyze an algorithm for finding an optimal walk from a source vertex to each vertex in directed temporal graphs. Our algorithm (provided in Sect. 3) runs in quasi-linear time in the number of vertices plus the number of time-arcs. This implies that the additional modeling of waiting-time constraints does not increase the asymptotic computational complexity of finding optimal temporal walks. Moreover, our algorithm can compute optimal walks not only for single optimality criteria but also for any linear combination of these. In experiments (see Sect. 4) on real-world data sets, we demonstrate that in terms of efficiency our algorithm can compete with state-of-the-art algorithms by Wu et al. [15]. Due to the lack of space, many details (including the discussion of the optimality criteria *shortest*, *minimum waiting time*, and *most-likely*) are deferred to the full arXiv version.

2 Modeling of Optimal Temporal Walks

Before we introduce our basic concepts relating to temporal graphs and walks, we start with a more extensive discussion of a motivating example from the disease spreading context.

Pandemic spread of an infectious disease is a great threat to global health, potentially associated with high mortality rates as well as economic fallout [14]. A large part of the legwork required to understand the dynamics of infectious diseases is the analysis of transmission routes through proximity networks [14]. Classic graph theory can be used to model the main structure of a network: Each person in the network is represented by a node and an edge between two

² Refer to the next section for definitions of these and further optimality criteria.

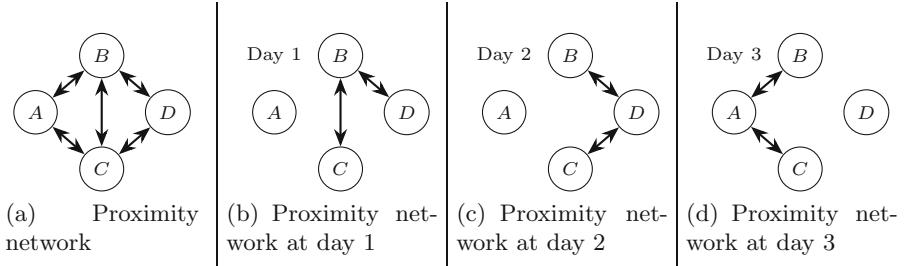


Fig. 2. A proximity network modeled as a static graph (Fig. 2(a)) and a closer look at the days in which the proximity contacts appear (Fig. 2(b) to (d)).

nodes indicates at least one proximity contact between these persons. However, the time component plays a crucial role in the analysis of transmission routes of a potential disease, as shown in the following example:

Example 1. Studying a proximity network as shown in Fig. 2(a), there are several transmission routes from A to D , e.g., $A \rightarrow B \rightarrow D$ and $A \rightarrow C \rightarrow D$, by which a disease could have spread. If we extend our model by the points in time of proximity contacts in Fig. 2(b) to (d), then we reach the conclusion that a disease could not have spread from A to D . The proximity contacts $A \xrightarrow{3} B$ and $A \xrightarrow{3} C$ occurred on day three whereas the contacts $B \xrightarrow{1} D$ and $C \xrightarrow{2} D$ occurred on days one and two, respectively. Thus, A could only have infected B and C after proximity contact with D . \square

In addition to what has been said so far, the infectious period of a disease also has to be taken into account when computing potential transmission routes through the network, implying the minimum time a person has to be infected before she becomes contagious herself and the maximum time a person can be infected before she is no longer contagious:

Example 2. If person B was infected by person A on day four ($A \xrightarrow{4} B$) and the infectious period of the disease starts after one day and ends after the fourth day, then person B could not have infected person C she met on day ten ($B \xrightarrow{10} C$). \square

Temporal Graphs. These are capable of representing both properties elaborated in the two examples above. Temporal graphs are a frequently used model in the prediction and control of infectious diseases [5, 10].

In this paper, we will consider the following model: A *temporal graph* $\mathcal{G} = (V, E, T, \alpha, \beta)$ is a five-tuple consisting of a lifetime $T \in \mathbb{N}$, a vertex set V , a time-arc set $E \subseteq V \times V \times \{1, \dots, T\} \times \{0, \dots, T\}$, a minimum waiting time $\alpha: V \rightarrow \{0, \dots, T\}$, and a maximum waiting time $\beta: V \rightarrow \{0, \dots, T\}$. A time-arc $(v, w, t, \lambda) \in E$ is a directed connection from v to w with *time stamp* t and *transmission time* λ , that is, a transmission from v to w starting at time step t

and taking λ time steps to cross the arc. The *arrival time* in vertex w is then $t + \lambda$. The two waiting-time functions $\alpha: V \rightarrow \mathbb{N}$ and $\beta: V \rightarrow \mathbb{N}$ assign each vertex a minimum and maximum waiting time, respectively; these functions can reflect the infectious period in our previous example. We set V_t to be the vertex subset $V_t \subseteq V$ at time t , that is, $V_t := \{v \mid (v, w, t, \lambda) \in E \vee (w, v, t, \lambda) \in E\}$; E_t to be the time-arc subset at time t , that is, $E_t := \{(v, w) \mid (v, w, t, \lambda) \in E\}$; and G_t to be the directed static graph $G_t := (V_t, E_t)$.

In our running disease spreading example, we are interested in transmission routes of an infectious disease. These transmission routes can revisit a person in the proximity network due to possible reinfection [1].

Temporal Walks & Optimal Temporal Walks. In temporal graphs, temporal walks are the fundamental concept that implements transmission routes.

A temporal walk is a sequence of time-arcs which connects a sequence of vertices and which are non-decreasing in time. In our model, a temporal walk additionally ensures that it remains the minimum waiting time in each intermediate vertex and does not exceed the maximum waiting time in any intermediate vertex of the walk.

Definition 1 (Temporal Walk). *Given a temporal graph $\mathcal{G} = (V, E, T, \alpha, \beta)$ and two vertices $s, z \in V$, a temporal walk from s to z is a sequence of time-arcs (e_1, e_2, \dots, e_k) with $e_i = (v_i, w_i, t_i, \lambda_i) \in E$ such that $s = v_1$, $z = w_k$, $w_j = v_{j+1}$, and $t_j + \lambda_j + \alpha(w_j) \leq t_{j+1} \leq t_j + \lambda_j + \beta(w_j)$ for all $j \in \{2, \dots, k - 1\}$.*

Example 3. Continuing Example 2, a valid temporal walk (transmission route) from A to D is $A \xrightarrow{4} B \xrightarrow{8} D$. Person A could have infected person B on day four. Due to the infectious period of four days, B was still contagious on day eight when she had contact with person C . \square

A *temporal path* is a temporal walk where all vertices are pairwise distinct. Maximum-waiting-time constraints have significant impact on temporal walks. As a consequence, there can be two vertices A and C such that any temporal walk from A to C is not a path, as shown in Fig. 1.

We are interested in temporal walks within our proximity network in general, but wish to place emphasis on temporal walks that optimize certain properties. A plethora of criteria can be optimized as a consequence of the time aspect (see full version on arXiv for an extensive discussion).

Definition 2 (Optimal Temporal Walk). *Let $\mathcal{G} = (V, E, T, \alpha, \beta)$ be a temporal graph, $c: E \rightarrow \mathbb{N}$ be a cost function, and $s, z \in V$ be two vertices. A temporal walk $P = (e_1, e_2, \dots, e_k)$ from s to z with $e_i = (v_i, w_i, t_i, \lambda_i)$ for all $i \in \{1, \dots, k\}$ is called optimal if it minimizes or maximizes a certain value among all temporal walks from s to z . We consider the following variants:*

criterion	min/max	optimization value
<i>foremost</i>	<i>min</i>	$t_k + \lambda_k$
<i>reverse-foremost</i>	<i>max</i>	t_1
<i>fastest</i>	<i>min</i>	$(t_k + \lambda_k) - t_1$
<i>min hop-count</i>	<i>min</i>	k
<i>cheapest</i>	<i>min</i>	$\sum_{i=1}^k c(e_i)$

To highlight the relevance of the different optimality criteria as well as their linear combinations, we briefly discuss the criteria *foremost* and *min hop-count* in our running disease spreading example.

A *foremost* walk is a temporal walk that has the earliest arrival time possible. Computing a *foremost* walk from a source vertex to all vertices in the proximity network signifies the speed with which an infectious disease could spread.

Example 4. Continuing Examples 1 to 3, person *A* can infect persons *B* and *C* on day 3; however, person *D* can only be infected on day 8. Consequently, the infectious disease could have permeated the entire system by day 8. \square

A *min-hop-count* walk is a temporal walk that minimizes the number of time-arcs. This can be seen as the most-likely transmission route of an infectious disease because with each intermediate contact the probability of contagion declines. Using a *linear combination* of these two criteria, we can compute a more realistic speed with which an infectious disease could spread within the network because the transmission routes determining the speed are more likely.

Transformations. To simplify the presentation of the forthcoming algorithm in Sect. 3, we designed it to run only on *instantaneous* temporal graphs, that is, temporal graphs with no transmission times ($\lambda = 0$ for all $(v, w, t, \lambda) \in E$) and no minimum-waiting-time constraints ($\alpha(v) = 0$ for all $v \in V$). This is no restriction since we can eliminate these with the following construction.

Transformation 1 (Remove α and λ). Let $\mathcal{G} = (V, E, T, \alpha, \beta)$ be a temporal graph. Transform \mathcal{G} into an instantaneous temporal graph $\mathcal{G}' = (V', E', T, \beta')$

- $V' = V \cup V_E$ with $V_E = \{v_e \mid e \in E\}$,
- $E' = \{(v, v_e, t, 0), (v_e, w, t + \lambda + \alpha(w), 0) \mid (v, w, t, \lambda) \in E\}$, and
- $\beta': V' \rightarrow \mathbb{N}$ with

$$\beta'(v) = \begin{cases} \beta(v) & \text{for } v \in V \\ T & \text{for } v \in V_E. \end{cases}$$

Proposition 1. Any temporal graph can be transformed by Transformation 1 into an equivalent instantaneous temporal graph in linear time.

Proof. It is easy to verify that Transformation 1 runs in $O(|V| + |E|)$ time. It remains to show that any time-arc sequence $P = (e_1, e_2, \dots, e_k)$ with

$e_i = (v_i, w_i, t_i, \lambda_i)$ is a temporal walk in \mathcal{G} if and only if $P' = (e_1^s, e_1^a, \dots, e_k^s, e_k^a)$ with $e_i^s = (v_i, v_{e_i}, t_i, 0) \in E'$, $e_i^a = (v_{e_i}, w_i, t_i + \lambda_i + \alpha(w_i), 0) \in E'$ is a temporal walk in \mathcal{G}' .

To this end, let $P = (e_1, \dots, e_k)$ be a temporal walk in \mathcal{G} . By Transformation 1, it holds that $e_i^s, e_i^a \in E'$. It is easy to verify that the time-arc sequence (e_i^s, e_i^a) is a valid temporal walk in \mathcal{G}' due to $\beta'(v_{e_i}) = T$. Furthermore, for every $e_i, e_{i+1} \in P$ and, consequently, for $e_i^a, e_{i+1}^s \in P'$ it holds that $t_i + \lambda_i + \alpha(w_i) \leq t_{i+1} \leq t_i + \lambda_i + \beta(w_i)$. Thus, the sequence (e_i^a, e_{i+1}^s) is a temporal walk in \mathcal{G}' if and only if the sequence (e_i, e_{i+1}) is a temporal walk in \mathcal{G} .

Altogether, we can conclude that every P in \mathcal{G} is a temporal walk if and only if P' is a temporal walk in \mathcal{G}' . \square

We have shown that Transformation 1 does not influence the existence of any temporal walk in the original graph. It neither changes the departure time and exactly doubles the number of time-arcs of a temporal walk. However, Transformation 1 influences the arrival time in a vertex. All walks arriving at a vertex v are delayed by the same value $\alpha(v)$. Thus, it does not change foremost or fastest walks but only their objective values.

3 Computing Optimal Temporal Walks

In this section, we explain the main idea of our algorithm (Algorithm 1) that computes an optimal walk from a given source to each vertex in the instantaneous temporal graph. Algorithm 1 optimizes for any linear combination of our optimality criteria and, hence, also for any single criterion.

Algorithm Structure. Given an instantaneous temporal graph $\mathcal{G} = (V, E, T, \beta)$ and a source vertex $s \in V$, for each $t \in \{1, \dots, T\}$, Algorithm 1 performs three main steps that we will discuss in detail: *GraphGeneration*, *ModDijkstra*, and *Update*.

Efficiently storing and accessing the value of an optimal walk from s to v that arrives at a certain time step t is the heart of the algorithm. We can maintain this information in $O(|E|)$ time during a run of Algorithm 1 such that this information can be accessed in constant time. This leads to the following main theorem.

Theorem 1 (³). *With respect to any linear combination of the optimality criteria, an optimal temporal walk from a source vertex s to each vertex in a temporal graph can be computed in $O(|V| + |E| \log |E|)$ time.*

If we restrict ourselves to the single criterion *foremost*, then we can improve upon the running time to $O(|V| + |E|)$ assuming a sorted time-arc list as input.

³ For a full proof of the theorem, see full arXiv version.

Algorithm 1. Computes optimal walks.

Input: An instantaneous temporal graph \mathcal{G} and a source vertex $s \in V$.

Output: For each $v \in V$ the specific length of an optimal $s-v$ walk.

Variables:

$\text{opt}(v)$ stores the value of an optimal walk from s to v within $[0, t]$;

$L(v)$ is a sorted list $[(\text{opt}_{a_1}, a_1), \dots, (\text{opt}_{a_k}, a_k)]$ where opt_{a_i} is an optimal value of a walk from s to v that arrives at time a_i with $t + \beta(v) \leq a_i \leq t$. We will sort the list such that: $\text{opt}_{a_1} < \dots < \text{opt}_{a_k}$ and $a_1 < \dots < a_k$;

\mathcal{G} is an instantaneous temporal graph with a sorted time-arc list;

$\delta_1, \dots, \delta_5$ are constants for the optimality criteria *foremost*, *reverse-foremost*, *fastest*, *cheapest*, and *min hop-count*, respectively.

```

1 Initialize  $\text{opt}(v) = \infty$  and  $L(v)$  as empty list for all  $v \in V \setminus \{s\}$ 
2 for  $t = 1, \dots, T$  with  $E_t \neq \emptyset$  do
3    $G, d_t, d_r \leftarrow \text{generateGraph}(G_t, L, s)$ 
4    $V', \text{opt}_t \leftarrow \text{modDijkstra}(G, d_t, d_r, s)$ 
5   for  $v \in V'$  do
6      $\text{opt}(v) = \min\{\text{opt}(v), \delta_1 \cdot t + \delta_3 \cdot (t - T) + \text{opt}_t(v)\}$ 
7      $L(v) \leftarrow \text{append } (\text{opt}_t(v), t) \text{ and delete redundant tuples}$ 
8 return  $\text{opt}$ 
9 function  $\text{generateGraph}(G_t, L, s)$ :
10   Initialize  $E_r = \emptyset$ 
11   for  $v \in V_t \setminus \{s\}$  do
12     delete tuples  $(\text{opt}_a, a)$  in  $L(v)$  with  $a + \beta(v) < t$ 
13     if  $L(v)$  not empty then
14        $E_r \leftarrow E_r \cup \{(s, v)\}$ 
15        $d_r(s, v) = \text{opt}_a$  with  $\text{opt}_a = \min\{\text{opt}_a \mid (\text{opt}_a, a) \in L(v)\}$ 
16     for  $(v, w) \in E_t$  do
17        $d_t(v, w) = \begin{cases} (\delta_2 + \delta_3) \cdot (T - t) + \delta_4 \cdot c(v, w) + \delta_5 & \text{if } v = s \\ \delta_4 \cdot c(v, w) + \delta_5 & \text{else} \end{cases}$ 
18   return  $((V_t \cup \{s\}, E_t \cup E_r), d_t, d_r)$ 
19 function  $\text{modDijkstra}((V, E_t \cup E_r), d_t, d_r, s)$ :
20   initialize  $\text{opt}_t(v) = \infty$ ,  $r(v) = \infty$  for all  $v \in V_t$ , and  $r(s) = 0$ 
21   initialize  $Q = V$  and  $V' \neq \emptyset$ 
22   while  $Q \neq \emptyset$  do
23      $v \leftarrow \text{vertex in } Q \text{ with min } r(v)$ 
24     remove  $v$  from  $Q$ 
25     for  $(v, w) \in E_t \cup E_r$  do
26        $r(w) = \min\{r(w), r(v) + \min\{d_t(v, w), d_r(v, w)\}\}$ 
27       if  $(v, w) \in E_t$  then
28          $\text{opt}_t(w) = \min\{\text{opt}_t(w), r(v) + d_t(v, w)\}$ 
29          $V' \leftarrow V' \cup \{w\}$ 
30   return  $V', \text{opt}_t$ 

```

Algorithm Details. Let $\mathcal{G} = (V, E, T, \beta)$ be an instantaneous temporal graph and let $s \in V$ be the source. In the beginning, $\text{opt}(v) = \infty$ and $L(v)$ is initialized with an empty list (Line 1 in Algorithm 1). Then, for each time step t , Algorithm 1 computes the optimal value for a walk from the source s to a vertex that arrives in time step t (if it exists). For each $t \in \{1, \dots, T\}$, it performs the following steps:

GraphGeneration. Generate a static graph G (Line 3 and Lines 9 to 18). This graph consists of the static graph $G_t = (V_t, E_t)$, that is, the static graph induced by all time-arcs with time stamp t , and the source vertex s . The weight of each arc in E_t is set depending on the optimality criterion. Additionally, non-existing arcs from s to each vertex $v \in V_t$ are added if there exists a temporal walk from s to v that arrived in the last $\beta(v)$ time steps. The weight of such an arc is then set to the optimal value among all walks from s to v that arrived in the last $\beta(v)$ time steps (see Line 15).

ModDijkstra. Run a modified Dijkstra Algorithm on the static graph G (Line 4 and Lines 19 to 30) that computes a shortest walk among all walks that end in an arc of E_t . This represents a temporal walk that arrives in time step t . It returns the set V' of vertices that can be reached within G via an arc in E_t and the function $\text{opt}_t: V' \rightarrow \mathbb{N}$ that maps each vertex $v \in V'$ to its optimality value an walk from s to v that arrives exactly at time t .

Update. For each $v \in V'$, set the optimum $\text{opt}(v)$ to the minimum of its current value and the value of a newly computed walk (Line 6). Add the tuple $(\text{opt}_t(v), t)$ to list $L(v)$ (Line 7).

After the *Update* step for time step t , the list $L(v)$ contains all tuples (opt_a, a) such that there exists a walk from s to v that arrives in $a \in [t - \beta(v), t]$ with its optimality value. We want to have constant-time access to the optimal value of a walk from s to v within time interval $[t - \beta(v), t]$ (see Line 15). This can be achieved by safely removing tuples from list $L(v)$ that are redundant for the optimal walk computation, that is, these tuples are nonmeaningful for the correct computation of optimal walks. Let

$$L(v) = [(\text{opt}_{a_1}, a_1), \dots, (\text{opt}_{a_k}, a_k)]$$

be such a list with $t - \beta(v) \leq a_1 < \dots < a_k \leq t$. A tuple (opt_a, a) is *redundant* if there exists a tuple with an arrival time greater than a and an optimality value smaller than opt_a . This is shown with the following lemma:

Lemma 1. *For a time step $t \in \{1, \dots, T\}$ and a vertex $v \in V$, if there are two tuples $(\text{opt}_{a_i}, a_i), (\text{opt}_{a_j}, a_j) \in L(v)$ with $a_i < a_j$ and $\text{opt}_{a_i} \geq \text{opt}_{a_j}$, then (opt_{a_i}, a_i) can be removed from $L(v)$.*

Proof. After time step t , Algorithm 1 only considers time-arcs with a time stamp $t' > t$. In the generated graph G (Line 3), the algorithm adds an arc from s to $v \in V_{t'}$ if a walk from s arrives in v within time interval $[t' - \beta(v), t']$.

If $a_i \in [t' - \beta(v), t']$, then $a_j \in [t' - \beta(v), t']$ because $a_i < a_j < t'$. Furthermore, let opt^* be the optimal value of a walk to v that arrives within time interval $[t' - \beta(v), t']$. The weight of the arc (s, v) is set to d . Due to $a_i, a_j \in [t' - \beta(v), t']$ and $\text{opt}^* \leq \text{opt}_{a_j} \leq \text{opt}_{a_i}$, the tuple (opt_{a_i}, a_i) is not needed in the list L_v at time step t anymore and can be safely removed. \square

If $L(v)$ does not contain any redundant tuples, then it also holds that $\text{opt}_{a_1} < \dots < \text{opt}_{a_k}$. Hence, (a_1, opt_{a_1}) has the optimal value of a walk that arrives within time interval $[t - \beta(v), t]$. It follows that finding $\text{opt}_a = \min\{\text{opt}_a \mid (\text{opt}_a, a) \in L(v)\}$ in Line 15 takes constant time. The deletion of redundant tuples takes $O(|E|)$ time during the whole run of Algorithm 1. Apart from that, for a $t \in \{1, \dots, T\}$ the functions generateGraph (Line 3) and modDijkstra (Line 4) run in $O(|E_t|)$ and $O(|E_t| \log |E_t|)$ time, respectively. These are the important pieces to derive a running time of $O(|V| + |E| \log |E|)$ for Algorithm 1.

4 Experimental Results

We implemented Algorithm 1 and performed some experimental studies including comparisons to existing state-of-the-art algorithms by Wu et al. [15] for single criteria. We show that our algorithm can compete with these algorithms on real-world instances when computing temporal walks with no maximum-waiting-time constraints. We further examine the influence of different maximum-waiting-time values on the existence and structure (e.g., number of cycles) of optimal temporal walks and on the running time of Algorithm 1.

Setup and Statistics. We implemented Algorithm 1 in C++ (v11) and performed our experiments on an Intel Xeon E5-1620 computer with 64 GB of RAM and four cores clocked at 3.6 GHz each. The operating system was Debian GNU/Linux 7.0 where we compiled the program with GCC v7.3.0 on optimization level -O3. We compare Algorithm 1 to the algorithms of Wu et al. [15] using their C++ code and testing it on the same hardware and with the same compiler. We tested our algorithm on the same freely available data sets as Wu et al. [15] from the well-established SNAP library [8]. The graphs have between 7,000 to $2 \cdot 10^6$ vertices and between 10,000 to $10 \cdot 10^6$ edges. For each optimization criterion, each $\beta \equiv c, c \in \{1, 2, 4, 8, \dots, 2^{\lceil \log T \rceil}\}$, and each data set, Algorithm 1 ran for 100 fixed source vertices of the data set chosen independently and uniformly at random to ensure comparability.⁴

⁴ Open source code is freely available at <https://fpt.akt.tu-berlin.de/temporalwalks>.

Comparison with Wu et al. [15]. When comparing with the algorithms by Wu et al. [15], we only use the runs with no maximum-waiting-time constraints ($\beta \equiv T$) and we tested all algorithms on the same set of randomly chosen starting vertices.

Algorithm 1 has a larger variance and is therefore more dependent on the choice of starting vertices in comparison to Wu et al. [15]. This is due to the fact that Algorithm 1 only considers arcs that start in vertices that were already visited while the algorithm by Wu et al. [15] always considers the whole sorted time-arc list and therefore has almost no variance in the running time. We mention in passing that we observed that even for $\beta \equiv T$, not all vertices can reach all other vertices by temporal walks in the considered graphs. If one takes the running time of an average run of each algorithm, that is, the median value of running times, then both algorithms have comparable running times. If one takes the average running time of each algorithm, then the running time of Algorithm 1 is higher than the running time of the algorithm by Wu et al. [15] by a factor of roughly ten (averaged over all optimization criteria). Despite the fact that this is a weakness of our algorithm, we believe it to be a valuable contribution as it solves more general problems: it can easily combine multiple optimization criteria and it can cope with maximum waiting times and instantaneous arcs, that is, arcs with $\lambda = 0$.

Effect of different β -values. We next analyze the impact that the maximum-waiting-time constraint β has on Algorithm 1. Decreasing β can have two different effects: First, it can make temporal walks invalid as the maximum allowed waiting time in a vertex is exceeded. Thus, with small β -values certain vertices can only reach few vertices by temporal walks. The second effect is that a temporal walk is invalidated but can be fixed by a detour that starts and ends in the vertex in which the maximal waiting time was exceeded.

We first investigate the second effect. To this end, we partition the optimization criteria into two categories: The first category contains all optimization criteria for which a detour has no negative effect on the solution. These are *foremost*, *reverse-foremost*, *fastest*, and *minimum waiting time*. Since the solution for, e. g., *fastest* is only depending on the first and last edge of the temporal walk, adding a cycle somewhere in between does not change the solution. The second category contains all other optimization criteria, that is, those for which a detour has a negative effect on the solution. These are *minimum hop count*, *cheapest*, and *shortest*. Due to the lack of space, we only display the effect for *cheapest* in Fig. 3 (left side). Note that we could not observe significant differences for the different optimization criteria within a category.

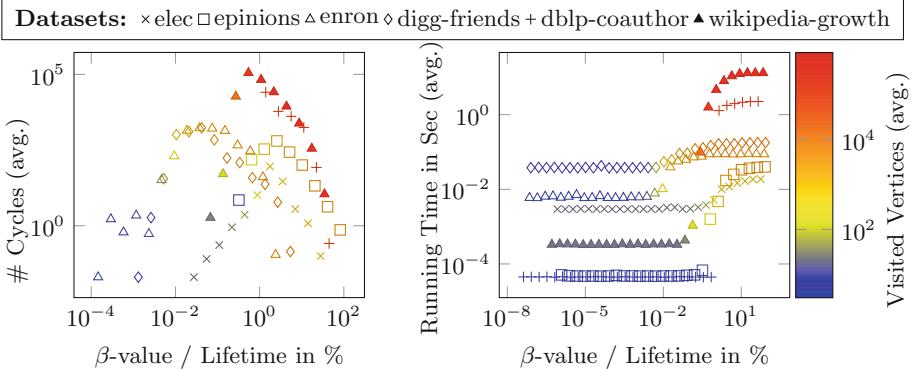


Fig. 3. Impact of different β -values on the number of cycles in a *cheapest* walk and on the number of vertices that can be reached by temporal walks from the chosen starting vertices (left diagram) and on the running time (right diagram).

Figure 3 (right side) shows the running-time dependence on the value of β and is representative for all criteria. It seems to be more likely that the first effect we described in the beginning (that decreasing β -values can make temporal walks invalid as the maximum allowed waiting time in a vertex is exceeded) is more important for explaining the running times. With increasing β -values, there seems to be a critical value (around 0.1%–10% of the lifetime of the temporal graph) where suddenly much more connections appear and hence the running time increases drastically. We observed that (almost) independently of the input graph, the running time linearly depends on the number of visited vertices.

5 Conclusion

Building on and widening previous work of Wu et al. [15], we provided a theoretical and experimental study of computing optimal temporal walks under waiting-time constraints. The performed experiments indicate the practical relevance of our approach. As to future challenges, recall that moving from walks to paths would yield NP-hard optimization problems. Hence, for the path scenario the study of approximation, fixed-parameter, or heuristic algorithms is a natural next step. For the scenario considered in this work, note that we did not study the natural extension to Pareto-optimal walks (under several optimization criteria). Moreover, for (temporal) network centrality measures based on shortest paths and walks, counting or even listing *all* temporal walks or paths would be of interest.

References

1. Barabási, A.L.: Network Science. Cambridge University Press, Cambridge (2016)
2. Bast, H., Delling, D., Goldberg, A., Müller-Hannemann, M., Pajor, T., Sanders, P., Wagner, D., Werneck, R.F.: Route planning in transportation networks. In: Algorithm Engineering, pp. 19–80. Springer (2016)
3. Casteigts, A., Himmel, A.S., Molter, H., Zschoche, P.: The computational complexity of finding temporal paths under waiting time constraints. arXiv preprint [arXiv:1909.06437](https://arxiv.org/abs/1909.06437) (2019)
4. Dean, B.C.: Algorithms for minimum-cost paths in time-dependent networks with waiting policies. Networks **44**, 41–46 (2004)
5. Holme, P.: Temporal network structures controlling disease spreading. Phys. Rev. E **94**(2), 022305 (2016)
6. Holme, P., Saramäki, J.: Temporal networks. Phys. Rep. **519**(3), 97–125 (2012)
7. Kivelä, M., Cambe, J., Saramäki, J., Karsai, M.: Mapping temporal-network percolation to weighted, static event graphs. Sci. Rep. **8**(1), 12357 (2018)
8. Leskovec, J., Krevl, A.: SNAP Datasets: stanford large network dataset collection. <http://snap.stanford.edu/data> (2014)
9. Lightenberg, W., Pei, Y., Fletcher, G., Pechenizkiy, M.: Tink: a temporal graph analytics library for Apache Flink. In: Proceedings of WWW 2018, pp. 71–72. International World Wide Web Conferences Steering Committee (2018)
10. Masuda, N., Holme, P.: Predicting and controlling infectious disease epidemics using temporal networks. F1000prime Rep. **5**, 6 (2013)
11. Modiri, A.B., Karsai, M., Kivelä, M.: Efficient limited time reachability estimation in temporal networks. arXiv preprint [arXiv:1908.11831](https://arxiv.org/abs/1908.11831) (2019)
12. Newman, M.E.J.: Networks. Oxford University Press, Oxford (2018)
13. Pan, R.K., Saramäki, J.: Path lengths, correlations, and centrality in temporal networks. Phys. Rev. E **84**(1), 016105 (2011)
14. Salathé, M., Kazandjieva, M., Lee, J.W., Levis, P., Feldman, M.W., Jones, J.H.: A high-resolution human contact network for infectious disease transmission. Proc. Nat. Acad. Sci. **107**(51), 22020–22025 (2010)
15. Wu, H., Cheng, J., Ke, Y., Huang, S., Huang, Y., Wu, H.: Efficient algorithms for temporal path computation. IEEE Trans. Knowl. Data Eng. **28**(11), 2927–2942 (2016)
16. Xuan, B.B., Ferreira, A., Jarry, A.: Computing shortest, fastest, and foremost journeys in dynamic networks. Int. J. Found. Comput. Sci. **14**(02), 267–285 (2003)
17. Zhao, A., Liu, G., Zheng, B., Zhao, Y., Zheng, K.: Temporal paths discovery with multiple constraints in attributed dynamic graphs. World Wide Web, 1–24 (2019)



Roles in Social Interactions: Graphlets in Temporal Networks Applied to Learning Analytics

Raphaël Charbey¹(✉), Laurent Brisson¹, Cécile Bothorel¹, Philippe Ruffieux²,
Serge Garlatti¹, Jean-Marie Gilliot¹, and Antoine Mallégol¹

¹ IMT Atlantique, Lab-STICC UMR CNRS 6285, 29238 Brest, France
`{raphael.charbey,laurent.brisson,cecile.bothorel,serge.garlatti,
jean-marie.gilliot,antoine.malleol}@imt-atlantique.fr`

² Usages du numérique et Didactique de l'informatique (MUNDI), Av. des Bains 21,
1014 Lausanne, VD, Switzerland
<http://www.sqily.com>

Abstract. There is a growing interest in how data generated in learning platforms, especially the interaction data, can be used to improve teaching and learning. Social network analysis and machine learning methods take advantage of network topology to detect relational patterns and model interaction behaviors. Specifically, small induced subgraphs called graphlets, provide an efficient topological description of the way each node is embedded in the meso-scale structure of a network. Here we propose to detect the roles occupied by the different participants, students and teachers, in the successive phases of courses modeled by a sequence of static snapshots. The detected positions, obtained thanks to graphlet enumeration combined with a clustering method, reveal the different roles observed in each snapshot. We also track the role changes through the overall sequence of snapshots. We apply our method to the Sqily platform and describe the mutual skill validation process. The detected roles, the transitions between roles and a overall visualization through Sankey diagrams help interpreting the course dynamics. We found that some roles act like necessary steps to engage students within an active exchange process with their classmates.

Keywords: Temporal networks · Social interactions · Motifs · Graphlets · Learning analytics · Role detection

1 Introduction

Networks provide a framework for studying complex systems in a broad set of fields, ranging from biology, telecommunication to social networks and e-learning. Actors and their relationships are modeled as graphs, where edges represent friendship or any kind of interaction. For example in the field of e-learning, as it is now well accepted that a student's collaboration [7] is central for facilitating the learning process, learning platforms propose discussion forums and

social networking facilities [6], and teachers are encouraged to design pedagogical activities involving interactions between learners. Social Learning Analytics [9] denotes a growing interest in managing interaction data, in order to improve teaching and learning, to assist educational institutions in increasing student retention and improving student success.

Network properties such as density, centrality, and degrees are very helpful for understanding the structure of a peer network, e.g. how students engage in learning and performance, both as individuals and as groups [22]. But detecting and interpreting patterns of exchanges that occur between the students is another step to understand the dynamics and identify or predict problematic situations. Understanding these patterns will help us understanding the course organisation and the role of each individual within the overall organisation.

The patterns of exchanges are likely to be detected through induced subgraph enumeration, that offers a competitive as well as an intuitive insight of the meso-scale structure of networks and allows for characterizing them relatively from one another. The induced subgraphs, also known as graphlets, also provide a native *role* detection support since nodes can be characterized by their positions in the graphlets. By studying the node roles in the successive phases of a temporal network, which can be modeled by a sequence of static graphs, one can understand their evolution and the way they transit from a role to another.

Contributions. In this article, we provide a methodology for describing interaction networks in a dynamic way at node level. We first show how to find graphlets in static networks. We then consider the problem of role detection. The position of a node in a graphlet describes the place it occupies in its neighborhood. We use a graphlet-based node embedding method which explicitly counts its different positions occupied in interaction schemes involving 3 nodes. Then we provide a data-driven way to build interaction profiles as a combination of positions representing the different role specific to the considered network.

In temporal networks, nodes are involved in different graphlets, and therefore roles, as time goes by. We discretize time by converting temporal information on edges, i.e. the timestamp when the interactions occurred, into a sequence of static networks called *snapshots*. By assigning roles to nodes during each snapshot, our analytical approach explores how they move from one role to another. Finally we propose a longitudinal visualization, with Sankey diagrams, of the transitions from one role to another throughout the life of a temporal network.

We apply this methodology to e-learning data from the Sqily platform dedicated to the mutual validation of skills. We thus illustrate how interaction patterns (based on the validation of skills) can make it possible to associate roles with learner behaviors. Static and dynamic analyses of these roles thus provide students and teachers with useful insights.

2 Related Work

Networks and Learning Analytics. Social network analysis (SNA) has proven extremely powerful at describing and analysing network behaviors in

e-learning. A systematic review of the literature on SNA in this field [4] covers more than 30 studies which analyze interaction patterns in forums and where centrality and density measures are mostly used. For example, in [26], low or high in- or out-degree centrality scores exhibit popular students who provide comments to others, who are reflectors and good communicators in the learning process or play knowledge broker's roles. Let us note that most of these studies deal with very small graphs (only a few dozen nodes).

The visualization of networks is an approach to explain the nature of community dynamics. KISSME is an example of an interactive visualization tool of content-aware interactions among learners [27]. Behavioral links (contributions such as “reply”, “reference” and “annotate”) can be overlaid on the learner-time display to show how patterns of interaction change over time.

Motifs in Temporal Networks. While dealing with large complex networks, one might need automatic ways to count and detect relevant, sometimes non obvious, behaviors. This is included in the scope of application of subgraph pattern detection in networks [12, 20]. In temporal networks, there is the constraint that edges have a timestamp, and the graphlets sought are induced subgraphs occurring in a wider network structure during specified time windows, with nodes interacting in an ordered sequence [13, 16]. When looking for motifs, we are interested in how many times each pattern occurs. An efficient counting of temporal graphlets is proposed in [21], along with a framework to compare the structure of several complex systems, e.g. they show that email exchanges and Facebook wall posts do not involve the same motifs and therefore reveal different communication behaviors.

In this work we are not interested in the sequences of interactions and temporal motifs, but more in the static motifs we observe at different phases of a teaching. Instantiated in the learning analytics field, our goal is to understand *with whom* learners interact, and the evolution of their position in the group at different times during a course.

Subgraph Enumeration and Nodes Embedding. If we discretize time and study interactions on successive snapshots, i.e. successive static graphs that are the result of an aggregation of interactions occurring during specified time windows [1, 8], the detection of motifs then consists in searching for static patterns within these snapshots. Hulovatyy et al. have introduced patterns that cover several consecutive snapshots and the persistence of edges through time [14]. Braha and Bar-Yam have enumerated the subgraphs of consecutive networks as if they were static ones [3] which is the strategy we adopt since the nodes we are dealing with are not always active in consecutive snapshots.

Graphlets are very useful to exhaustively list interaction patterns through, for example, a systematic enumeration of all possible subgraphs of a chosen size. A network may be characterized by a vector composed of the relative frequency of each graphlet, the ratio between its number of appearances and the total number of graphlets in the network [23]. Another possible vector is based on the graphlet representativity, that is the ratio between its relative frequency within a network and its relative frequency within a set of networks [5]. Moreover they allow for

easy interpretation of the results. They also can be used as an embedding method to model nodes' neighborhood [11, 19]. By enumerating the positions (or orbits) in which nodes appear, the graphlets offer a way to compare their topological similarity.

3 Method

We define a temporal oriented graph $G = (V, E)$ in which each edge (u, v, ts) means that the user $u \in V$ interacts with the user $v \in V$ at time ts . We discretize time by converting temporal information into a sequence of T *snapshots*. The interactions that occur during the period of each snapshot are aggregated leading to a set of T graphs $\{G_1, \dots, G_T\}$ where G_t is the directed graph representing the interactions between active nodes that occurred within the t -th time slice.

The first step of our proposal is to enumerate the different possible graphlets of size k and then search for them in the static graphs G_t . The second step is to compute position-based embeddings for the nodes in each G_t relatively to their positions within the different graphlets. Finally, a clustering step over the nodes vectors produce complex interaction profiles, mixing positions and exhibiting roles as distributions of frequencies of positions in graphlets.

Size k graphlet enumeration consists on visiting every possible combination of k nodes such that the subgraph induced by these nodes (the nodes themselves and all edges between them) is connected. Each of these subgraphs is then identified as isomorphic to one of the graphlets and the relative enumeration counter is incremented. The list of size-3 directed graphlets is presented in Fig. 1.

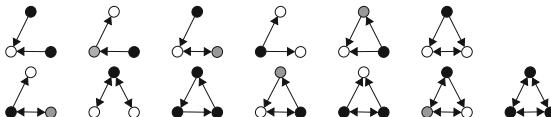


Fig. 1. The 13 directed graphlets of size three and their respective positions depicted by shades of grey. In each graphlet, the nodes with the same colour are in the same position.

Every position is interesting to observe. Intuitively, appearing in the central ↗ or peripheral ↗ orbit of the star does not correspond to the same role, even though the graphlet is the same.

As our goal is to describe the different positions that nodes occupy over time, and in order to build snapshots that will lead to the role dynamics, we focus on the temporal separation between two events as a referee for time discretization. We consider that a time window δ without any interaction is the marker of a break in the network dynamics which would result in an other organization. Each snapshot G_t therefore corresponds to a sequence of events so that two consecutive events are separated with less than δ .

On each snapshot, we enumerate the positions of each of their nodes using Fanmod process [28]. Then comes the step of nodes embedding. For each node in each snapshot, we compute the frequency of each of its positions as follows: the frequency of a vertex v in a position p is the ratio of the number of appearances of v in p with the total number of appearances of v in any position. This embedding is an adaptation of a popular graphlet-based network embedding [23] to the position level.

The set of position-based vectors are the exhaustive list of the different distributions of positions held by nodes, compiling the exhaustive set of behaviors detected at the different periods of the lifetime of the current temporal network. By applying the kMeans algorithm [17, 18] to these vectors, we obtain clusters of similar distributions of positions. The clustering produced by this algorithm is dependent on its initialization step and the number of clusters k that is given as a parameter. Therefore we run the algorithm a hundred times for each value of k between 1 and 20 and kept the best result in term of the silhouette score [24].

The resulting clusters define sets of nodes whose embeddings are similar. In our mutual validation course case-study, a user's embedding describes how s/he collaborate with others, by validating or being validated, and by whom. Through the experiment detailed in the next section, we will illustrate the power of expressiveness of positions in graphlets.

4 Case Study from the Sqily Platform

This case study focuses on interactions on the Sqily¹ learning platform, whose particularity is to promote the mutual validation of skills. We first show how we constructed a dataset from the available information on the mutual validation of skills, then we present the topological roles extracted by our method by linking them to the corresponding student behaviors. Finally, we conduct a longitudinal study that highlights the value of these roles in exploring the dynamics of interactions.

4.1 Mutual Validation of Skills

In this context we use the term skills in the broad sense, i.e. knowledge, capacities, know-how and professional skills. The principle of mutual validation of skills is to encourage the learner to adopt a reflective approach: s/he must mobilize her/his newly acquired skills in order to explain them and help other students acquiring them. The learner thus puts herself/himself in the role traditionally assigned to the teacher and deepens her/his skills.

The mutual validation of skills has two inspirations: mutual teaching [10] (verbalizing knowledge, exchanging with peers, improving understanding through repetition) and knowledge trees [2] (graphic representation of competencies,

¹ <http://www.sqily.com>.

knowledge management, valuing expertise). According to these principles, the Sqily platform:

- allows to define a tree of skills,
- allows students who master a skill to create assessments and validate skills of others students.

4.2 Dataset Description

In this case study we analyze the behavior of students and teachers in terms of skills validation. Each course is modeled as a graph G ; edges (u, v, ts) represent mutual skill validations where the user u validates one skill of the user v at time ts .

Our dataset is comprised of 11 courses that have an average of 58 students. There is a strong involvement in mutual validation: 30% of the skills were created by students and 70% of the evaluations were done through peer-reviewing. However, students validated only 40% of the skills for which they started the activity sequence.

We arbitrarily set the size of the separating window between two snapshots at 15 days (parameter δ). It generated a total of 64 graphs distributed differently over the courses: they are divided in at least 1 and up to 19 phases, with an average of 3.6 phases. The phases last at least 1 day and up to 103 days, with a mean of 32 days. The snapshots have an average of 21 vertices and 60 edges but they are also very heterogeneous with a minimum of 3 vertices and 2 edges and a maximum of 102 vertices and 370 edges.

4.3 Role Characterization

The topological framework based on position enumeration led us to obtain roles, which positions frequencies are summarized in Table 1. We describe these roles by characterizing them with variables that were not used to define them: the proportion of teachers/students (see Table 2), the student peer-reviewed rate (average rate of evaluations made by peer students) and the commitment level (rank of skill acquisition in the course) (see Table 2). We present the roles obtained in two sets: the first includes the roles most involved in validating the skills of other learners, while the second includes roles concerned about getting validated.

Assessor Oriented Roles

- **Expert** The Expert has mastered a skill and validates a large number of other learners. These learners generally are validated by the Expert alone. This role mainly brings together teachers and learners who will embody the role of the teacher in one or more skills. It mainly corresponds to an assessor's behavior: 87% of its positions are the source of the edges. Moreover, they are those with the lowest skill acquisition rank, which highlights their early involvement in courses.

Table 1. Roles description: Frequencies of positions within graphlets (multiplied per 100). The four left positions reflect active validation, the 3 positions in the middle of the table are intermediate ones, whereas the 4 ones on the right of the table represent passive postures in the validation process. The roles are also classified vertically, depending on their ratio of teachers.

	Assessor's positions				Mixed positions			Assessee's positions			
											
Peer Assessed Learner (Late)					1	2		27	54	11	1
Expert Assessed Learner					1	1		91	4	1	1
Peer Assessed Learner	1	2			4	5	1	60	16	7	2
Proxy Teacher	17	19	4	54		2					
Committed Learner	18	13	1	11	14		2	22	5	3	1
Expert	64	11	5	7	5		1	3	1		

Table 2. The left side of the table presents the distribution of roles among all courses according to student or teacher status. The right side characterizes them upon their commitment to the mutual validation of skills. Student peer-reviewed rate is the average rate of evaluations made by students. Skill acquisition rank orders the learners of a course in a normalized way (between 0 and 1).

	Students	Teachers	Student-reviewed rate (mean)	Skill acquisition rank (median)
Expert	14	40	0.46	0.27
Proxy Teacher	44	3	0	0.45
Committed Learner	146	19	0.75	0.40
Expert Assessed Learner	239	1	0.34	0.43
Peer Assessed Learner	248	4	0.79	0.53
Peer Assessed Learner (Late)	39	0	0.94	0.74

- **Proxy Teacher** While holding the Proxy Teacher role, a user (mostly a student) focuses exclusively on validating the skills of her/his peers. The positions are concentrated on assessor's ones, with a majority of co-validation with another assessor (the Experts are conversely the only assessor most of the times). While being themselves validated, the users in this role never ask for peer-validation, which is consistent with their early involvement in the courses (like the Experts).
- **Committed Learner** The Committed Learner is balanced in her/his approach to mutual validation of competence: s/he quickly validates her/his skills (0.40 skill acquisition rank) and adopts a reflective approach over the same period by validating other learners. It is the second role with the highest number of teachers but also one of the most demanding in terms of student-review.

Assessee Oriented Roles

- **Expert Assessed Learner** The Expert Assessed Learner is focused on getting assessed, and relies almost exclusively on an Expert. Indeed, 91% of their positions are in the target of the star  and they have a low student-reviewed rate (0.34). This role is played by students (all occurrences but one) who obtain their skills rather quickly (0.43 skill acquisition rank) which may indicate that they have had no choice but to ask their teacher for validation.
- **Peer Assessed Learner** The Peer Assessed Learner is mainly focused on receiving validation of her/his skills and relies mainly on students to do so. S/He's mostly validated within the star pattern , but also appears in  and . This role is played by students (248 over 252) who acquire their skills more slowly (0.53 skill acquisition rank) than Expert Assessed Learners.
- **Peer Assessed Learner (Late)** This role includes a sub-category of Peer Assessed Learners who validate their skills later relatively to others. Their positions are distributed among the assessee's positions with a high rate of  (transitively validated). These students are the last ones to validate skills (0.74 skill acquisition rank), which allows their validators (94% of students) to validate some skills before proposing their own assessments.

This topological description, combined with behavioral indicators, highlights the variety of roles taken by students and teachers throughout a course. These roles make it possible to suggest several hypotheses about learners' behavior. To do so, it is interesting to investigate the changes in roles during the different phases of the course. Therefore we provide a longitudinal study through a Sankey diagram, a visualization tool allowing to describe node flows between different steps [15].

4.4 Examples of a Course's Dynamics

Figure 2 depicts the evolution of roles (each color line) within the phases (each vertical line) of one of the courses. It is interesting to note that the course begins with a first phase where the teacher, in Expert role, validates skills to some Expert Assessed Learners. In a second phase, more students arrive. They occupy all different roles and many Expert Assessed Learners from the first phase are now acting as Experts and Committed Learners.

We also may focus on some specific phases, for instance phases 9 et 10, whose networks G_9 and G_{10} are showed in Fig. 2. In phase 9, that lasts 10 days, a student, user 54, is acting as an Expert, i.e. as a source of many stars , since he or she is the only assessor of several other students. The teacher, user 5, appears as a Committed Learner since s/he shares a lot of co-assessor's positions  (all the students s/he validated are co-validated by other users) and in the first position of the transitivity subgraph  with user 54 next to him .

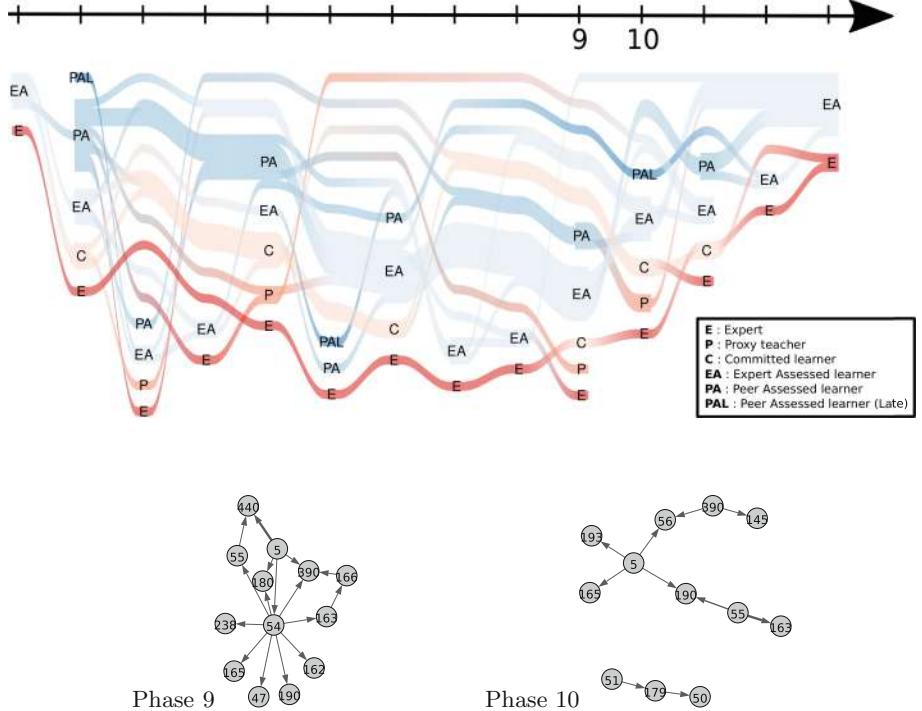


Fig. 2. The Sankey diagram representing the evolution of roles through the 14 phases of the course and the networks representing the validations that occurred during phases 9 and 10. Edge width depicts the number of validations (not taken into account in our measures).

Phase 10 arrives 35 days after phase 9 and lasts 11 days. As the teacher got a central position, s/he also occupies the Expert role again. Meanwhile, phase 9's very active user 54 has disappeared. Two students, 55 and 390, act as Proxy Teachers since they appear in the central position of star and in the co-assessor's position. Note that between the two phases, most of the students that were considered as Peer Assessed Learners have moved to Proxy Teachers, and are then validating their classmates.

4.5 Overall Role Evolution

In a complementary way, we finally consider the overall transitions between roles. For each user u , we filter the snapshots G_t where u is active. From the roles occupied by u in each of these graphs, we get a sequence of roles. For each user, we can thus compute the number of times he or she passes from a role to another.

The overall role transition, depicted in Fig. 3. A first remark is that most of the roles are somehow flexible even though most of the students seem to oscillate



Fig. 3. Percentage of evolution from one role to another

between roles Expert Assessed Learner and Peer Assessed Learner, which mainly differ from their propensity to be validated by the teacher. Note however that these roles also partly lead to the more active ones—Proxy Teacher and Committed Learner. Thus following the evolution of these roles could be a deeper insight.

Let us notice that the role Expert Assessed Learner seems to be a sink for other student roles: more than 50% of the Expert Assessed Learners (Late or not) and the Peer Assessed ones become Expert Assessed Learners in the next phase. In the meanwhile, as said before, this role is not trapping the students in an posture of being evaluated by a teacher and leads also to assessor's positions. Its central position in the transition table invites for further investigations.

Concerning the assessors, it is notable that the occupiers of roles Expert and Committed Learners flow to every other roles except for the Expert Assessed Learner (Late) who are the last to validate their skills. Here it could be interesting to separate the students from the teachers in order to see if the latter ones are even more strictly confined to assessor's positions (and which ones), or conversely if teachers sometimes become assesses.

5 Conclusion

We have introduced a methodology to detect roles in temporal networks. We have proposed a general framework for the development of a position-based embedding of nodes in snapshots. Thus a sequence of embedding reflects the evolution of topological roles that nodes occupy at different stages of an interaction network. We have illustrated this methodology on a case study based on the learning analytics field and demonstrated the power of expressiveness of graphlets and position-based embeddings.

Our work opens a number of avenues for additional research. Within a course, the teacher may authorize the self-assessment of skills. Enumerating graphlets with loops would reveal the influence of self-evaluation on the peer review mechanism. Within a curriculum it could be interesting to observe the evolution of a student's behaviour, and especially to use graphlets for the early detection of drop-out risk, which is a particularly important subject in the field of learning analytics that we have not yet addressed.

Another follow-up is the use of another kind of graphlets to track role through snapshots in order to compare the results. These graphlets could be size-4 graphlets, others graphlets from literature [14] or disconnected graphlets [25], for instance. Another possible approach would be to focus on position transitions to characterize the role of each user through time.

References

1. Araujo, M., Papadimitriou, S., Günnemann, S., Faloutsos, C., Basu, P., Swami, A., Papalexakis, E.E., Koutra, D.: Com2: fast automatic discovery of temporal ('comet') communities. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 271–283. Springer (2014)
2. Authier, M., Lévy, P.: Les arbres de connaissances. La Découverte, Paris (1999). <https://www.cairn.info/les-arbres-de-connaissances--9782707130440.htm>
3. Braha, D., Bar-Yam, Y.: Time-dependent complex networks: dynamic centrality, dynamic motifs, and cycles of social interactions. In: Adaptive Networks, pp. 39–50. Springer (2009)
4. Cela, K.L., Sicilia, M.Á., Sánchez, S.: Social network analysis in e-learning environments: a preliminary systematic review. Educ. Psychol. Rev. **27**(1), 219–246 (2015)
5. Charbey, R., Prieur, C.: Stars, holes, or paths across your facebook friends: a graphlet-based characterization of many networks. Network Sci. 1–22 (2019)
6. Dalgaard, C.: Social software: E-learning beyond learning management systems. Eur. J. Open, Distance e-learning **9**(2) (2006)
7. Dillenbourg, P.: What do you mean by collaborative learning? (1999)
8. Dunlavy, D.M., Kolda, T.G., Acar, E.: Temporal link prediction using matrix and tensor factorizations. ACM Trans. Knowl. Discov. Data (TKDD) **5**(2), 10 (2011)
9. Ferguson, R., Shum, S.B.: Social learning analytics: five approaches. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, pp. 23–33. ACM (2012)

10. Gartner, A., Kohler, M., Riessman, F., Grosjean, M.: Des enfants enseignent aux enfants: apprendre en enseignant. Hommes et groupes, Epi (1973). <https://books.google.fr/books?id=y8DuPAAACAAJ>
11. Gu, S., Milenkovic, T.: Graphlets versus node2vec and struc2vec in the task of network alignment. arXiv preprint [arXiv:1805.04222](https://arxiv.org/abs/1805.04222) (2018)
12. Holland, P.W., Leinhardt, S.: Local structure in social networks. Soc. Methodol. **7**, 1–45 (1976)
13. Holme, P., Saramäki, J.: Temporal networks. Phys. Reports **519**(3), 97–125 (2012)
14. Hulovatyy, Y., Chen, H., Milenković, T.: Exploring the structure and function of temporal networks with dynamic graphlets. Bioinformatics **31**(12), i171–i180 (2015)
15. Komarek, A., Pavlik, J., Sobeslav, V.: Network visualization survey. In: Computational Collective Intelligence, pp. 275–284. Springer (2015)
16. Kovanen, L., Karsai, M., Kaski, K., Kertész, J., Saramäki, J.: Temporal motifs in time-dependent networks. J. Stat. Mech. Theory Exp. **2011**(11), P11005 (2011)
17. Lloyd, S.: Least squares quantization in PCM. IEEE Trans. Inf. Theory **28**(2), 129–137 (1982)
18. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. Oakland, CA, USA (1967)
19. Milenković, T., Pržulj, N.: Uncovering biological network function via graphlet degree signatures. Cancer informatics **6**, CIN–S680 (2008)
20. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science **298**(5594), 824–827 (2002)
21. Paranjape, A., Benson, A.R., Leskovec, J.: Motifs in temporal networks. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, pp. 601–610. ACM (2017)
22. Paredes, W.C., Chung, K.S.K.: Modelling learning & performance: a social networks perspective. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. LAK 2012, pp. 34–42 ACM, New York (2012). <http://doi.acm.org/10.1145/2330601.2330617>
23. Pržulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geometric? Bioinformatics **20**(18), 3508–3515 (2004)
24. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
25. Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., Borgwardt, K.: Efficient graphlet kernels for large graph comparison. In: Artificial Intelligence and Statistics, pp. 488–495 (2009)
26. Suh, H., Kang, M., Moon, K., Jang, H.: Identifying peer interaction patterns and related variables in community-based learning. In: Proceedings of the 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: The Next 10 Years! CSCL 2005, pp. 657–661. International Society of the Learning Sciences (2005). <http://dl.acm.org/citation.cfm?id=1149293.1149379>
27. Teplovs, C., Fujita, N., Vatrapu, R.: Generating predictive models of learner community dynamics. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge. LAK 2011, pp. 147–152. ACM (2011)
28. Wernicke, S., Rasche, F.: Fanmod: a tool for fast network motif detection. Bioinformatics **22**(9), 1152–1153 (2006)



Enumerating Isolated Cliques in Temporal Networks

Hendrik Molter^(✉), Rolf Niedermeier, and Malte Renken

Faculty IV, Algorithmics and Computational Complexity, TU Berlin,
Berlin, Germany
`{h.molter,rolf.niedermeier,m.renken}@tu-berlin.de`

Abstract. Isolation is a concept from the world of clique enumeration that is mostly used to model communities that do not have much contact to the outside world. Herein, a clique is considered *isolated* if it has few edges connecting it to the rest. Motivated by recent work on enumerating cliques in temporal networks, we bring the isolation concept to this setting. We discover that the addition of the time dimension leads to six distinct natural isolation concepts. Our main contribution is the development of fixed-parameter enumeration algorithms for five of these six clique types employing the parameter “degree of isolation”. On the empirical side, we implement and test these algorithms on (temporal) social network data, obtaining encouraging preliminary results.

Keywords: Community detection · Dense subgraphs · Social network analysis · Time-evolving data · Fixed-parameter tractability

1 Introduction

“Isolation is the one sure way to human happiness.” – Glenn Gould

Clique detection and enumeration is a fundamental primitive of complex network analysis. In particular, there are numerous approaches (both from a more theory-based and from a more heuristic side) for listing all maximal cliques (that is, fully-connected subgraphs) in a graph.¹ It is well-known that finding a maximum-cardinality clique is computationally hard (NP-hard, hard in the approximation sense and hard when parameterized by the clique cardinality). Hence, heuristic approaches usually govern computational approaches to clique finding and enumeration. There have been numerous efforts to provide both theoretical guarantees and practically useful algorithms [4, 9–11]. In particular, to simplify (in a computational sense) the task on the one hand and to enumerate more meaningful maximal cliques (for specific application contexts) on the

¹ *Network* and *Graph* are used interchangeably.

Full version available on arXiv (<https://arxiv.org/abs/1909.06292>).

Supported by the DFG, project MATE (NI 369/17).

other hand, Ito and Iwama [10] introduced and investigated the enumeration of maximal cliques that are “isolated”. Roughly speaking, isolation means that the connection of the maximal clique to the rest of the graph is limited, that is, there are few edges with one endpoint in the clique and one endpoint outside the clique; indeed, the degree of isolation can be controlled by choosing specific values of a corresponding isolation parameter. For instance, think of social networks where one wants to spot more or less segregated sub-communities with little interaction to the world outside but intensive interaction inside the community. We mention in passing that, recently, there have been (only) theoretical studies on the concept of “secludedness” [2, 3] which is somewhat similar to the older isolation concept: whereas for isolation one requests “few outgoing edges”, for secludedness one asks for “few outneighbors”; while finding isolated cliques becomes tractable [10], finding secluded ones remains computationally hard [2].

Ito and Iwama [10] showed that in static networks isolated cliques often can be enumerated efficiently; the only exponential factor in the running time depends on the “isolation parameter”, and so fairly isolated cliques can be enumerated quite quickly. In follow-up work, the isolation concept then was significantly extended and more thorough experimental studies (also with financial networks) have been performed [9, 11]. However, analyzing complex networks more and more means studying time-evolving networks. Hence, computational problems known from static networks also need to be solved on temporal networks (mathematically, these are graphs with fixed vertex set but a time-dependent edge set) [8, 12–14]. Thus, not surprisingly, the enumeration of maximal cliques has recently been brought to the temporal setting [1, 7, 15, 16]. While becoming algorithmically more challenging than in the static network case, nevertheless the empirical results that have been achieved are encouraging. In this work, we now fill a gap by proposing to bring also the isolation concept to the temporal clique enumeration context, otherwise using the same modeling of temporal cliques as in previous work.

We focus on two basic isolation concepts described by Komusiewicz et al. [11] for the static setting. More specifically, we consider “maximal isolation” (every vertex has small outdegree) and “average isolation” (vertices have small outdegree on average). We face a much richer modeling than in the static case since isolation can happen in two “dimensions”: vertices and time; for both we can consider maximum and average isolation. With this distinction, we end up with eight natural ways to model isolation, where two “pairs” of isolation models turn out to be equivalent, finally leaving six different temporal isolation concepts for further study.

Our main contributions are as follows: First, as indicated above, we identify six mathematically formalized concepts of isolation for temporal networks. Second, building on and extending the algorithmic framework of Komusiewicz et al. [11] for static networks, for small isolation values we provide efficient algorithms for five of our six isolated clique enumeration models and prove worst-case

performance bounds for them.² In this context, a main algorithmic contribution is the development of tailored subroutines (that are only partially shared between different isolation concepts). Finally, on the empirical side we contribute an encouraging first experimental analysis of our algorithms based on social network data. Our preliminary experiments indicate differences (mostly in terms of running time) but also (sometimes surprising) accordances between the concepts.

2 Preliminaries

In this section, we provide some basic notation and terminology, recall the isolation concepts for static graphs, and transfer them to temporal graphs.

Static Graphs. Graphs in this paper are assumed to be undirected and simple. To clearly distinguish them from temporal graphs, they are sometimes referred to as *static* graphs. Let $G = (V, E)$ be a static graph. We denote the vertex set of G with $V(G)$ and the edge set of G with $E(G)$. For $v \in V(G)$ we use $\deg_G(v)$ for the number of edges ending at v . For $v \in A \subseteq V$, $\text{outdeg}_G(v, A)$ denotes the number of edges with one endpoint v and the other one outside of A . We further use $\text{outdeg}_G(A) := \sum_{v \in A} \text{outdeg}_G(v, A)$ and $\delta_G(A) := \min_{v \in A} \deg_G(v)$. In all these notations, we omit the index G if there is no ambiguity.

Temporal Graphs and Temporal Cliques. A *temporal graph* is a tuple $\mathcal{G} = (V, E_1, \dots, E_\tau)$ of a vertex set V and τ edge sets $E_i \subseteq \binom{V}{2}$. The graphs $G_i := (V, E_i)$ are called the *layers* of \mathcal{G} . The *time edge set* $\mathcal{E}(\mathcal{G})$ (or \mathcal{E} if \mathcal{G} is clear from the context) is the disjoint union $\bigsqcup_{t=1}^\tau E_t$ of the edge sets of the layers of \mathcal{G} . For any $1 \leq a \leq b \leq \tau$ we define the (static) graphs $\bigcup_{t=a}^b G_t := (V, \bigcup_{t=a}^b E_t)$ and $\bigcap_{t=a}^b G_t := (V, \bigcap_{t=a}^b E_t)$.

Following the definition of Viard et al. [15], a Δ -*clique* (for some $\Delta \in \mathbb{N}$) of \mathcal{G} is a tuple $(C, [a, b])$ with $C \subseteq V$ and $1 \leq a \leq b \leq \tau$ such that C is a clique in $\bigcup_{i=t}^{t+\Delta} G_i$ for all $t \in [a, b - \Delta]$.

One easily observes that $(C, [a, b])$ is a Δ -clique in \mathcal{G} if and only if $(C, [a, b - \Delta])$ is a 0-clique in $\mathcal{G}' = (V, E'_1, \dots, E'_{\tau-\Delta})$ where $E'_i := \bigcup_{t=i}^{i+\Delta} E_t$. Due to this, in our theoretical results we will only concern ourselves with $\Delta = 0$ and simply refer to 0-cliques as *temporal cliques*.

Temporal Isolation. We first introduce the isolation concepts for static graphs and then describe how we transfer them to the temporal setting. In a (static) graph G , a clique $C \subseteq V(G)$ is called *avg-c-isolated* if $\text{outdeg}_G(C) < c \cdot |C|$ where $c \in \mathbb{Q}$ is some positive number [10]. Further, it is called *max-c-isolated* if $\max_{v \in C} \text{outdeg}_G(v, C) < c$. Clearly max- c -isolation implies avg- c -isolation.

Moving to temporal graphs, we aim at defining an isolation concept for temporal cliques. Recall that a temporal clique consist of a vertex set and a time

² In terms of the language of parameterized algorithmics, we show that these cases are fixed-parameter tractable when parameterized by isolation value.

interval. We apply the isolation requirement both on a vertex and on a time level, meaning that for each dimension we can either require the average outdegree (as for static avg- c -isolation) or the maximum outdegree (as for static max- c -isolation) to be small. To make this more clear, we next provide some examples. For instance, we can require that, on average over all layers, the maximum outdegree in a layer is small. Or we can require that the average outdegree must be small in every single layer. Note that the ordering of the requirements for the time dimension and the vertex dimension also matters. Requiring the average outdegree to be small in every layer is different from requiring that, on average over all vertices, the maximum degree over all time steps must be small. Having two isolation requirements (avg and max) for two dimensions with two possible orderings, we arrive at eight canonical temporal isolation types. However it turns out that if we use the same requirement for both dimensions, they behave commutatively, so it boils down to six *different* temporal isolation types. In the following, we give a formal definition for each of the six temporal isolation types. To make the names less confusing, we use “usually” to refer to the avg isolation requirement in the time dimension and “alltime” to refer to the max isolation requirement in the time dimension.

Definition 1 (Temporal Isolation). Let $c \in \mathbb{Q}$. A temporal clique $(C, [a, b])$ in a temporal graph $\mathcal{G} = (V, E_1, \dots, E_\tau)$ is called

- alltime-avg- c -isolated if $\max_{i \in [a, b]} \sum_{v \in C} \text{outdeg}_{G_i}(v, C) < c \cdot |C|$,
- alltime-max- c -isolated if $\max_{v \in C} \max_{i \in [a, b]} \text{outdeg}_{G_i}(v, C) < c$,
- avg-alltime- c -isolated if $\sum_{v \in C} \max_{i \in [a, b]} \text{outdeg}_{G_i}(v, C) < c \cdot |C|$,
- max-usually- c -isolated if $\max_{v \in C} \sum_{i=a}^b \text{outdeg}_{G_i}(v, C) < c \cdot (b + 1 - a)$,
- usually-avg- c -isolated if $\sum_{i=a}^b \sum_{v \in C} \text{outdeg}_{G_i}(v, C) < c \cdot |C| \cdot (b + 1 - a)$, and
- usually-max- c -isolated if $\sum_{i=a}^b \max_{v \in C} \text{outdeg}_{G_i}(v, C) < c \cdot (b + 1 - a)$.

We define the set of *isolation types* as $\mathcal{I} = \{\text{alltime-max}, \text{alltime-avg}, \text{max-usually}, \text{usually-avg}, \text{avg-alltime}, \text{usually-max}\}$.

For all isolation types $I \in \mathcal{I}$, an I - c -isolated temporal clique $(C, [a, b])$ is called *time-maximal* if there is no other I - c -isolated clique $(C', [a', b'])$ with $C' \supseteq C$ and $[a', b'] \supset [a, b]$. If there is no I - c -isolated temporal clique $(C', [a', b'])$ with $C' \supset C$ and $[a', b'] \supseteq [a, b]$, then we call $(C, [a, b])$ *vertex-maximal*. We call $(C, [a, b])$ *maximal* if it is time-maximal and vertex-maximal.

Subsequently we give some intuition about the different isolation concepts. Note that for sufficiently small c they all converge to disallowing *any* outgoing edges. We start with the most restrictive and perhaps also most straightforward isolation type, that is, **alltime-max-isolation**. Here all vertices are required to have little or no outside contact at all times—think of a quarantined group. Slightly less restrictive is the notion of **avg-alltime-isolation**. Here it would be possible to have some distinguished “bridge” vertices inside the clique with relatively much outside contact, as long as most vertices never have many outgoing edges. If we reorder the terms, then we obtain **all-timeavg-isolation**. In contrast to the previous case, now the set of “bridge” vertices may be different

at any point in time. A typical situation where this could occur is that there is a low bandwidth connection between the clique and the rest of the graph, only allowing a limited number of communications to occur at any given moment. The next isolation concept, **usually-max-isolation**, can be seen as allowing short bursts of activity, in which some or even all vertices have many outgoing edges, as long as the entire clique is isolated most of the time. Again, if we reorder the terms, then we get a less restrictive concept (**max-usually-isolation**). Here, the bursts of activity may happen at different times for different vertices. Finally, **usually-avg-isolation** is the least restrictive of these notions, only limiting the total number of outside contacts over all vertices and layers that are part of the temporal clique.

We can make the following observation about the relations between different types of isolation. It is easily checked using the definitions above.

Observation 2. Let $\mathcal{G} = (V, E_1, \dots, E_\tau)$ be a temporal graph. The following nine implications hold for any $a \leq b$, any clique C in $\bigcap_{t=a}^b G_t$, and any $c > 0$:

$$\begin{array}{ccc}
(C, [a, b]) \text{ alltime-max-}c\text{-isolated} & \Longrightarrow & (C, [a, b]) \text{ avg-alltime-}c\text{-isolated} \\
& \Downarrow & \Downarrow \\
(C, [a, b]) \text{ usually-max-}c\text{-isolated} & & (C, [a, b]) \text{ alltime-avg-}c\text{-isolated} \\
& \Downarrow & \Downarrow \\
(C, [a, b]) \text{ max-usually-}c\text{-isolated} & \Longrightarrow & (C, [a, b]) \text{ usually-avg-}c\text{-isolated} \\
& \Downarrow & \Downarrow \\
C \text{ is max-}c\text{-isolated in } \bigcap_{i=a}^b G_i & \Longrightarrow & C \text{ is avg-}c\text{-isolated in } \bigcap_{i=a}^b G_i
\end{array}$$

Note that Observation 2 does not hold for *maximal* isolated temporal cliques: A maximal alltime-max- c -isolated clique is not necessarily a maximal usually-avg- c -isolated clique. In the full version, we prove several results that help to confine the search space of our algorithms. Here, we only mention the following two lemmata, which we will use in Sect. 3 to prove the correctness of our algorithm for the enumeration of avg-alltime-isolated cliques.

Lemma 3. Let G be a static graph and let C be a clique in G . Then, any avg- c -isolated subset $C' \subseteq C$ has size $|C'| > \delta(C) - c + 1$.

Lemma 4. Let C be a clique and let $C' \subseteq C$ be a maximal avg- c -isolated subset. Let $\tilde{C} \subseteq C$ be the $\delta(C) - c + 2$ vertices of minimal degrees. Then $\tilde{C} \subseteq C'$.

3 Enumerating Maximal Isolated Temporal Cliques

We develop efficient algorithms to enumerate maximal isolated temporal cliques for five out of the six introduced temporal isolation concepts (all except usually-max).³ In this section, we focus on one case (avg-alltime) to present some of the

³ One may wonder why usually-max-isolation was dropped here. The answer is that, even though the same approach as for the other isolation concepts also works for usually-max-isolation, we found no way to limit the work that would be required in the `isolatedSubsets()` subroutine significantly below $\Omega(2^n)$.

Table 1. Running time of our maximal isolated temporal clique enumeration algorithms for the different temporal isolation types.

alltime-avg	alltime-max	avg-alltime	max-usually	usually-avg
$\mathcal{O}(c^c \tau^2 \cdot V \cdot \mathcal{E})$	$\mathcal{O}(2.89^c c \tau \cdot \mathcal{E})$	$\mathcal{O}(5.78^c c \tau \cdot \mathcal{E})$	$\mathcal{O}(2.89^c c \tau^3 \cdot \mathcal{E})$	$\mathcal{O}(5.78^c c \tau^3 \cdot \mathcal{E})$

key ideas. Our algorithms have *fixed-parameter tractable* (FPT) running times for the isolation parameter c , that is, for fixed c , the running time is a polynomial whose degree does not depend on c . The following theorem summarizes our main result. Its proof is partially given at the end of this section. The remainder, including the proof of the running times listed in Table 1, can be found in the full version.

Theorem 5. *Let a temporal graph \mathcal{G} with τ layers, an isolation type $I \in \mathcal{I} \setminus \{\text{usually-max}\}$, and an isolation parameter $c \in \mathbb{Q}$ be given. Then all maximal I - c -isolated temporal cliques in \mathcal{G} can be enumerated in FPT-time for the isolation parameter c . The specific running times depend on I and are given in Table 1.*

Our algorithms are inspired by the algorithms for static isolated clique enumeration [10, 11] and build upon the fact that every maximal I - c -isolated temporal clique $(C, [a, b])$ is contained in some vertex-maximal c -isolated clique C' of $G_{\cap} := \bigcap_{t=a}^b G_t$ (by Observation 2). Algorithm 1 constitutes the top level algorithm. Here, we iterate over all possible time windows $[a, b]$ and apply the so-called trimming procedure developed by Ito and Iwama [10] to G_{\cap} to obtain, for each so-called *pivot vertex* v , a set $C_v \subseteq N[v]$ containing all avg- c -isolated cliques of G_{\cap} that contain v . Subsequently, we enumerate all maximal cliques within C_v and test each of them for maximal I - c -isolated subsets. For this step, we employ our theoretical results to quickly skip over irrelevant subsets. The details depend on the choice of I , as does the strategy for the last step, that is, removing non-maximal elements from the result set. Remember that we have to pay attention to both, time- and vertex-maximality. For the latter we can, in most cases, utilize an idea by Komusiewicz et al. [11].

Enumerating Isolated Subsets. We now discuss the `isolatedSubsets()` subroutine of Algorithm 1 (Line 11). While the details depend on the isolation type, there are two main flavors. For alltime-max-isolation and max-usually-isolation, it is possible to determine a single vertex that must be removed in order to obtain an isolated subset. By repeatedly doing so, one either reaches an isolated subset or the size threshold set by Lemma 3. In particular, each maximal clique contains at most one maximal isolated subset.

For usually-avg-isolation, avg-alltime-isolation (Function 1), and alltime-avg-isolation, multiple vertices are removal candidates. However, one can show that their number is still limited (for the case of avg-alltime, see Lemma 4). We therefore build a search tree, iteratively exploring removal sets of growing size. The case of alltime-avg-isolation is somewhat special, as here the set of removal candidates is different for each layer.

Algorithm 1. Enumerating maximal I - c -isolated cliques for $I \in \mathcal{I} \setminus \{\text{usually-max}\}$

Input: A temporal graph $\mathcal{G} = (V, E_1, \dots, E_\tau)$, a $c \in \mathbb{Q}$, and an isolation type $I \in \mathcal{I} \setminus \{\text{usually-max}\}$.

Output: All maximal I - c -isolated cliques in \mathcal{G} .

```

1 result ← {}
2 foreach  $a = 1 \dots \tau$  do
3   foreach  $b = a \dots \tau$  do
4     /* Here we are looking for cliques with lifetime  $[a, b]$ . */
5      $G_{\cap} \leftarrow \bigcap_{i=a}^b G_i$ 
6     Sort vertices by ascending degree in  $G_{\cap}$ 
7     foreach vertex  $v$  do
8       /* Vertex  $v$  is the pivot vertex. */
9        $C_v \leftarrow$  candidate set for pivot  $v$  after trimming stage (in  $G_{\cap}$ )
10       $k \leftarrow \lfloor \deg_{G_{\cap}}(v) - c + 2 \rfloor$  /* By Lemma 3, all isolated cliques
11        are at least this large. */
12       $\mathcal{C} \leftarrow$  set of all maximal cliques of size at least  $k$  in  $C_v \subseteq G_{\cap}$ 
13      foreach  $C \in \mathcal{C}$  do
14        subsets ←  $I$ -isolatedSubsets( $C$ ,  $[a, b]$ ,  $\deg_{G_{\cap}}(v)$ )
15        result ← result ∪  $\{(C, [a, b]) \mid C \in \text{subsets}\}$ 
16      end
17    end
18  end
19  foreach  $(C, [a, b]) \in \text{result}$  do
20    if  $I$ -isMaximal( $C, [a, b]$ ) then
21      output  $(C, [a, b])$ 
22    end
23  end

```

Checking for Maximality. We now informally discuss the `isMaximal()` subroutine of Algorithm 1 (Line 18). For a fully detailed description we refer to the full version. Note that, while each temporal clique $(C, [a, b])$ returned by `isolatedSubsets()` is vertex-maximal within its respective set C_v , it may be not vertex-maximal with regard to the entire graph. Moreover, we need to check for maximality with regard to cliques with a larger time window. The naïve approach of pairwise comparing all elements of the result set is feasible but inefficient. Instead, for alltime-max-isolation, alltime-avg-isolation, and avg-alltime-isolation it is sufficient to only check whether the time window can be extended in either direction, and whether a larger clique exists within the same time window (i.e., checking vertex-maximality). Except for the case of alltime-avg-isolation, the latter can again be implemented more efficiently than by using pairwise comparisons. We modify the maximality test developed by Komusiewicz et al. [11] which searches for cliques within the common neighborhood of C and then checks whether these can be used to build a larger isolated clique.

Function 1. avg-alltime-isolatedSubsets(C , $[a, b]$, δ)

```

1  $\forall v \in C : s_v := \max_{i \in [a, b]} \deg_{G_i}(v)$ 
2  $d := \lfloor |C| - \delta + c - 2 \rfloor /*$  By Lemma 3, we only remove top  $d$  vertices. */
3  $\{v_i \mid 1 \leq i \leq d\} :=$  the  $d$  vertices in  $C$  with the highest values of  $s_v$ 
4  $\mathcal{D}' \leftarrow \{\emptyset\}$ 
5 result  $\leftarrow \emptyset$ 
6 while  $\mathcal{D}' \neq \emptyset$  do
7    $\mathcal{D} \leftarrow \mathcal{D}'$ 
8    $\mathcal{D}' \leftarrow \emptyset$ 
9   foreach  $D \in \mathcal{D}$  do
10    |  $C' \leftarrow C \setminus D$ 
11    | if  $\sum_{v \in C'} s_v \geq |C'| \cdot (|C'| - 1 + c)$  then
12    | |  $j := \max\{0, i \mid v_i \in D\}$ 
13    | |  $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{D \cup \{v_i\} \mid j < i \leq d\}$ 
14    | end
15    | else
16    | | result  $\leftarrow$  result  $\cup \{C \setminus D\}$ 
17    | end
18  end
19 end
20 return result

```

Correctness. We now show the correctness of our algorithm. We first prove that the `isolatedSubsets()` function (Function 1) behaves as intended.

Lemma 6. Let $\mathcal{G} = (V, E_1, \dots, E_\tau)$ be a temporal graph, $c \in \mathbb{Q}$, and $I \in \mathcal{I} \setminus \{\text{usually-max}\}$. Let C be a clique in $G_\cap := \bigcap_{i=a}^b G_i$ and $\delta = \delta_{G_\cap}(C)$. Then I -`isolatedSubsets`(C , $[a, b]$, δ) returns all maximal sets $\tilde{C} \subseteq C$ such that $(\tilde{C}, [a, b])$ is I - c -isolated.

Proof (for the case $I = \text{avg-alltime}$). For the sake of brevity, we will simply write that some set $X \subseteq C$ is, say, alltime-avg-isolated to denote that $(X, [a, b])$ is alltime-avg-isolated. Let $B := \{v_i \mid 1 \leq i \leq d\}$ and let $\tilde{C} \subseteq C$ be any maximal avg-alltime- c -isolated subset. Note that any subset of C is avg-alltime- c -isolated if and only if the same set was avg- c -isolated in a static graph where each vertex' degree was set to $\max_{i \in [a, b]} \deg_{G_i}(v)$. By applying Lemma 4 to this auxiliary graph, we see that \tilde{C} must contain $C \setminus B$. Thus, we observe that the algorithm will eventually check \tilde{C} . \square

The algorithms for the other isolation types and the corresponding proofs of correctness are part of the full version. Thus, we now have all the necessary pieces to prove the correctness of Algorithm 1.

Proof (of Theorem 5). The running time of Algorithm 1 is proven in the full version, as is the correctness of the `isMaximal()` subroutine. Thus it remains to show that the result set contains (at least) all maximal I - c -isolated temporal

cliques. To this end, let $(C, [a, b])$ be any maximal I - c -isolated clique. Then, C is an avg- c -isolated clique in $G_{\cap} = \bigcap_{a \leq i \leq b} G_i$ by Observation 2. Ito and Iwama [10] showed that we then have $C \subseteq C_v$ where $v \in C$ is of minimum degree. Further $|C| \geq |C_v| - k$ by Lemma 3. Thus, $C \subseteq C'$ for some $C' \in \mathcal{C}$, so $(C, [a, b])$ is added to the result set by Lemma 6. \square

4 Experimental Evaluation

In this section, we empirically evaluate the running times of our enumeration algorithms for maximal isolated Δ -cliques (Algorithm 1) on several real-world temporal graphs. In particular, we investigate the effect of different isolation concepts as well as different values for isolation parameter c and Δ (now allowing $\Delta > 0$, see the definition of Δ -cliques in Sect. 2) on the running time and on the number of cliques that are enumerated. We also draw some comparisons concerning running times to a state-of-the-art algorithm to enumerate maximal (non-isolated) Δ -cliques by Bentert et al. [1].

Setup and Statistics. We implemented our algorithms⁴ in Python 3.6.8 and carried out experiments on an Intel Xeon E5-1620 computer clocked at 3.6 GHz and with 64 GB RAM running Debian GNU/Linux 6.0. The given times refer to single-threaded computation. Bentert et al. [1] implemented their algorithm in Python 2.7.12.

For the sake of comparability we tested our implementation on four freely available data sets; however, due to space restrictions, we only present a subset of the results here. The remainder can be found in the full version.

- Face-to-face contacts between high school students (“highschool-2011” [5], 126 vertices, 28560 time edges, 272330 time steps),
- Spatial proximity between persons in a hospital (“tij_pres_LH10” [6], 73 vertices, 150126 time edges, 259180 time steps).

The “highschool-2011” data set was also used by Bentert et al. [1].

We chose five roughly exponentially increasing values ε , 1, 5, 25, 125 for the isolation parameter c , where $\varepsilon := 0.001$ effectively requires complete isolation and $125 \approx |V|$ imposes little or no restriction. We chose our Δ -values in the same fashion as Bentert et al. [1]. In order to limit the influence of time scales in the data and to make running times comparable between instances, the chosen Δ -values of 0, 5^3 , and 5^5 were scaled by $L/(5 \cdot |\mathcal{E}|)$, where L is the temporal graph’s lifetime in seconds [7, Section 5.1].

Experimental Results. In Figs. 1 and 2 the number of maximal isolated Δ -cliques and the running time are plotted for each of the five isolation types and a range of isolation values c . Missing values indicate that the respective instance exceeded

⁴ The code of our implementation is freely available at <https://www.akt.tu-berlin.de/menue/software/>.

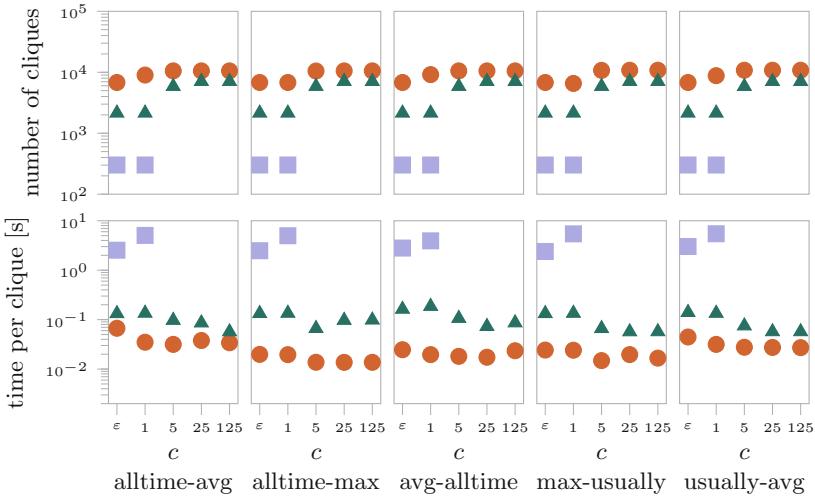


Fig. 1. Plot for the data set “highschool-2011” showing the number of cliques (top) and the computing time per clique (bottom) for the different temporal isolation types and different values of c and Δ . The different Δ -values are visualized by the different markers, with circles, triangles and squares denoting values of 0, 5^3 , and 5^5 , respectively.

the time limit of 1 h. In general, the different isolation types produce surprisingly similar outputs. This suggests that the degrees of the vertices forming an isolated Δ -clique are typically rather similar and remain constant over the lifetime of the clique. Unsurprisingly, raising the value of c increases the number of maximal cliques as the isolation restriction is weakened. However, this effect ceases roughly at $c = 5$. Increasing c further does not produce additional cliques, suggesting that the vertices in Δ -cliques we found in the data sets mostly have out-degree at most five. Furthermore, we can generally observe that the number of maximal cliques decreases with increasing values of Δ , which might seem unexpected at first glance, but is a consequence from finding many small cliques (with few vertices as well as short time intervals) for small Δ -values that “merge together” for larger Δ -values. This behavior is consistent across all data sets we investigated.

Regarding running time, our algorithm is generally slower than the non-isolated clique enumeration algorithm by Bentert et al. [1], even for small values of c . For comparison, the algorithm by Bentert et al. [1] solved the “highschool-2011” instance for the same values for Δ that we considered in less than 10 seconds per instance, while we needed up to 20 min for Δ -values 0 and 5^3 and more than one hour for 5^5 . We believe that the two main reasons for our algorithm to be slower are the following. On the one hand, the maximality check we perform is much more complicated than the one of the algorithm of Bentert et al. [1], which is an issue that also occurs in the static case [9, 11]. On the other hand, we have to explicitly iterate through more or less all possible intervals in which we could find an isolated Δ -clique, which seems unavoidable

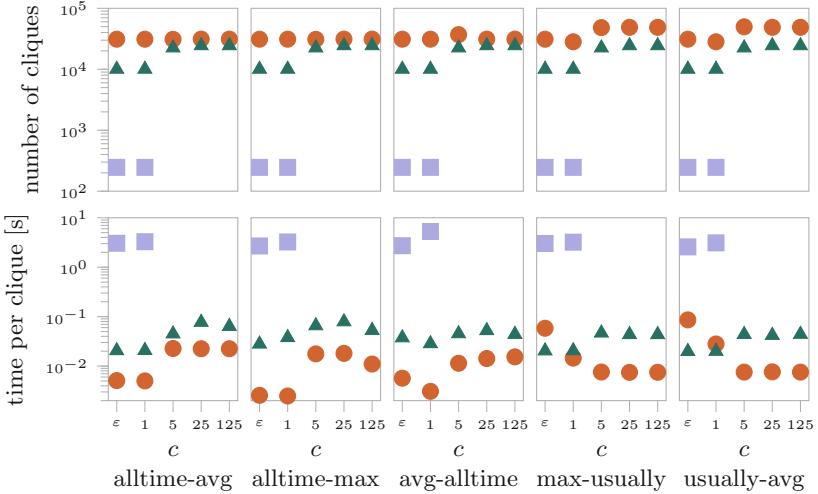


Fig. 2. Plot for the data set “tij-pres_LH10” (see also description of Fig. 1).

in our setting. A particular consequence of this is that our algorithm is not *output-sensitive*, that is, the running time can be much larger than the number of maximal isolated Δ -cliques in the input graph. In the case of (non-isolated) Δ -clique enumeration, there are ways to circumvent these issues and in particular, the algorithm of Bentert et al. [1] is output-sensitive. Both algorithms have a similar running time behavior with respect to Δ , that is, the running time increases with Δ , once Δ reaches moderately large values. Since higher values of Δ create a more dense graph after the preprocessing step, this behavior is expected. The algorithm of Bentert et al. [1] is slow for very small values of Δ that are close to zero (compared to itself for larger values of Δ). We do not observe this phenomenon in most of our algorithms. In the variants for max-usually and usually-avg, however, we experience a similar issue for small values of c , especially visible in the “tij-pres_LH10” data set (Fig. 2), where the running time is surprisingly high for $\Delta = 0$ and $c = \varepsilon$. A possible explanation is that the “usually-variants” use a different maximality check than the “alltime-variants”. Interestingly, no universal trend can be observed for the running time taken per resulting clique with respect to c , which stands in contrast to our theoretical running time analysis.

5 Conclusion

We have brought the concept of isolation from the static to the temporal setting, introducing six different types of temporal isolation. For five out of those we developed algorithms and showed that enumerating maximal temporally isolated cliques is fixed-parameter tractable with respect to the isolation parameter. This leaves one case (usually-max-isolation) open for future research on computational

complexity classification. As a rule of thumb, if there are no specific requests from the use case, we recommend to choose the alltime-max concept as a default, since overall it allowed for the fastest running times without huge differences in terms of enumerated maximal cliques.

From an algorithm engineering perspective there is still room for improvement. So far the practical running times make it hard to analyze larger data sets as done for example by Bentert et al. [1] in the “non-isolated” setting.

Finally, as in the static case, it would be natural to apply the isolation concepts to further community models such as for example temporal k -plexes [1, 11].

Acknowledgments. We want to thank our student assistant Fabian Jacobs for his work on the implementation of our algorithms and anonymous reviewers for helpful feedback.

References

1. Bentert, M., Himmel, A.S., Molter, H., Morik, M., Niedermeier, R., Saitenmacher, R.: Listing all maximal k -plexes in temporal graphs. ACM J. Exp. Algorithm. **24**(1), 1.13:1–1.13:27 (2019)
2. van Bevern, R., Fluschnik, T., Mertzios, G.B., Molter, H., Sorge, M., Suchý, O.: The parameterized complexity of finding secluded solutions to some classical optimization problems on graphs. Discret. Optim. **30**, 20–50 (2018)
3. Chechik, S., Johnson, M.P., Parter, M., Peleg, D.: Secluded connectivity problems. Algorithmica **79**(3), 708–741 (2017)
4. Eppstein, D., Strash, D.: Listing all maximal cliques in large sparse real-world graphs. ACM J. Exp. Algorithm. **18**, 1–3 (2013)
5. Fournet, J., Barrat, A.: Contact patterns among high school students. PLoS ONE **9**(9), 1–17 (2014)
6. Génois, M., Barrat, A.: Can co-location be used as a proxy for face-to-face contacts? EPJ Data Sci. **7**(1), 11 (2018)
7. Himmel, A.S., Molter, H., Niedermeier, R., Sorge, M.: Adapting the Bron-Kerbosch algorithm for enumerating maximal cliques in temporal graphs. Soc. Netw. Anal. Min. **7**(1), 35:1–35:16 (2017)
8. Holme, P., Saramäki, J.: Temporal networks. Phys. Rep. **519**(3), 97–125 (2012)
9. Hüffner, F., Komusiewicz, C., Moser, H., Niedermeier, R.: Isolation concepts for clique enumeration: comparison and computational experiments. Theor. Comput. Sci. **410**(52), 5384–5397 (2009)
10. Ito, H., Iwama, K.: Enumeration of isolated cliques and pseudo-cliques. ACM Trans. Algorithms **5**(4), 40:1–40:21 (2009)
11. Komusiewicz, C., Hüffner, F., Moser, H., Niedermeier, R.: Isolation concepts for efficiently enumerating dense subgraphs. Theor. Comput. Sci. **410**(38–40), 3640–3654 (2009)
12. Latapy, M., Viard, T., Magnien, C.: Stream graphs and link streams for the modeling of interactions over time. Soc. Netw. Anal. Min. **8**(1), 61 (2018)
13. Michail, O.: An introduction to temporal graphs: an algorithmic perspective. Internet Math. **12**(4), 239–280 (2016)
14. Rossetti, G., Cazabet, R.: Community discovery in dynamic networks: a survey. ACM Comput. Surv. **51**(2), 35 (2018)

15. Viard, T., Latapy, M., Magnien, C.: Computing maximal cliques in link streams. *Theor. Comput. Sci.* **609**, 245–252 (2016)
16. Viard, T., Magnien, C., Latapy, M.: Enumerating maximal cliques in link streams with durations. *Inf. Process. Lett.* **133**, 44–48 (2018)

Networks in Finance and Economics



A Partially Rational Model for Financial Markets: The Role of Social Interactions on Herding and Market Inefficiency

Lorenzo Giannini¹, Fabio Della Rossa^{1,2}, and Pietro DeLellis^{1(✉)}

¹ Department of Electrical Engineering and Information Technology,
University of Naples Federico II, Via Claudio 21, 80125 Naples, Italy
pietro.delellis@unina.it

² Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

Abstract. This work investigates how social influence affects the collective behavior of interconnected financial agents in an artificial market. Each agent bases her trading decisions on her perceived value of the traded asset. If the interconnections between agents are not considered, an efficient market emerges, where the intrinsic value of the traded asset is correctly estimated. In the presence of social interactions, modeled through a scale-free network, the trading decisions of each agent also depends on the perception her neighbors have on the asset value. We illustrates how sociality can yield herding, which in turn degrades market efficiency and stability. Then, we propose a control strategy to mitigate herding so as to reduce volatility and regain market efficiency.

Keywords: Herding control · Fundamentalist market models · Market bubble · Agent-based modeling · Complex networks

1 Introduction

Any tradable good has an associated intrinsic value, depending on the demand/offer, the riskiness, availability and many other non-trivial factors. This value is often referred as the *fundamental value* of the asset, which is not always easy to be determined. If the good is traded in a market, a proxy of its value is its *market price*, as it reflects the general opinion that market participants have on its value. Neoclassical economics assumes that any financial agent is driven by rationality, using the “utility maximization” principle when taking trading decisions. Moreover, it postulates that each agent acts independently on the basis of full and relevant information [23]. Under these assumptions, *market efficiency*, defined as the ability of a market price to fully reflect all available information, is achieved. Indeed, financial agents form rational expectations about future price variations, thus buying or selling accordingly and driving the market price to the correct fundamental value. These considerations are the foundations of the so-called *efficient market hypothesis*, which explains the price formation in financial markets and was formalized by Eugene Fama in 1970 [14].

This theory, widely accepted in the second half of the 20th century, in the last twenty years has been criticized as it fails to explain periods when the market consistently overestimates the value of an asset, in spite of the available information [7]. In fact, the formation of the expectations of financial agents is not purely rational, and are not independent of the opinions and choices of the other investors. For instance, if speculators are noticing an increased demand of a specific asset, they are more prone to buy it to make profit from the positive trend, thus driving the price even higher, while for the same reason an investor will probably refrain from selling. Similar behaviors can be observed in several financial bubbles in the economics history. For example, at the beginning of the 21th century, a collective market phenomenon, later named *dot-com boom*, took place. The rising evaluation of internet companies paired with high confidence in future profits led many investors to overlook traditional metrics of profitability [22]. To illustrate the impact of this phenomenon on market prices, in Fig. 1 we report the time series of the closing values of the NASDAQ Composite and S&P 500 indexes from 1990 to 2008. We can observe that the NASDAQ Composite, which is more focused on IT companies, in 1999 started to diverge from the S&P 500 to than abruptly become comparable again from 2001, due to the dramatic increase and subsequent collapse of IT companies.

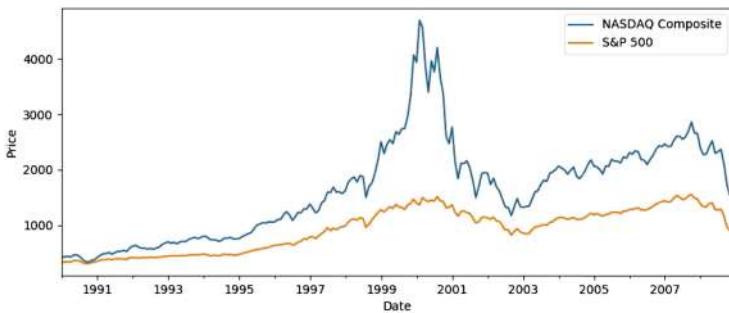


Fig. 1. NASDAQ composite and S&P 500 closing values (Source: Yahoo Finance).

The study of this kind of phenomena, referred to as *market bubbles*, has attracted great research interest due to their potentially disruptive socio-economic impacts [17, 20]. Market crashes are observed when a significant difference between the market prices and the real value of the traded assets, thus they cannot be explained under the efficient market assumption. The introduction of the *human factor* in economic theory, intended as the intrinsic irrationality of human decision making process, has been a fundamental step towards a better understanding on how market inefficiencies, bubbles and crisis arise, giving rise to new branches of economics such as behavioral [2] and experimental [12] economics. Recently, it has been proposed that one of the possible explanations of market bubbles lies in the fact that a financial agent, which is not aware of the

intrinsic value of a resource, uses the opinion of her peers as a proxy for its value. This phenomenon of collective imitation, referred to as *herding*, is blamed as one of the possible causes of market instability, which reinforces agent expectations [10], and leads to increased volatility [4].

Several models tried to reproduce empirical the so-called *stylized facts* observed in real financial markets, but none of them allowed to establish a direct link between sociality and market efficiency. The authors of [19] used statistical tools and calibration on real data to reproduce market dynamics, thus providing an accurate descriptions of the financial market under analysis, but they did not focus on how investor decision making impacts on market dynamics. In [6, 18], the investors are divided into fundamentalists-agents that base their decisions on their estimation of the fundamental price- and chartists-agents that follow the market trends. These models show that these assumptions explains the market volatility increases, but do not explicitly consider herding. Recently, an attempt to introduce the imitation mechanism in a market model has been done in [21]. However, the fundamental value of the asset is not explicitly modeled, thus preventing the possibility to use the model to characterize the effect of herding on efficiency. Herding behavior has also been studied in [8], but each agent was randomly assigned to her *herding group*, leaving the communication structure present between the agents unmodelled. Other approaches, instead, leveraging tools from complex networks theory, decided to neglect the complexity of the agent individual behavior to focus on the impact of herding [11].

The aim of this work is to elucidate the impact of social interactions on price formation in a realistic agent-based artificial market model. Specifically, we consider a centralized market with a double auction order book, where social interactions are modeled through the use of a static directed scale-free network. Agent decision making is mainly fundamentalist; compared to other models with fundamentalists, the agents are not perfectly aware of the current fundamental price, but they can form their own prediction. To the best of our knowledge, this is the first model that pairs non-trivial agent decision making with an explicit modeling of the herding behavior.

2 Novel Artificial Financial Market Model

2.1 Brief Overview on Double Auction Markets

Most computerized public markets are *order driven*. A centralized entity collects all bids and offers made by traders willing to partake in the market and match them to allow the trade to happen. All requests to trade, or *orders*, are stored in informative structures called *books*. This kind of structure is called *double auction* [15]. Typically, the centralized entity also takes care of finalizing the settlement between market participants, acting as a *clearer*. To mitigate the risk associated to the *trading* activity, the clearer requests a precautionary cash deposit, called *margin*, from anyone who is willing to trade. The traders express their willingness to buy (sell) a specific amount of goods at a *limit price*, which is the highest (lowest) price they are willing to set for a trade. The order is stored in the book

until another trader is willing to meet the same price request, then there is the order *execution* as the clearer finalizes the transaction. An order waiting in the order book for an execution is called *limit order*. When a trader is willing to buy (sell) at a price another agent has already submitted a sell (buy) order, then the order submitted is referred to as a *market order*, and it is immediately executed.

The orders stored in the books also determine the price of an asset. Specifically, the highest price any buyer is willing to pay, so the highest price of all the N_{bid} buy limit orders, is the *bid price* p_{bid} , while the lowest price a seller is willing to get paid, i.e., the lowest price of all the N_{ask} sell limit orders, is the *ask price* p_{ask} . These are often also called *best quotes* as they represent the best price available for a market order. The difference between the bid price and ask price is called *bid/ask spread*, while their average is called *mid price*. Limit orders with the best price for the counterpart are the first to be executed; since bid and ask are the best available prices, market orders cause executions only on limit orders with limit price equal to the current bid and ask prices. When multiple orders have the same price, the first submitted is the first to be executed. The price of the last transaction is the current asset price p_{cur} . The asset price at the end of day t is called the market price $p(t)$.

2.2 Our Model

The financial agents trade via the order book, submitting market or limit orders by optimizing their expectation about future prices. We assume that they possess limited resources, consisting in cash amount $c_i(t)$ and asset availability $a_i(t)$. The wealth $W_i(t)$ of agent i at the end of day t is the sum of all the available cash and assets, namely

$$W_i(t) = c_i(t) + a_i(t)p(t) \quad (1)$$

Different from existing literature, our model accounts for the social interaction among the investors. Therefore, we assume that the agents interact on a weighted digraph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ where each of the N nodes in \mathcal{V} is a financial agent, and the presence of an edge $(i, j) \in \mathcal{E}$ implies that node i has an influence over node j . Furthermore, we associate to each pair of agent $(i, j) \in \mathcal{E}$ a weight $w_{ij}(t)$ representing the strength of the influence. This implies that, when forming their investment decision, the i -th agent will only consider the decisions of the agents in her *in-neighborhood* $\mathcal{N}_i^{\text{in}}$, that is, the set of nodes having outgoing edges towards i . Among her in-neighbors, agent i will take in greater consideration the decision of the wealthier through the weights w_{ij} , as wealth is considered as a proxy of success. Accordingly, the time-varying edge weights are updated at the end of each day of trading according to the relative wealth between the agents

$$w_{ij}(t) = W_i(t)/W_j(t). \quad (2)$$

In a day t , each agent is picked in a random sequence to compute a prediction of her future returns, then submits market orders, and can decide to cancel her non-executed orders from the order book.

Return Prediction. At each day t , the i -th agent makes a prediction on the expected log-return $\hat{r}_i(t)$. The expected log-return is a convex combination of the agent's prediction on the fundamental price $\hat{p}_{f,i}(t)$ and a social prediction of the log-return $\hat{r}_{s,i}(t)$, which is defined as a weighted average of the expectations of the in-neighbors, with weights computed as in (2).

Individual Prediction. The agents form their individual expectations on future prices based on the available information, which might differ from one agent to another. The dynamic price prediction $\hat{p}_{f,i}(t)$ of agent i is modelled as the following Ornstein-Uhlenbeck process:

$$\partial \hat{p}_{f,i}(t) = \gamma_i(p_f(t) - \hat{p}_{f,i}(t))\partial t + \sigma_i \partial B_i(t), \quad (3)$$

where $\gamma_i(p_f(t) - \hat{p}_{f,i}(t))\partial t$ is a mean reversion term on the fundamental price, $\partial B_i(t) \sim \mathcal{N}(0, 1)$ is a Wiener process; γ_i represents the strength of attraction towards the mean, which can be interpreted as a measure of the agent *expertise*, intended as her ability of correctly estimating the fundamental price. The variance σ_i^2 of the process B_i models the *insecurity* of the agent. Indeed, a high value of σ_i implies that agent i will be more prone to randomly change her estimation of the expected returns. The log-return $r_{f,i}$ is then computed as

$$\hat{r}_{f,i}(t) = \log(\hat{p}_{f,i}(t)/p_{\text{cur}}), \quad (4)$$

where we recall that p_{cur} is the asset current price, i.e. the price of the last transaction of the current asset.

Social Prediction. The social prediction $\hat{r}_{s,i}$ of the log-return is computed as the weighted average of the predictions of the neighbours. Namely,

$$\hat{r}_{s,i}(t) = \frac{1}{d_i^{\text{in}}} \sum_{(j,i) \in \mathcal{E}} A_{ji} w_{ji}(t-1) \hat{r}_j(t-1), \quad (5)$$

where A_{ji} is the element (j, i) of the binary adjacency matrix of graph \mathcal{G} , and d_i^{in} is the in-degree of agent i , computed as the cardinality of her in-neighborhood set $\mathcal{N}_i^{\text{in}}$.

Return Prediction. The estimation that agent i performs of the future log-return is then computed as a convex combination of the individual and social predictions, that is,

$$\hat{r}_i(t) = (1 - \zeta) \hat{r}_{f,i}(t) + \zeta \hat{r}_{s,i}(t), \quad (6)$$

where the parameter $\zeta \in [0, 1]$ balances the relevance of $\hat{r}_{s,i}$ against the individual prediction of the return $\hat{r}_{f,i}$. The values 0 and 1 are representative of the extreme cases of non-social agents, that do not consider the opinions of the others, and fully-social agents, who completely disregard their individual predictions in favor of those of the agents in their neighborhood.

Order Placement. Once the expected return has been determined, agent i places the order in the book. The sign of the expected return defines the *direction* of the order: if $\hat{r}_i(t) > 0$ agent i places a buy order, otherwise a sell order.

Price Determination. Similarly to [1], the price determined by agent i is distributed around the bid price (if she decided to buy, $\hat{r}_i(t) > 0$) or the ask price (if she decided to sell, $\hat{r}_i(t) < 0$). Specifically,

$$p_i(t) = \begin{cases} p_{\text{bid}} - (\eta_i(t) - \varepsilon) & \text{if } \pi_i(t) > 0 \text{ (buy order)} \\ p_{\text{ask}} + (\eta_i(t) - \varepsilon) & \text{if } \pi_i(t) < 0 \text{ (sell order)} \end{cases} \quad (7)$$

where $\eta_i(t)$ is the realization of a log-normal distributed random variable. The shift factor ε determines the amount of market orders the agents place compared to limit orders. The choice of the log-normal distribution is linked to empirical observations on the distribution of orders in the book.

Demand Determination. Once the expected return and the price have been determined, similarly to [6], the agents will decide if buying or selling the asset, and which quantity they are willing to trade by maximizing the following Constant Absolute Risk Aversion utility function [13]

$$U_i(W_i, \alpha_i) = -e^{-\alpha_i W_i}, \quad (8)$$

where α_i is the risk aversion of agent i . Maximizing (8) under the assumption of an expected return $\hat{r}_i(t)$ leads to a theoretical demand of

$$\pi_i(t) = \hat{r}_i(t)/\alpha_i \hat{V}_i(t) p_i(t), \quad (9)$$

where $\hat{V}_i(t)$ is the estimated volatility at the time the decision is taken. Specifically, at each time step the volatility is estimated on a time windows of length τ_i (the *time horizon* of agent i) based on past prices data as follows:

$$\hat{V}_i(t) = \frac{1}{\tau_i} \sum_{i=0}^{\tau_i} (r(t-i) - \bar{r}_i(t))^2, \quad (10)$$

where $r(t-i) = \ln(p(t-i)/p(t-i-1))$ is the spot return of the market price at day $t-i$, while \bar{r} is the average spot return computed as

$$\bar{r}_i(t) = \frac{1}{\tau_i} \sum_{i=0}^{\tau_i} r(t-i) = \frac{1}{\tau_i} \sum_{i=0}^{\tau_i} \ln \left(\frac{p(t-i)}{p(t-i-1)} \right). \quad (11)$$

Then, agent i compares the theoretical assessment $\pi_i(t)$ of the demand with her resource availability, thus obtaining the actual demand $d_i(t)$ as

$$d_i(t) = \begin{cases} \lfloor \min(\pi_i(t), c_i/p_i(t)) \rfloor & \text{if } \pi_i(t) > 0 \text{ (buy order),} \\ \lfloor \min(-\pi_i(t), a_i) \rfloor & \text{if } \pi_i(t) < 0 \text{ (sell order),} \end{cases} \quad (12)$$

where $\lfloor \cdot \rfloor$ is the floor function. A buy/sell order of $d_i(t)$ units of the asset at price $p_i(t)$ is finally placed in the order book.

Order Cancellation. Finally, agent i can remove limit orders she made in the past. The reasons why an agent may be willing to remove an order stored in the book include the order age (an old order might have been made under very different return expectations and market conditions) or low likelihood of execution (the trader may decide to replace it with another with higher chances to be executed, or even with a market order). Once an order has been cancelled, the resources held by the order book as a margin for that order are given back to the agent, who can then use them to place subsequent orders.

Following the approach proposed in [19], we relate the order probability of cancellation to (i) its rank in the book with respect to the best offer of the same direction, (ii) its age, and (iii) the order book imbalance. Specifically, agent i automatically remove orders older than their time horizon τ_i . Then, for each orders that has not expired yet, she considers the order price p_o and computes the quantity

$$P^c = \begin{cases} \left(1 - \exp\left(\frac{p_o - p_{\text{bid}}}{p_o}\right)\right) \left(\frac{N_{\text{bid}}}{N_{\text{bid}} + N_{\text{ask}}} + \Delta_{\text{imb}}\right) & (\text{for a buy order}), \\ \left(1 - \exp\left(\frac{p_{\text{ask}} - p_o}{p_o}\right)\right) \left(\frac{N_{\text{ask}}}{N_{\text{bid}} + N_{\text{ask}}} + \Delta_{\text{imb}}\right) & (\text{for a sell order}). \end{cases} \quad (13)$$

The first term of the products measures the distance of the considered order with respect to the best order in the same direction, while the second term measures the imbalance between buy and sell orders, shifted by a term $\Delta_{\text{imb}} = 0.2$ to best fit empirical data [19]. She then cancels the order with probability $\min(1, P^c)$. Note that when an order is at the best market quote, the first term vanishes, so its probability of cancellation is 0.

Traded Asset. The output of a day t of trading is the vector $\pi^{\text{ex}}(t)$, where each agent stores the traded asset in day t (positive if the asset was bought, negative if sold). Note that $\pi_i^{\text{ex}}(t)$ may refer both to executed orders submitted by agent i at time t , or already present in the book.

3 Sociability and Market Behavior

Here, we perform extensive numerical simulations to establish the impact of sociability, quantified by the parameter ζ , on market behavior. Specifically, we aim at assessing whether there is a relationship between the presence of herding and the emergence of instabilities and inefficiency in the market. Therefore, we introduce the following metrics:

- the **herding intensity** H is defined as $H = \|\rho\|_2$, where ρ is the sample estimation of the correlation matrix of the time series of the traded assets π_i^{ex} , whose element ij is

$$\rho_{i,j} = \sigma_{ij}/\sigma_{ii}\sigma_{jj},$$

where

$$\sigma_{ij} = \frac{1}{T-1} \left(\sum_{t=1}^T (\pi_i^{\text{ex}}(t) - \bar{\pi}_i^{\text{ex}})(\pi_j^{\text{ex}}(t) - \bar{\pi}_j^{\text{ex}}) \right)^{1/2}, \quad i, j = 1, \dots, N,$$

Table 1. Initial state and simulations' parameters. d_i^{out} is the out-degree of agent i .

Initial value	Value	Parameter	Value
$c_i(0)$	$3000(1 + d_i^{\text{out}})$	α_i	$\mathcal{N}(5, 1)$
$a_i(0)$	$10(1 + d_i^{\text{out}})$	τ_i	$\mathcal{N}(50, 10)$
$r_i(0)$	0	γ_i	$\mathcal{N}(0.005, 0.001)$
		σ_i	$\mathcal{N}(0.75, 0.1)$

T is the length of the available time series, and $\bar{\pi}_i^{\text{ex}}$ is the sample mean of the trading sequence of the i -th agent;

- introducing the error $e(t) := p_f(t) - p(t)$ between the market price and the fundamental price, we define the **market efficiency** \bar{e} and **market instability** σ_e as

$$\bar{e} = \frac{1}{T} \sum_{t=0}^T e(t)^2, \quad \sigma_e = \frac{1}{T} \sum_{t=0}^T (e(t) - \bar{e})^2,$$

respectively.

To test the impact of sociability on the three quantities defined above, we perform a total of 900 simulations (30 for each of the 30 pairs (\mathcal{G}, p_f) of randomly generated graph topologies and fundamental prices) for each of the 11 selected values of ζ , equally spaced in the interval $[0, 1]$. Each simulation considers 300 financial agents trading for $T = 5000$ consecutive days. The graph topologies are scale-free like [5] to mimic the topological structure of real-world information spreading [16], while the fundamental price $p_f(t)$ is taken as a realization of a geometric Wiener process [6, 18]. The agents' initial states, as well as the random distributions from which the model parameters are extracted, are reported in Table 1. Finally, to obtain an order book history that is qualitatively similar to [1], in Eq. (7) we selected $\varepsilon = 1.3$ and $\eta_i(t)$ as a realization of a log-normal random variable, whose natural logarithm has 0 mean and variance equal to 0.5.

In Fig. 2, we report the box-plot of H , \bar{e} , and σ as a function of the sociability ζ . Furthermore, to assess the impact of sociability, for each metric and for each value of ζ , we performed a one-way ANOVA to test whether the metrics were significantly different from the case $\zeta = 0$, where sociality is absent. Both a graphical inspection and the statistical tests show the impact of sociality on the three metrics. We observe that, as sociality increases, herding emerges, in turn yielding an exponential increase of both market inefficiency and instability.

4 Control of Herding

Even though herding has been often blamed as one of the possible causes of financial crises [3], there are limited results in the literature on herding control and mitigation [24]. As illustrated in Sect. 3, our model suggests that herding is the transmission belt between sociality and market inefficiency and instability.

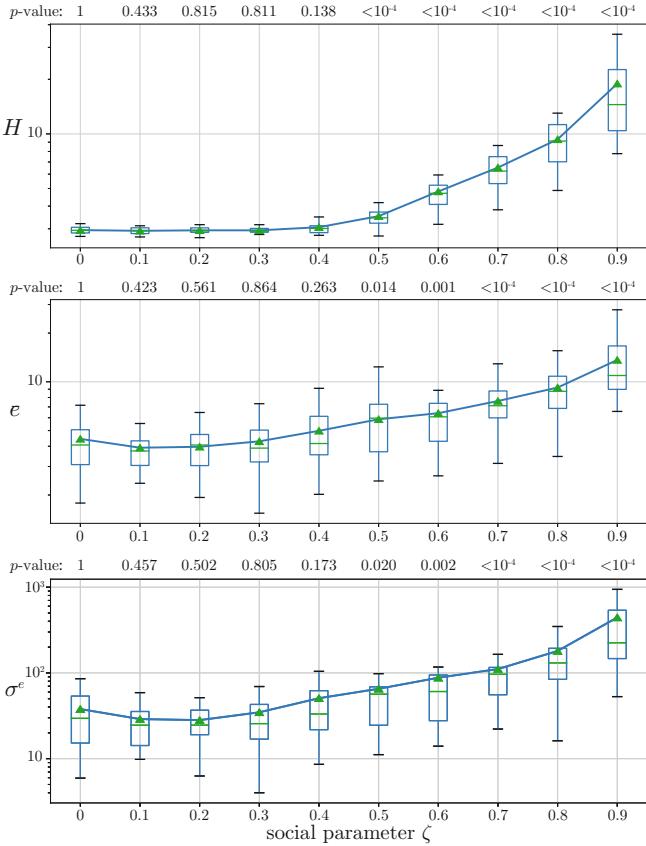


Fig. 2. Box plots of the realizations of the herding intensity H (top panel), the market efficiency \bar{e} (middle panel), and the market instability σ_e (bottom panel) obtained from 900 simulations of our model. On top of each panel, the p -values associated to the result of the ANOVA tests against the null-hypothesis ‘the effect of ζ is neglectable’ are reported.

Therefore, we propose a control strategy that acts on the network topology to mitigate the impact of sociality. Specifically, we leverage the vulnerability of scale-free networks to targeted attacks, which may disconnect the network by only removing around the 2% of the most connected nodes [9]. In formal terms, the control strategy we propose selects the subgroup of the most connected nodes, the ones with highest out degree, and isolates them from the network by reducing the weight of their outgoing edges.

To test the effectiveness of the proposed strategy in mitigating volatility and inefficiencies in our market model, we consider the same numerical setup as in Sect. 3. In particular, we focus on the case in which ζ is 0.7, that is, the lowest value such that has been observed to have a significant effect (with $p < 10^{-4}$) on all the three market measures, see Fig. 2. Then, we start to control (i.e. isolate)

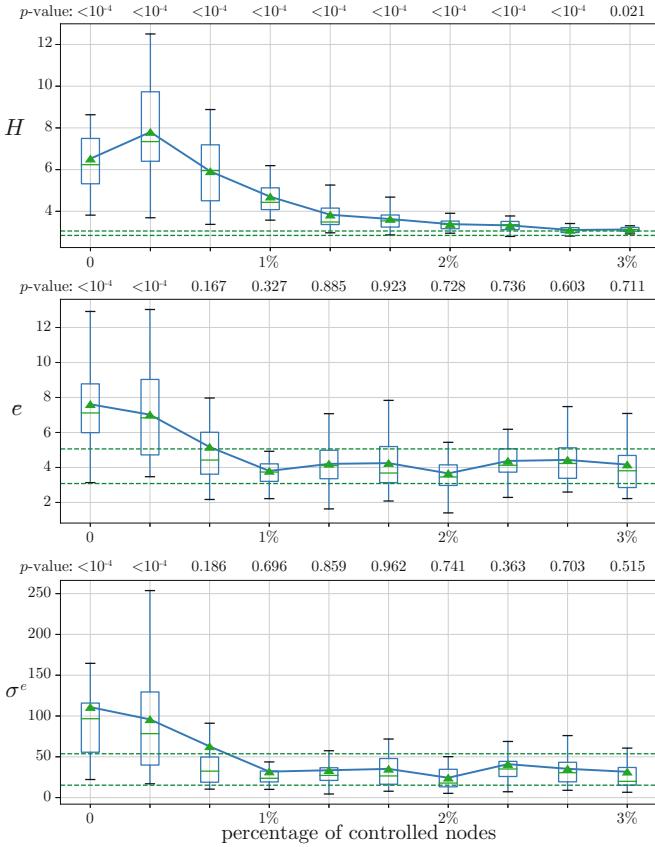


Fig. 3. Box plots of the realizations of the herding intensity H (top panel), the market efficiency \bar{e} (middle panel), and the market instability σ_e (bottom panel) from 900 simulations of our model for increasing percentages of the controlled nodes. On top of each panel, the p -values associated to the result of the ANOVA tests against the null-hypothesis ‘the presence of social interaction is neglectable’ are reported.

an increasing percentage of the most connected (i.e. influential) nodes, and test if there is a significant difference with the case of absence of sociality ($\zeta = 0$). From Fig. 3, we notice that, as the number of controlled nodes increases, the market regains efficiency, and volatility reduces. Interestingly, when only the first percentile of the most influential nodes is controlled, \bar{e} and σ_e become not statistically different from the case of absence of sociality, while H is still statistically different. This means that, although information still diffuses in the network, as the herding intensity is still clearly distinct from the case of absence of sociality, the detrimental impact of herding in terms of inefficiency and volatility is successfully mitigated.

Using the same simulation setting of the previous section, we report the herding, efficiency and instability measures obtained when $\zeta = 0.7$ (the first for

which the p -value of the ANOVA test was $<10^{-4}$ for all the three measures) for an increasing amount of controlled node in Fig. 3. On top of each panel we report the p -value of the ANOVA test against the data obtained with $\zeta = 0$. In each panel, we report with green dashed lines the limits of the box obtained from our simulations with $\zeta = 0$ (i.e. the first and the third quartile). The top panel shows that if we control up to the 3% of the nodes, the herding measure H still be distinguishable from the one obtained without sociality: this results says that the effect of the network, just controlling less than the 3% of the nodes is still present and detectable. Pinning more nodes ($>4\%$) brings the obtained box between the dashed-green bounds, and the p -value of the ANOVA test at more than 0.5. Looking at the other two measures we notice that the undesirable effects of herding, i.e. market inefficiency and instability, are negligible even when only controlling the 1% most connected nodes of the network.

5 Conclusions

We presented a novel agent-based model that, leveraging tools from complex networks theory, explicitly models the effect of social interactions and information diffusion on a centralized market with a double auction order book. A numerical exploration of the model parameter space showed that, when the agents take their trading decisions also considering the expectations of their neighbors, the market price drifts away from the fundamental price and volatility increases. These results are in line with the empirical observation that herding can drive the market towards instability.

Then, to mitigate the detrimental effect of herding, we introduced a control strategy that directly acts on the graph where information diffuses across the agents. Specifically, the strategy consists in reducing the outgoing weights of the most influential nodes in the network. We observed that, isolating only a small fraction (1%) of the nodes suffices to regain market efficiency and reduce volatility, while preserving a significant information flow among the agents.

We emphasize that the proposed control strategy could be potentially translated into market policies and regulations. Indeed, in financial markets, the most influential nodes could be identified as the major investors, or the main financial institutions, and reducing their outgoing edges could be implemented by simply forbidding the diffusion of their trading strategies. Clearly, in view of an implementation in real-world markets, more sophisticated policies should be adopted to attenuate herding while preserving market transparency.

References

1. Bartolozzi, M.: A multi agent model for the limit order book dynamics. *Eur. Phys. J. B* **78**(2), 265–273 (2010)
2. Berg, N., Gigerenzer, G.: As-if behavioral economics: neoclassical economics in disguise? *Hist. Econ. Ideas* **18**, 133–165 (2010)

3. Bikhchandani, S., Sharma, S.: Herd behavior in financial markets. *IMF Staff Pap.* **47**(3), 279–310 (2000)
4. Blasco, N., Corredor, P., Ferreruela, S.: Does herding affect volatility? Implications for the Spanish stock market. *Quant. Finan.* **12**(2), 311–327 (2012)
5. Bollobás, B., Borgs, C., Chayes, J., Riordan, O.: Directed scale-free graphs. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 132–139. Society for Industrial and Applied Mathematics (2003)
6. Chiarella, C., Iori, G., Perelló, J.: The impact of heterogeneous trading rules on the limit order book and order flows. *J. Econ. Dyn. Control* **33**(3), 525–537 (2009)
7. Colander, D.: The death of neoclassical economics. *J. Hist. Econ. Thought* **22**(2), 127–143 (2000)
8. Cont, R., Bouchaud, J.P.: Herd behavior and aggregate fluctuations in financial markets. *Macroecon. Dyn.* **4**(2), 170–196 (2000)
9. Crucitti, P., Latora, V., Marchiori, M., Rapisarda, A.: Efficiency of scale-free networks: error and attack tolerance. *Phys. A: Stat. Mech. Appl.* **320**, 622–642 (2003)
10. De Long, J.B., Shleifer, A., Summers, L.H., Waldmann, R.J.: Positive feedback investment strategies and destabilizing rational speculation. *J. Finan.* **45**(2), 379–395 (1990)
11. DeLellis, P., DiMeglio, A., Garofalo, F., Iudice, F.L.: The evolving cobweb of relations among partially rational investors. *PLoS ONE* **12**(2), e0171891 (2017)
12. Dequech, D.: Neoclassical, mainstream, orthodox, and heterodox economics. *J. Post Keynes. Econ.* **30**(2), 279–302 (2007)
13. Eeckhoudt, L., Gollier, C., Schlesinger, H.: The risk-averse (and prudent) newsboy. *Manag. Sci.* **41**(5), 786–794 (1995)
14. Fama, E.F.: Efficient capital markets: a review of theory and empirical work. *J. Financ.* **25**(2), 383–417 (1970). <http://www.jstor.org/stable/2325486>
15. Friedman, D.: The double auction market institution: a survey. *Double Auction Market Inst. Theor. Evid.* **14**, 3–25 (1993)
16. Ganesh, A., Massoulié, L., Towsley, D.: The effect of network topology on the spread of epidemics. In: *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp. 1455–1466. IEEE (2005)
17. Lei, V., Noussair, C.N., Plott, C.R.: Nonspeculative bubbles in experimental asset markets: lack of common knowledge of rationality vs. actual irrationality. *Econometrica* **69**(4), 831–859 (2001)
18. Lux, T., Marchesi, M.: Volatility clustering in financial markets: a microsimulation of interacting agents. *Int. J. Theor. Appl. Finan.* **3**(04), 675–702 (2000)
19. Mike, S., Farmer, J.D.: An empirical behavioral model of liquidity and volatility. *J. Econ. Dyn. Control* **32**(1), 200–234 (2008). Applications of statistical physics in economics and finance
20. Smith, V.L., Suchanek, G.L., Williams, A.W.: Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econom. J. Econom. Soc.* **56**, 1119–1151 (1988)
21. Tedeschi, G., Iori, G., Gallegati, M.: Herding effects in order driven markets: the rise and fall of gurus. *J. Econ. Behav. Organ.* **81**(1), 82–96 (2012)
22. Teeter, P., Sandberg, J.: Cracking the enigma of asset bubbles with narratives. *Strateg. Organ.* **15**(1), 91–99 (2017)
23. Weintraub, E.R.: *Neoclassical Economics. The Concise Encyclopedia of Economics* (2002)
24. Zhang, J.Q., Huang, Z.G., Wu, Z.X., Su, R., Lai, Y.C.: Controlling herding in minority game systems. *Sci. Rep.* **6**, 20925 (2016)



A Network Structure Analysis of Economic Crises

Maximilian Göbel^(✉) and Tanya Araújo

ISEG, University of Lisbon and UECE, Lisbon, Portugal
maximilian.goebel@phd.iseg.ulisboa.pt

Abstract. Do countries with similar macroeconomic dynamics during pre-crisis times experience a common subsequent crisis-, respectively non-crisis-, status? Based on the Euclidean distance, a community structure detection algorithm generates the network topology of four distinct pre-crisis periods between 1990 and 2008 comprising 27 countries. The desired outcome is a clear-cut separation of future crisis from non-crisis economies. The approach succeeds in uncovering prominent cluster formations, whereas period-specific heatmaps reveal the time-varying importance of the considered indicators. The heterogeneous cluster-formation does not allow to infer any dynamics, which would unambiguously hint at an upcoming crisis event.

Keywords: Clustering · Crisis prediction · Macroeconomic dynamics

1 Introduction

In their World Economic Outlook of October 2018, the IMF published a study on the macroeconomic developments of the preceding decade. In the year after the events of September 2008, ninety-one economies, making up sixty-six percent of the world's GDP, suffered from declining output [10]. Still ten years after the breakdown of Lehman Brothers, sixty percent of those economies, which had not experienced a banking-crisis back then, struggled with output trends, ranging below their pre-crisis levels [10]. Besides the negative economic consequences, the effects of such a turmoil may spread beyond the boundaries of financial markets or the real economy and even challenge a country's social and political status quo. IMF economists found a positive correlation between declining GDP, due to the events of the Great Recession, and income inequality within a country [10]. A study by Nunn *et al.* [22] even detected a significantly positive correlation between negative economic growth and the probability of political turnover in a country. Despite seminal theoretical [15, 16, 23, 24] and empirical works [2, 3, 13], which inspired subsequent studies, there is still no consensus about the key determinants of economic turmoil. Whereas Reinhart and Rogoff [27] conclude that “[w]hile each financial crisis no doubt is distinct, they also share striking similarities” ([27], p. 11), others emphasize the changing nature of crises over time [6, 29].

Hence, this study tries to shed further light on the existence of specific macroeconomic dynamics, which drive countries onto a crisis- or non-crisis-trajectory, and whether certain patterns tend to recur. Despite several already existing non-parametric studies [8, 20, 31] this paper applies a new approach to the notion of similar macroeconomic development over a certain period of time to determine a common subsequent economic status.

2 Methodology

A community structure detection algorithm, fed with the Euclidean distance between any two countries, measured on a number of macroeconomic indicators throughout eight quarters prior to a *reference date*, is intended to generate a cluster formation, telling future crisis economies apart from non-crisis countries. A measure of consistency, as applied in Araújo and Göbel [1], quantifies the quality of the partitioning, followed by an evaluation of corresponding heatmaps, which examines the existence of certain macroeconomic characteristics, which may make a country prone to suffer from a subsequent crisis, as well as their persistence over time.

2.1 Measuring Similarity

The Euclidean distance, which qualifies as a *metric* [19], serves as the measurement of similarity among countries. Following the literature [2, 3, 20, 31], the quarterly raw data, described in Sect. 4.1, enters the analysis as percentiles of each indicator's *entire* time-series distribution. These percentiles $p_{i,n,t}$ then form the input for the row-vector $\mathbf{v}_{n,t}$, where $p_{i,n,t}$ is the percentile of indicator i for country n at quarter t . $\mathbf{v}_{n,t}$ describes the $t \times I$ row-vector for country n , where I is the total number of indicators i , which the model is composed of. The Euclidean distance, as described in Gan *et al.* [9], between country n and country z at a particular quarter t is defined as:

$$d_t(\mathbf{v}_{n,t}, \mathbf{v}_{z,t}) = \left[\sum_{i=1}^I (p_{i,n,t} - p_{i,z,t})^2 \right]^{\frac{1}{2}} = \left[(\mathbf{v}_{n,t} - \mathbf{v}_{z,t}) (\mathbf{v}_{n,t} - \mathbf{v}_{z,t})^T \right]^{\frac{1}{2}}, \quad (1)$$

where $p_{i,n,t}$ and $p_{i,z,t}$ are the percentiles of the i^{th} variable, respectively indicator, of country n , respectively z , at quarter t .

So far, this formula only measures the distance between any two countries n and z at one specific point in time, i.e. the bilateral distance in one specific quarter t . Extending the similarity measure to several periods, $\Delta t = [t, t + l]$, transforms the $t \times I$ row-vector $\mathbf{v}_{n,t}$ into the matrix $\Delta t \times I$. Thus, Eq. (1) is augmented by a time dimension as follows:

$$d_{\Delta t}(\mathbf{v}_{n,\Delta t}, \mathbf{v}_{z,\Delta t}) = \left[\sum_{t=1}^l \sum_{i=1}^I (p_{i,n,t} - p_{i,z,t})^2 \right]^{\frac{1}{2}} = \left[(\mathbf{v}_{n,\Delta t} - \mathbf{v}_{z,\Delta t}) (\mathbf{v}_{n,\Delta t} - \mathbf{v}_{z,\Delta t})^T \right]^{\frac{1}{2}}. \quad (2)$$

The drawback of Eq. (2) is the necessity for the matrix, $\mathbf{v}_{n,\Delta t}$, to be of equal dimension in the cross-section of countries, i.e. a balanced dataset is crucial. With Δt comprising only eight observations and accounting for the existence of non-linear relationships, other well-established distance measures [19] are ruled out for this setting.

2.2 Generating Sparse Networks

Other than the analysis of social networks (e.g. [25]), in which vertices do not maintain interrelations with all of the other nodes, this study's setup generates a complete network. Such a network structure with N nodes exhibits a total of $\frac{N(N-1)}{2}$ weighted edges. As a complete network does not conform with this paper's theoretical considerations of only the shortest distances to reveal economic similarity among countries, a method to reduce the number of edges and generate a *sparse* network is inevitable.

Measures like the well-accepted *Minimal Spanning Tree* (MST) [19] or adaptations of it [32] generate a sparse and connected network, in which, however, not necessarily the strongest ties are represented only. According to this paper's theory, countries, displaying a large mutual distance, shall neither cluster in the same community nor form any visible mutual link in the resulting network structure. Furthermore, the MST generates a connected network with no disconnected components. However, allowing for disconnected nodes or even whole clusters - *islands* or *cliques* - can reveal important insights, with the theoretically desired outcome favoring crisis-clusters to be decoupled from non-crisis communities.

Piccardi *et al.* [26] simply eliminate the 90% weakest links of the complete network. While this procedure might sound arbitrary, it serves as a benchmark to filter the strongest ties within a network encompassing a relatively small number of nodes. With the number of countries increasing, the 10% strongest links may still create a far too dense network structure. Several robustness checks on the basis of 10%, respectively 20%, of the $\frac{N(N-1)}{2}$ total interlinkages and on the $1.5 \times N$ shortest linkages, saw higher *modularity* measures, Q^G ¹ emerging when keeping the number of edges close to the number of the network's nodes N . Judging by the evaluation of these robustness checks, the upcoming empirical analysis is computed on the 10% shortest distances of the $\frac{N(N-1)}{2}$ total edges. Important to bear in mind is that such a threshold may fall short of conveying all the necessary information for separating crisis from non-crisis countries, but reduces the probability of insignificant links perturbing the information content of the shortest distances and grouping *false friends* into the same community, i.e. assigning countries to the same cluster even though their direct interconnection ranges among the network's *largest* distances².

¹ See Sect. 2.3 for an explanation of the *modularity* algorithm and the measure Q^G .

² The size of the threshold for the *largest* distances equals the one for the *shortest* distances.

2.3 Modularity - A Community Structure Detection Algorithm

Among the most prominent clustering algorithms are Newman's [21] *modularity* and the so-called *k-means* approach, dating back to MacQueen [18]. The method of *modularity* quantifies the quality of a specific cluster formation within a network. The underlying principle is not to minimize the number of edges between clusters, but to generate a structure which exhibits fewer between-cluster edges than expected from a random network structure, when keeping the overall number of links constant [21]. Whereas a random network does not exhibit any community structure at all [11], a large *modularity* measure, $Q \in [0; 1]$, suggests an unambiguous cluster formation. The advantage over other algorithms, such as *k-means*, which requires the number of clusters k to be fixed and exogenously specified at initialization [8], is *modularity*'s capability to derive the optimal number of clusters endogenously from the underlying dataset [21].

A refinement of Newman's [21] specification, Q , is implemented in the software package *Gephi*, which will be referred to in the forthcoming analysis. This refinement, Q^G , by Blondel *et al.* [4] improves the efficiency of computation and reduces the number of communities in short time. Furthermore, the *resolution limit problem*, describing the original measures's [21] struggle to detect small clusters, is also mitigated by the modified measure Q^G [7, 11].

2.4 Quantitative Evaluation

Once the network structure is generated, a quantitative assessment evaluates the topology's compliance with the research question, i.e. the extent as to which upcoming crisis countries separate from their non-crisis peers. The hard-coded *Crisis-Cluster-Consistency* figure (*CCC*) measures the degree of *correctly classified* countries. We follow Araújo and Göbel [1] in defining a country to be *correctly classified* and in calculating the extent as to which the network's clustering is consistent with a homogeneous crisis-flagging partitioning.³ Hence, a country is *correctly classified*, if it is assigned to the same community as at least half as many of those countries, with which it shares the same flagging. A single and disconnected node is only then marked as *correctly classified*, if no other country in the network shows the same crisis-/non-crisis-flag. The number of *correctly classified* countries is then divided by the total number of nodes, represented in the network. This ensures $CCC \in [0; 1]$, with $CCC = 1$ representing the desired network structure. However, our *CCC* does only consider the exact matching of flags to qualify for a *correct classification*. It does not account for the possibility of pre- and post-crisis stages - either of the same or a different type of crisis - to actually reveal similar dynamics, which would justify a heterogeneous cluster-formation. Furthermore, the chance that macroeconomic dynamics, which result in a certain crisis-status, are different for advanced and developing economies, is neither incorporated in our *CCC* measure. These issues are, however, no shortcomings of the *CCC*, as such findings have not been proven

³ See Table 1 for the country- and period-specific flagging.

theoretically nor empirically, but could turn out as a potential outcome of the upcoming analysis.

Last but not least, heatmaps are intended to give a more profound idea of the characteristics of the emerging clusters. They serve as a tool for validating established theories about ongoing macroeconomic dynamics during the run-up phases to economic crises and allow to infer potentially recurring patterns.

3 Pre-crisis Periods and Crisis Events

Most published papers have focused on explaining either a single type of crisis or a combination of banking- and currency-crises, which became also known as a *twin-crisis* [12]. Although a uniform definition for any type of crisis does not exist, the literature reveals the frequent use of certain quantitative and qualitative conventions. The crisis events for this paper's 27 countries were identified by matching the databases of [17, 28] and [29] and by adapting the *Exchange Market Pressure Index* following [13]. To prevent the latter from being driven by large price variations, the literature applies differing approaches to distinguish between low-inflation and high-inflation periods [13, 20, 31]. This paper follows [20] and [31], by dividing the time-series into periods of low and high inflation and calculating the *Exchange Market Pressure Index* for each case separately. The merging of the crisis dating databases and mechanisms reveals crises to not necessarily occur in the same quarter, but in clusters over time. In order to preserve the length of the pre-crisis horizon, the analysis is based on the following procedure: at first, certain time intervals, in which crisis-occurrences agglomerate, are considered as periods worth analyzing. A specific quarter t , around which several crises tend to cluster, serves as the *reference quarter* of the interval. The literature assumes a specific pre-crisis period to usually comprise 18 [30] to 24 months [2, 13] prior to the crisis-date t . The remainder of the paper adapts the latter and assumes the pre-crisis period to last for eight quarters. Thus the pre-crisis period comprises the quarters $t - 8$ to $t - 1$. For a country, experiencing a crisis in $t - q$ or $t + q$, the pre-crisis period will be shifted to cover the quarters $t - q - 8$ to $t - q - 1$ and $t + q - 8$ to $t + q - 1$ accordingly.⁴ The interval's non-crisis countries enter the analysis with the pre-crisis period running from quarter $t - 8$ to the last quarter, $t - 1$, prior to the *reference date* t .

Table 1 lists the country-specific *reference dates* by period, as well as the corresponding crisis, respectively non-crisis flags. Period 1 is listed, but has to be excluded from the upcoming analysis due to a lack of data availability.

⁴ We tried to stick to $q \in [-4; 4]$, i.e. keep the intervals of crisis-occurrences rather tight in order to not veil the existence of time-specific dynamics. However, this interval can be changed arbitrarily.

4 Pre-crisis Macroeconomic Similarities and Crisis Occurrence

4.1 The Model

For the upcoming analysis, we adapt the indicators selected by Berg and Patillo [3]. Their model has repeatedly served as a benchmark for validating early-warning models based on various clustering techniques [20,31]. For making the data cross-sectionally comparable, we follow [3], respectively most of the existing literature on early-warning models, and transform the quarterly-observed variables into percentiles. As this paper's dataset differs from the referenced ones

Table 1. Assessment periods 1–5

	Reference dates					Crisis flags				
	Period 1	Period 2	Period 3	Period 4	Period 5	Period 2	Period 3	Period 4	Period 5	
Argentina	Q1 1983	Q4 1991	Q1 1995	Q4 1997	Q3 2008	t	B; c	b	0	
Australia	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	b*	b*	C	C	
Austria	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	0	0	0	B	
Belgium	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	0	0	0	B	
Brazil	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	t	B; c	b	C	
Canada	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	0	0	0	0	
Colombia	Q2 1984	Q4 1991	Q4 1994	Q2 1998	Q3 2008	c	C	T	0	
Finland	Q1 1983	Q3 1991	Q4 1994	Q4 1997	Q3 2008	B; c1	0	0	0	
France	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	0	B*	0	B	
Germany	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	0	0	0	B	
Greece	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	B*; c	b*	0	B	
India	Q1 1983	Q4 1991	Q4 1993	Q4 1997	Q3 2008	0	B	b*	C	
Indonesia	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	B*	b*	T	C	
Italy	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	b*; C	b*	0	B	
Japan	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	B	b	b	0	
Mexico	Q4 1982	Q4 1991	Q4 1994	Q4 1997	Q3 2008	B	T	b*	C	
Norway	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	B; c1	0	0	C	
Netherlands	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	0	0	0	B	
Portugal	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	0	0	0	B	
South Korea	Q1 1983	Q4 1991	Q4 1994	Q3 1997	Q3 2008	0	0	T	0	
Spain	Q1 1983	Q4 1992	Q4 1994	Q4 1997	Q3 2008	C	0	0	B	
Singapore	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2008	0	0	C	0	
Sweden	Q1 1983	Q3 1991	Q4 1994	Q4 1997	Q3 2008	B; c1	b	0	T	
Thailand	Q1 1983	Q4 1991	Q4 1994	Q3 1997	Q3 2008	0	0	T	0	
Turkey	Q4 1982	Q4 1991	Q1 1994	Q4 1997	Q3 2008	B*	C	0	C	
United Kingdom	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q3 2007	B*	B*	b*	B; c1	
United States	Q1 1983	Q4 1991	Q4 1994	Q4 1997	Q4 2007	b	0	0	B	

Note: B = banking-crisis at the end of the period; C = currency-crisis at the end of the period; T = banking- & currency-crisis at the end of the period; b = banking-crisis occurrence within four quarters prior to the reference date; c = currency-crisis occurrence within four quarters prior to the reference date; t = twin-crisis occurrence within four quarters prior to the reference date; b1 = banking-crisis occurrence within four quarters after the reference date; c1 = currency-crisis occurrence within four quarters after the reference date.

above, the computation relies on modified indicators. For the *Real Exchange Rate Overvaluation* relative to trend, a country's exchange rate vis-à-vis the U.S. dollar is detrended using *R*'s HP-filter-code *hpfilter*. For the United States, the raw-data series on the exchange rate is substituted by end-of-quarter values of the *Real Trade Weighted U.S. Dollar Index*, retrieved from the FRED database. Following the literature [2, 3, 31], which refer to Kaminsky *et al.* [13], an increase in the *Real Exchange Rate* is associated with a *real depreciation*. Thus, low percentiles hint at an *overvalued Real Exchange Rate*. The ratio of *Short-Term Debt to International Reserves* is proxied by the ratio of the Bank for International Settlements' *Outstanding International Debt coming due < 1 year* - retrieved from FRED - relative to *International Reserves*, proxied by the International Financial Statistics' *RAXG_XDR* time-series. Other than in Berg and Pattillo [3], the *Current Account Deficit relative to GDP* did not undergo any pre-processing. A comparison of feeding the model with the year-on-year growth-rates of *Exports* respectively *International Reserves* and their reported level values, favored the latter over the former.

4.2 The Network Structures

Figure 1 comprises both the network partition generated by *Gephi's modularity* algorithm on the right-hand side and the corresponding heatmap on the left-hand side. The nodes in the pictures on the lower right symbolize the different clusters, with the size being determined by each cluster's total degree, i.e. the number of their cross-cluster links. The numbers, displayed in each cell of the heatmaps, report the average percentile during each pre-crisis period for each indicator and for each cluster.⁵

Reviewing the cluster composition in each period gives a first idea about the general result and already allows to draw a preliminary conclusion: the overall topology suffers from a high degree of within-cluster heterogeneity regarding the accumulation of crisis and non-crisis countries. The low *CCC* values, displayed in Fig. 1a–d, further confirm this impression. On the contrary, the *modularity* values Q^G range among those found in several other applications [21].

Despite the poor distinction of crisis and non-crisis clusters, some prominent communities did emerge even if the interpretation is highly influenced by an ex-post point of view: the build-up of the Asian-crises formation in Period 3 (Fig. 1b) and Period 4 (Fig. 1c) or the distinction between the South-European states and the northern Euro Area economies in Period 5 (Fig. 1d) as well as the attachment of the United States to the later on troubled southern European economies. An interesting fact regarding the latter finding is the period-shift as depicted in Column 6 of Table 1. The pre-crisis period of the United States runs three quarters ahead of their cluster peers, which begs the question whether

⁵ EX = *Exports*; Reserves = *International Reserves*; R_ExchRate_DevTrend = *Real Exchange Rate Overvaluation relative to Trend*; N_STDebt/Reserves = *Nominal Short-Term Debt to International Reserves*; CA/GDP = *Current Account Deficit relative to GDP*.

a real-time monitoring might have indicated the southern European countries' trajectory? Nevertheless, the reasons for a rather different subsequent economic development in the United States compared to its community peers are left to speculation: is it the U.S.-Dollar's outstanding role as a reserve currency and the United State's perception as a safe haven? Were varying capabilities of and interventions by the monetary authority in question the reasons for differing future trajectories despite apparently similar fundamentals prior to the Great Recession? Three issues are, however, important to be aware of: first, we only looked at five macroeconomic indicators which were supposed to proxy the interplay of dozens of variables determining a country's economic development. Furthermore, the rather nice build-up of the Asian-crisis structures in Periods 3 and 4 shall not be surprising as this very event was the trigger for the initial calibration of the model. Last but not least, Period 3 reveals a shortcoming of the *modularity* algorithm regarding this paper's setting. Cluster 2 suffers from the existence of *false friends*. Greece and Belgium form a mutual link, ranging among the 10% *largest* distances, and may hence not be assigned to the same community.

Heatmaps are now intended to provide a better understanding of the cluster-specific dynamics and to compare the empirical results with the prevailing theoretical literature about the dynamics preceding a crisis event. The only homogeneous crisis community in Period 2 is depicted in Cluster 2. This community is characterized by CA/GDP , which rates on average at the third quartile and a rather low ratio of $N_STDebt/Reserves$. The combination of depleted *Reserves* but a still *overvalued Real Exchange Rate* is detrimental to theoretical reasoning about the aftermath of currency-crises [5]. A closer look at the crisis-dating in Reinhart and Rogoff [28] shows the pictured pre-crisis period - the early 1990s - to mark the end of a prolonged period of successive currency-crises combined with hyperinflationary tendencies in both countries [14]. Thus, Brazil and Argentina may not serve as a stereotype example of countries, which all of a sudden had to cope with an attack on their currencies. Furthermore, Cluster 2 seems to show rather unique dynamics compared to other communities, which conforms with its visualization as an *island*. The communities to show still the closest similarities, despite some visible differences, are Clusters 6 and 8. Rather low levels of CA/GDP ⁶ and an overvalued *Real Exchange Rate* do fit the theory of a possible target for a speculative attack [5], whereas *Reserves* rather indicate the Central Banks of Spain and Italy to have had the necessary resources to fight a threatening devaluation. Nevertheless, the Spanish Peseta and the Italian Lira got indeed under pressure around the *reference date* [5], leaving a puzzle to the prevailing understanding of currency-crisis and triggering the formation of the second-generation of currency-crisis models [5, 23, 24]. In contrast to Cluster 2, Cluster 5 makes a typical example for the aftermath of a currency-crisis: depleted *Reserves*, an elevated level of CA/GDP and a *depreciated* currency. Besides the $N_STDebt/Reserves$, this result pictures the stereotype differences between the

⁶ The raw-data reveals indeed a *Current Account Deficit*.

aftermath (Cluster 5) and the run-up period (Clusters 6 and 7) to a currency-crisis event [5]. Low values of *Reserves*, a low level of CA/GDP ⁷ combined with an overvalued *Real Exchange Rate* vis-à-vis its long-run trend, would suggest India to be on a trajectory into a currency-crisis. However, neither the *Exchange Market Pressure Index* nor [17] do report any event around the *reference date*. Only [28] state a currency-crisis for India in 1991 and a twin-crisis to start in 1993. This finding underscores the importance of accurate crisis-dating procedures.

Period 3 pictures the aftermath of the breakdown of the *European Exchange Rate Mechanism I*, an event, which was not to be explained by macroeconomic abnormalities [5]. Cluster 2 does indeed show characteristics of an aftermath of speculative attacks on a country's currency, even if the high levels of *Reserves* do not fit into such a scenario, matching the findings of several studies [5, 23, 24]. The figures in Clusters 6 and 7 are again stereotype for the aftermath of a currency-crisis: depleted *Reserves*, a recovered CA/GDP and a strongly depreciated currency. These results do indeed conform with the events regarding the Italian Lira as well as the currency-crisis in India, when following [28]. Eye-catching numbers are displayed for Cluster 9, which represents the run-up period to Mexico's twin-crisis. A large *Current Account Deficit* combined with an overvalued *Real Exchange Rate*, do match the theoretical considerations about an upcoming currency-crisis. Even if a direct inter-period comparison is not straightforward in this setting, the numbers resemble the pattern seen in Cluster 7 in Period 2, in which India was said to be in the run-up to a currency-crisis.

Cluster 0 of Period 4 is exclusively composed of non-crisis countries. High *Reserves*, a reasonable CA/GDP , and a median value for $N_STDebt/Reserves$ conform with the prevailing theory and do not hint at any upcoming speculative attacks on the currencies, despite tendencies of the *Real Exchange Rate* to be overvalued vis-à-vis its long-run trend. Cluster 3 rather shows the contrary. It comprises almost all Asian-crisis countries and the heatmap displays close similarities to Cluster 9 in Period 3 and Cluster 7 in Period 2. Hence comparable dynamics between the run-up phase to Mexico's currency-crisis a few years earlier, respectively India's pre-crisis quarters, also seemed to be at work prior to the Asian crises of 1997/98.

So far the findings may induce to assume that a rather overvalued *Real Exchange Rate* combined with either high levels of *Reserves* and/or a high percentile of CA/GDP , are sufficient for *not* being considered as a target for a speculative attack. However, despite high levels of *Reserves* and CA/GDP , the currencies of almost all economies in Clusters 3 and 4 of Period 5 got under pressure around the *reference date*. CA/GDP seems to separate the South-European countries from their northern peers.

A robustness check replaced *Exports* with the *Rate of Unemployment*, as the latter was regarded as the root cause for driving the expectations of market participants, culminating in the attacks on the *European Exchange Rate Mechanism I* at the beginning of the 1990s [5, 23]. Given that setting, the network structure

⁷ Cross-checking the raw-data reveals indeed a *Current Account Deficit*.

remains heterogeneous. Nevertheless, a high *Rate of Unemployment* seems to be characteristic for separating largely crisis-composed clusters from non-crisis communities in the early 1990s. However, this discriminatory power diminishes over time.

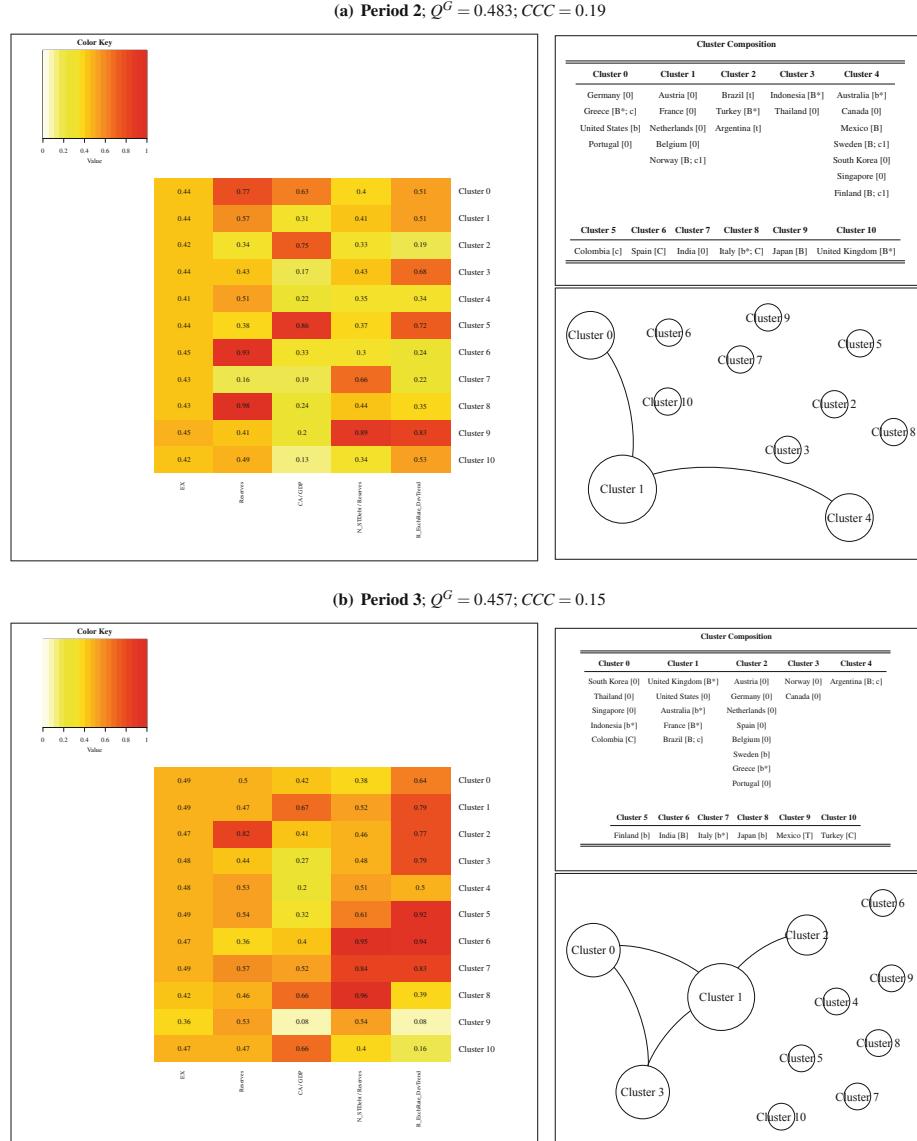
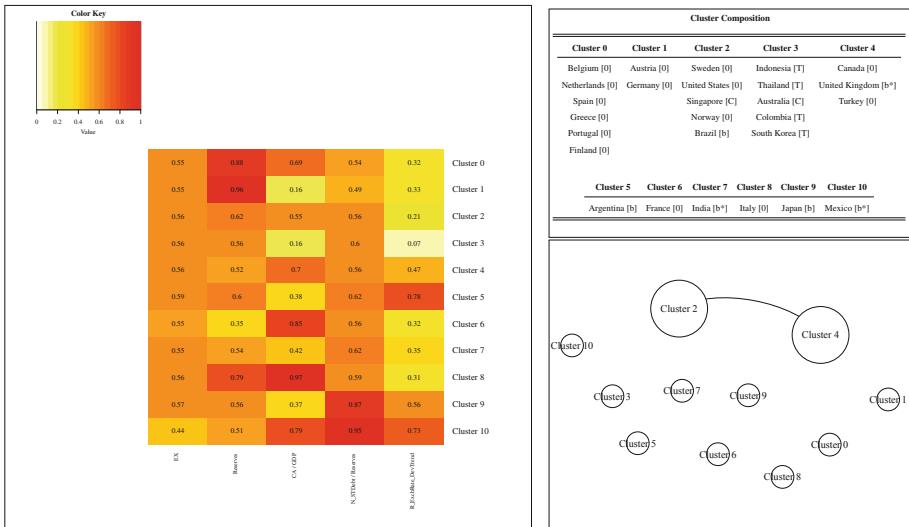
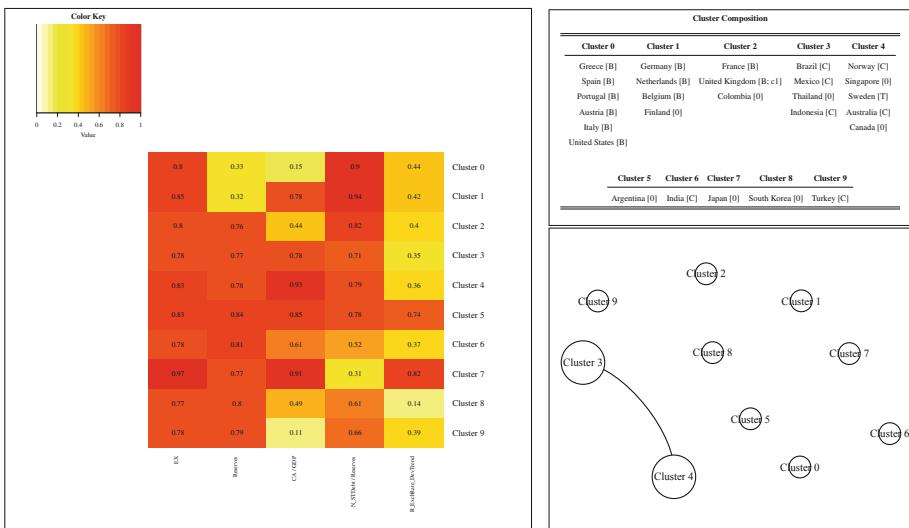


Fig. 1. Heatmap and cluster-composition

(c) Period 4; $Q^G = 0.689$; $CCC = 0.15$ (d) Period 5; $Q^G = 0.602$; $CCC = 0.22$ **Fig. 1. (continued)**

5 Conclusion

The present study tried to shed further light on whether certain macroeconomic dynamics during pre-crisis times do hint at a trajectory leading a country into a crisis or non-crisis status. After proxying cross-country similarities by the Euclidean distance, *Gephi's modularity* algorithm generated a network topology for four distinct eight-quarter long periods comprising 27 countries. The quantitative *CCC* measure confirmed the visible failure to identify pure crisis, respectively non-crisis, communities. Heatmaps were intended to give a better understanding of the cluster-specific dynamics and to compare the results with prevailing theoretical studies about the build-up of economic crises. Due to heterogeneous cluster formations, unambiguous or even timely invariant patterns of macroeconomic dynamics could not be detected. However, some observations are noteworthy: low percentiles of *CA/GDP* and an overvalued *Real Exchange Rate* combined with no more than median-level *Reserves*, seemed to be a strong indication for *emerging market economies* to be on a trajectory into a currency-crisis during the 1990s. This pattern vanished during the run-up to the Great Recession. The latter seemed to display its own dynamics. While the amount of *Reserves* seemed to be the decisive factor between the currency-crisis-clusters⁸ and the banking-crisis clusters⁹, it was the *CA/GDP* which separated the Mid-Northern European economies from their southern peers and the U.S. Even if this paper did not identify recurring macroeconomic dynamics during each eight-quarter long period prior to four distinct crisis events, the underlying methodology may serve as an impulse for further research and as a starting point for investigating the reasons for varying economic development despite allegedly similar initial circumstances. Upcoming research is targeted at dynamically searching for early-warning indicators, mitigating the sensitivity to the crisis-dating procedure and adapting the clustering algorithm to the study's requirements.

References

1. Araújo, T., Göbel, M.: Reframing the S&P 500 network of stocks along the 21st century. *Phys. A: Stat. Mech. Appl.* **526**, 121062 (2019). <https://doi.org/10.1016/j.physa.2019.121062>
2. Berg, A., Pattillo, C.: Predicting currency crises: the indicators approach and an alternative. *J. Int. Money Financ.* **18**, 561–586 (1999)
3. Berg, A., Pattillo, C.: What caused the asian crises: an early warning system approach. *Economic Notes by Banca Monte dei Paschi di Siena SpA* **28**(3), 285–334 (1999b)
4. Blondel, V.D., Guillaume, J.-L., Lambotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **10**, 1–12 (2008)
5. Eichengreen, B., Rose, A.K., Wyplosz, C.: Exchange market mayhem: the antecedents and aftermath of speculative attacks. University of California, Berkeley (1995). <http://faculty.haas.berkeley.edu/arose/erw3ep.pdf>. Accessed 16 July 2019

⁸ See Fig. 1d Clusters 3 and 4.

⁹ See Fig. 1d Clusters 0 and 1.

6. Fioramanti, M.: Predicting sovereign debt crises using artificial neural networks: a comparative approach. *J. Financ. Stab.* **4**, 149–164 (2008)
7. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* **104**(1), 36–41 (2007)
8. Fuertes, A.-M., Kalotychou, E.: Optimal design of early warning systems for sovereign debt crises. *Int. J. Forecast.* **23**, 85–100 (2007)
9. Gan, G., Ma, C., Wu, J.: Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM Series on Statistics and Applied Probability, p. 71. SIAM, Philadelphia (2007)
10. International Monetary Fund: World Economic Outlook: Challenges to Steady Growth, Washington DC, pp. 71–100 (2018)
11. Isogai, T.: Clustering of Japanese stock returns: statistical analysis of the correlation structure of fat-tailed returns. Dissertation, Japan Advanced Institute of Science and Technology (2015)
12. Kaminsky, G.L., Reinhart, C.M.: The twin crises: the causes of banking and balance-of-payments problems. *International Finance Discuss Paper* 544 (1996)
13. Kaminsky, G.L., Lizondo, S., Reinhart, C.M.: Leading indicators of currency crises. *IMF Staff Pap.* **45**(1), 1–48 (1998)
14. Kaminsky, G.L., Mati, A., Choueiri, N.: Thirty years of currency crises in Argentina: external shocks or domestic fragility? *NBER Working Paper* 15478 (2009)
15. Krugman, P.: A model of balance of payments crises. *J. Money Credit Bank* **11**, 311–325 (1979)
16. Krugman, P.: Balance sheets, the transfer problem, and financial crises. *Int. Tax. Public Financ.* **6**, 459–472 (1999)
17. Laeven, L., Valencia, F.: Systemic banking crises revisited. *IMF Working Paper WP/18/206* (2018)
18. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297 (1967)
19. Mantegna, R.N.: Hierarchical structure in financial markets. *Eur. Phys J. B.* **11**, 193–197 (1999)
20. Marghescu, D.R., Sarlin, P., Liu, S.: Early-warning analysis for currency crises in emerging markets: a revisit with fuzzy clustering. *Intell. Syst. Acc. Financ. Manag.* **17**, 143–165 (2010)
21. Newman, M.E.J.: Modularity and community structure in networks. *Proc. Natl. Acad. Sci.* **103**(23), 8577–8582 (2006)
22. Nunn, N., Quian, N., Wen, J.: Distrust and political turnover. *NBER Working Paper Series* 24187 (2018)
23. Obstfeld, M.: The logic of currency crises. *NBER Working Paper Series* 4640 (1994)
24. Obstfeld, M.: Models of currency crises with self-fulfilling features. *Eur. Econ. Rev.* **40**, 1037–1047 (1996)
25. Padgett, J.F., Ansell, C.K.: Robust action and the rise of the medici, 1400–1434. *Am. J. Sociol.* **98**, 1259–1319 (1993)
26. Piccardi, C., Calatroni, L., Bertoni, F.: Clustering financial time series by network community analysis. *Int. J. Mod. Phys. C* **22**(1), 35–50 (2011)
27. Reinhart, C.M., Rogoff, K.S.: Is the 2007 U.S. sub-prime financial crisis so different? An international historical comparison. *NBER Working Paper* 14587 (2008)
28. Reinhart, C.M., Rogoff, K.S.: This Time is Different: Eight Centuries of Financial Folly. Princeton University Press, Princeton (2009)

29. Ristolainen, K.: Predicting banking crises with artificial neural networks: the role of nonlinearity and heterogeneity. *Scand. J. Econ.* **120**(1), 31–62 (2018)
30. Sarlin, P.: On biologically inspired predictions of the global financial crisis. *Neural Comput. Appl.* **24**, 663–673 (2014)
31. Sarlin, P., Marghescu, D.R.: Visual predictions of currency crises using self-organizing maps. *Intell. Syst. Acc. Financ. Manag.* **18**(1), 15–38 (2011)
32. Spelta, A., Araújo, T.: The topology of cross-border exposures: beyond the minimal spanning tree approach. *Phys. A: Stat. Mech. Appl.* **391**(22), 5572–5583 (2012)



A Multiplier Effect Model for Price Stabilization Networks

Jun Kiniwa¹(✉) and Hiroaki Sandoh²

¹ School of Social Information Science, University of Hyogo,
8-2-1 Gakuen-nishi, Nishi, Kobe 651-2197, Japan

kiniwa@sis.u-hyogo.ac.jp

² School of Policy Studies, Kwansei Gakuin University,
2-1 Gakuen, Sanda, Hyogo, Japan
sandoh@kwansei.ac.jp

Abstract. We consider a multiagent network model consisting of nodes and edges as cities and their links to neighbors, respectively. Each network node has an agent and priced goods, where the agent can buy or sell goods in the neighborhood. Though every node may not have an equal price, we can show the prices will reach an equilibrium by iterating buy and sell operations. We introduce a protocol in which each buying agent makes a bid to the lowest priced goods in the neighborhood; and each selling agent selects the highest bid (if any). We are interested in how some conventional theory, for instance, the multiplier effect is reconstructed in our model. Thus, we develop an multiplier effect model which enables us to apply a fiscal policy to our price stabilization networks. Our model also includes reducing inventory stocks if there is an excess supply. Finally, we run simulation experiments and investigate the influence of network features on the reduction of inventory stocks.

Keywords: Self-stabilization · Multiagent network · Multiplier effect · Reduction of inventory stocks

1 Introduction

Background. Conventionally, the topic of price determination has been discussed from microeconomics approach [17]. In the presence of supply and demand curves, if the price is higher (resp. lower) than an equilibrium, there is excess supply (resp. excess demand) and thus the price moves to the equilibrium. At the equilibrium price, the quantity of goods sought by consumers is equal to the quantity of goods supplied by producers. Neither consumers nor producers have incentive to change the price/quantity of goods at the equilibrium.

In contrast, we considered a multiagent network model [8–10], in which each agent repeatedly makes auctions and the price of goods is eventually determined. Our network model consists of nodes and edges as cities and their links to neighbors, respectively. Each node contains exactly one agent which represents people living in the city. Agents who want to buy goods make bids to the

lowest-priced neighboring node, if any. Then, agents who want to sell the goods accept the highest bid. We have shown the reason of price determination by using the idea of self-stabilization in distributed systems [4]. From any initial state, self-stabilizing algorithms eventually lead to a legitimate state without any aid of external actions. Such a self-stabilization resembles the price determination, where the price reaches an equilibrium without external operations.

Motivation. Our first work was motivated by an intuition that simulating transactions between agents may stabilize the price instead of the supply-demand theory. We developed a trading model using auctions in which prices converge to a unique one [8,9]. We, however, were not able to explain why such a unique price is determined. After that, we assumed a relation $p_i = m_i/q_i$ among the price p_i , goods q_i and money m_i at each node i , and each agent exchanges money and goods. Then, it enables us to estimate an equilibrium price $P_e = M/T$, where $M = \sum_i m_i$ and $T = \sum_i q_i$ [10]. Further, we developed a method of expected optimal bidding [11] and then extended our model to an asynchronous model [12]. In the model, we can estimate the equilibrium price at $P_e = MV/T$ by using the velocity of money V . We also investigated the influence of money injection, which represents a monetary policy.

Problem. As stated, we were able to examine the influence of monetary policy of a central bank by which inflation/deflation may be caused. Our previous models were suitable for exploring how the increase/decrease of money affects price stabilization and the moves of money to each node. We, however, were not able to examine the influence of fiscal policy of a government by which GDP may be boosted. Our previous models were not suitable for exploring how the investment for producing goods affects GDP and the velocity of money.

Solution. Our solution to the problem above is to increase/decrease the quantity of goods at the selling node. Initially, suppose that neighboring agents A and B have distinct priced goods, where B 's goods are cheaper than A 's. If agent A wants to buy more goods beyond B 's inventory stocks with excess of money, A asks B to increase production or build a factory to produce goods. Then, A can buy all the goods newly produced and the prices reach an equilibrium. On the contrary, if agent A wants to buy no more goods at B with short of money, A asks B to reduce his inventory stocks or stop his planning production. Then, the prices reach an equilibrium.

In other words, our new idea is to increase/decrease the quantity of goods in strict accordance with the excess/lack of money. Then, we can discuss the multiplier effect of buying goods based on this idea.

Related Work. The classical theory of price determination is introduced in microeconomics [17], and that of the multiplier effect is categorized as macroeconomics [16]. Recent work on the multiplier effect [18,20] is also done using the theory. In contrast to the conventional work, we set up an agent-based model. Our agent model is simpler than the previous macroeconomic modeling [1,19]. There exist a large body of literature on social economic networks [2] containing

a network formation game [6] and a buyer-seller network [7, 13]. The network formation game considers the choice between agents [5], while the buyer-seller network considers the competition and exchange in bipartite networks [13]. Unlike their interest in maximizing economic surplus, our work focuses on price stabilization. Auction theory has been comprehensively studied in [14]. Our protocol in Sect. 2.2 may be considered as a consensus algorithm. The consensus algorithm is described in [15], and its self-stabilizing version is described in [3]. We, however, cannot regard their work as economics.

Contributions. In this paper, we propose a multiplier effect model, which represents a fiscal policy, for price stabilization networks. We first describe the idea of representing investments and the reduction of inventory stocks in our network model. Then, we consider our protocol and prove its correctness. We also show some additional properties. The advantage is to boost economy through the multiplier effect, while the disadvantage is to reduce inventory stocks. To compare the scale of reduction, we run simulation experiments for three networks with respect to two parameters. Briefly, the reduction of inventory stocks does not occur so much in dense networks and if the quantity of newly produced goods is small.

We organize the rest of this paper as follows. Section 2 states our model and protocols. Section 3 proves the correctness of our protocol and some additional results. Section 4 shows some results of simulation experiments for several networks. Finally, Sect. 5 concludes the paper.

2 Model

Here we describe our model consisting of a network in Sect. 2.1 and a protocol design in Sect. 2.2.

2.1 Network

Our system can be represented by a connected network $G = (V, E)$, consisting of a set of nodes V and edges E , where each node represents a city and each link between nodes represents a neighborhood. Let N_i be a set of neighboring nodes of $i \in V$. We assume that each node $i \in V$ has a single good with a distinct price. Let p_i be the price of the goods at node i . Each node $i \in V$ has exactly one representative agent a_i who always stays at i and can buy goods in the neighborhood N_i . Each agent a_i has disposable money m_i , savings s_i and the quantity q_i of goods, where he uses money in m_i and does not use money in s_i . Every income I_i of a_i is separated into $\beta \cdot I_i$, added to m_i , and $(1 - \beta) \cdot I_i$, added to s_i . Here, the ratio β is called a *marginal propensity to consume*. The price p_i is determined by the relation between the quantity of goods and the buying power, called a *supply-demand* balance. So we simply assume two properties at each node. First, the price is proportional to the amount of money and savings

for constant goods. Second, the price is inversely proportional to the amount of goods for constant money and savings. That is, we have a relation

$$p_i = \frac{m_i + s_i}{q_i}, \quad (\text{i})$$

where we assume $p_i \neq 0$ and $q_i \neq 0$. Summing up the variables for every node, we obtain the Fisher's quantity equation [17]. Thus, we can verify its correctness. We also assume a *synchronous model*, that is, every agent periodically exchanges messages and knows the states of neighboring agents at each step, called a *round*.

The *buy operation* is executed as follows. Each agent a_i assigns a *value* v_i^j to the goods of any neighboring node $j \in N_i$, where the value means the maximum amount an agent is willing to pay. Agent a_i compares its own goods price p_i with the neighboring price p_j . If the cheapest price in N_i is p_j ($< p_i$), agent a_i wants to buy it and makes a bid b_i^j to node j . We consider $v_i^j = p_i$ for any $j \in N_i$ because he can buy it at price p_i at his node [14].

The *sell operation* is executed as follows. After receiving bids, agent a_j *contracts* with $a_i \in N_j$, who made the highest bid b_i^j . Then, agent a_j passes a_i several units of goods, and conversely agent a_i passes a_j money b_i^j per a unit. We call the set of exchanges in a round a *block*. Further, we call a series of blocks (or a block), a *transaction*, carried out until the price p_j becomes equal to p_i (by relation (i)). We do not take the carrying cost of goods into consideration but focus on the change of prices.

Each node $i \in V$ has a state Σ_i represented by a tuple—goods, money and savings (q_i, m_i, s_i) . We call the state of all nodes a *configuration*. We describe the set of all configurations as $\Gamma = \Sigma_1 \times \Sigma_2 \times \dots \times \Sigma_{|V|}$. An *atomic step* consists of reading the states of neighboring agents, a buy/sell operation, and updating its own state. Then, the atomic step changes a configuration $\mathbf{c}_t \in \Gamma$ to $\mathbf{c}_{t+1} \in \Gamma$ in the t -th round. An *execution* \mathbf{E} is a sequence of configurations $\mathbf{E} = \mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_t, \mathbf{c}_{t+1}, \dots$ such that $\mathbf{c}_t \in \Gamma$ changes to $\mathbf{c}_{t+1} \in \Gamma$.

2.2 Protocol Design

In this section, we first describe a basic protocol, denoted by **Naive** [9], in which every agent can buy or sell goods from/to other agents until prices reach an equilibrium. Next, we consider a new protocol model, called a multiplier-effect protocol, denoted by **MultiplierEffect**, in which every agent is willing to expend the proportion β of his income and saves the proportion $1 - \beta$ of it. Notice that we add a priority rule (rule 5) to avoid confusion.

Naive

1. Each agent a_i makes a bid b_i^j to neighboring agent a_j which has the lowest-priced goods in N_i , where $p_j < b_i^j < p_i$. For example, $b_i^j = \frac{p_i + p_j}{c}$ for a constant c ($1 + p_j/p_i < c < 1 + p_i/p_j$).
2. Agent a_j contracts with the neighboring a_h who has made the highest bid $\max_{h \in N_j} b_h^j$.

3. Then, the goods moves from q_j to q_h and the money moves from m_h to m_j at h 's bidding price b_h^j as long as $p_h > p_j$. The new prices p_h and p_j after the exchange are updated by relation (i).
4. If several agents make bids to node j with the same highest price, agent a_j selects one of them at random.
5. (*priority rule*:) If concurrent buy (b_j^k to $k \in N_j$) and sell (b_h^j from $h \in N_j$) concentrate at agent a_j , the sell is given priority over the buy.

If the rule 3 above is replaced by the following rule 3', we call it a multiplier-effect protocol (**MultiplierEffect**).

- 3'. – If agent a_h has enough money beyond the goods of a_j (by condition **Increment**), agent a_h buy all q_j and newly produced goods at node j until the prices p_j and p_h become equal.
– On the contrary, if agent a_h does not have enough money to buy all the goods of a_j (by condition **Decrement**), agent a_j reduces his inventory stocks until the prices p_j and p_h become equal.
– In both cases above, the proportion β of money moves from m_h to m_j (the proportion $1 - \beta$ of money moves from m_h to s_j) at a_h 's bidding price b_h^j as long as $p_h > p_j$.

The conditions by which agent a_h asks agent a_j to increase/decrease goods are given as follows, denoted by $b = b_h^j$ for simplicity. Without loss of generality, we can assume that the savings, bidding price, etc., are positive values.

Increment: If $\frac{s_h}{q_h + m_h/b} \leq \frac{m_j + s_j + m_h}{q_j - m_h/b}$ holds, agent a_h buys $I = \frac{1}{s_h}(q_h + m_h/b)(m_j + s_j + m_h) - (q_j - m_h/b)$ units for new goods in addition to buying p_j .

Decrement: If $\frac{s_h}{q_h + m_h/b} > \frac{m_j + s_j + m_h}{q_j - m_h/b}$ holds, agent a_j reduces $R = (q_j - m_h/b) - \frac{1}{s_h}(q_h + m_h/b)(m_j + s_j + m_h)$ units from q_j , and a_h buys the rest.

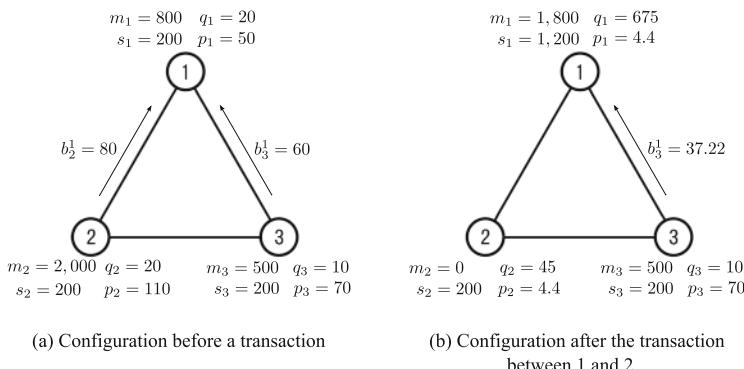


Fig. 1. An illustration of protocol **MultiplierEffect**

Example 1. Figure 1 shows an example of our network system consisting of 3 nodes $V = \{1, 2, 3\}$. Suppose that a transaction consists of a block and that each bid from a_i to a_j is $b_i^j = \frac{p_i + p_j}{2}$. At first, the prices of goods are $(p_1, p_2, p_3) = (50, 110, 70)$ as shown in Fig. 1(a). Each agent a_i wants to buy the lowest-priced goods at node $j \in N_i$ if its price is lower than p_i , that is, $p_i > \min_{j \in N_i} p_j$. Thus, both a_2 and a_3 make bids to node 1. Since agent a_2 beats a_3 , agent a_2 makes a contract with agent a_1 . Let x be the number of a_2 's buying units of goods. Since agent a_2 has enough money beyond q_1 , he asks a_1 to produce new goods so that he can pay all his disposable money. Then, a_1 produces more $I = 680$ units of goods and a_2 buys them all. The prices of nodes 1 and 2 reach an equilibrium $4.44\cdots$ by $p_2 = \frac{(m_2 - b_2^1)x + s_2}{q_2 + x} = \frac{m_1 + s_1 + b_2^1x}{q_1 + I - x} = p_1$, where $m_2 - b_2^1x = 0$. Agent a_2 's money 2,000 is divided into m_1 and s_1 1,000 each according as the marginal propensity to consume $\beta = 0.5$.

In Fig. 1(b), agent a_3 randomly selects agent a_1 and tries to make a bid $b_3^1 = 37.22$. However, $p_3 = \frac{s_3}{q_3 + m_3/b_3^1} > \frac{m_1 + s_1 + m_3}{q_1 - m_3/b_3^1} = p_1$ means the lack of agent a_3 's money. Then, agent a_1 has to reduce $R = 251.5$ units from $q_1 = 675$. The prices of nodes 1 and 2 will reach an equilibrium $\frac{m_1 + s_1 + m_3}{(q_1 - R) - m_3/b_3^1} = 8.53$. \square

3 Correctness

We concern about whether the prices of goods eventually reach an equilibrium price even if they are initially distinct. So we define the equilibrium price as a legitimate configuration, and then show the legitimate configuration will be eventually achieved.

Definition 1 (legitimate configuration). A configuration is legitimate if every node has equally priced goods. \square

In [9], we examined a sufficient condition for price stabilization in **Naive**. Suppose that agents a_i and a_j make bids to node $h \in N_i \cap N_j$. We say that *bids have the same order as values* if $v_i^h \leq v_j^h$ implies $b_i^h \leq b_j^h$ for the goods of node h . The following theorem further shows that an additional condition leads to the price stabilization.

Theorem 1 [9]. Suppose bids keep the same order as values. If any contract price lies between buyer's price and seller's price, price stabilization occurs. \square

Based on the condition above, we consider the stabilizing property of our multiplier-effect protocol. Let I and R be the quantity of produced goods by an investment and the quantity of reduced goods, respectively. Then, we have the following lemmas.

Lemma 1. Suppose that agents a_i and a_j ($p_i > p_j$) carry out a transaction. Then, agent a_j should newly produce goods

$$I = \frac{1}{s_i}(q_i + m_i/b_i^j)(m_j + s_j + m_i) - (q_j - m_i/b_i^j) \quad (\text{ii})$$

if agent a_i can buy all of q_j , that is,

$$\frac{s_i}{q_i + m_i/b_i^j} \leq \frac{m_j + s_j + m_i}{q_j - m_i/b_i^j}.$$

Proof. We denote by $b = b_i^j$ for simplicity. Let x be the number of units passed in a transaction. The condition that the price reaches an equilibrium and an agent spends all his disposable money is

$$\frac{(m_i - bx) + s_i}{q_i + x} = \frac{(m_j + s_j) + bx}{(q_j + I) - x}$$

and

$$m_i - bx = 0.$$

Then, we have $x = m_i/b$ and I above. \square

The following fact can be similarly shown.

Lemma 2. Suppose that agents a_i and a_j ($p_i > p_j$) carry out a transaction. Then, agent a_j should reduce inventory stocks

$$R = (q_j - m_i/b_i^j) - \frac{1}{s_i} (q_i + m_i/b_i^j)(m_j + s_j + m_i)$$

if agent a_i cannot buy all of q_j , that is,

$$\frac{s_i}{q_i + m_i/b_i^j} > \frac{m_j + s_j + m_i}{q_j - m_i/b_i^j}.$$

\square

In what follows, we call the set of same priced, neighboring nodes a *cluster*. We pay attention to the number of clusters in G during the execution of **MultiplierEffect**.

Lemma 3. After spending all the money, the number of clusters is monotone decreasing in the protocol **MultiplierEffect**.

Proof. By Lemmas 1 and 2, since the quantity of goods often changes, the price of goods also changes. Thus, the number of clusters may sometimes increase under stabilization. So we consider the situation when every agent has spent all his money, which will eventually be reached. There are three cases. Let agent a_j make a contract with a_h , where $p_j < p_h$.

1. Agent a_h 's money m_h is not empty, and goods q_j does not run out even if a_h paid all his money m_h .
2. Agent a_h 's money m_h is not empty, and goods q_j is not enough if a_h paid all his money m_h .
3. Agent a_h 's money m_h is empty.

In case 1, the new price is determined between p_j and p_h . In case 2, agent a_j 's quantity of goods increases and both the prices p_j and p_h fall. In both cases above, the number of clusters may increase. In case 3, agent a_h pays no money to a_j and agent a_j reduces his production. Then, the price p_j rises to p_h , that is, the number of clusters decreases.

Since every agent eventually spends all his disposable money, the number of clusters is monotone decreasing. \square

Theorem 2. *By the protocol MultiplierEffect, the legitimate configuration will be eventually achieved.*

Proof. We can say that the number of clusters is monotone decreasing after some point by Lemma 3. Furthermore, no agent let his transaction wait for other operations. So no deadlock occurs. There exist price gap(s) between clusters as long as the number of clusters is greater than 1. Then, the rules 1–3 of MultiplierEffect are applied and the configuration changes to the next one. If the number of clusters becomes 1, the legitimate configuration is achieved. \square

Lemma 4. *Some money may remain unused after stabilization.*

Proof. There are two cases in which some money remains unused. First, if much money enough to complete stabilization is given, it is clear that some money remains unused. Next, suppose that $p_i > p_j$ and $p_k = p_i$ for any $a_k \in N_i$. If $m_i = 0$, the goods q_j is asked to reduce until $p_i = p_j$ is achieved. Then, $m_j > 0$ remains unused. \square

Theorem 3. *Let M be a total money. Then, the total used money during the stabilization is at most*

$$\frac{M}{1 - \beta}.$$

Proof. Let m_i be the initial money of agent a_i . Agent a_i will pay m_i to agent a_j and βm_i is stocked as a disposable money. Then, agent a_j will pay βm_i to agent a_k and $\beta^2 m_i$ is stocked as a disposable money, and so on. Notice that some money may remain unused after the stabilization by Lemma 4. So, we can estimate that $m_i + \beta m_i + \beta^2 m_i + \dots \leq \frac{m_i}{1 - \beta}$ will be paid. Likewise, the initial money of any agent will be also paid. Thus, the total used money is at most

$$\sum_i \frac{m_i}{1 - \beta} = \frac{M}{1 - \beta}. \quad \square$$

4 Simulation

In this section, we run simulation experiments for our multiplier-effect protocol in path, grid and complete graph networks. We investigate the influence of the network topologies and other aspects on the consumption of money.

We compare the three networks with respect to how

- (1) the bid level, and
- (2) the block size

have great influence on the number of reduced inventory stocks. Notice that the reduction of inventory stocks, caused by an excess supply, is an undesirable process. So we have to evaluate when small number of reduced inventory stocks is achieved. More precisely, (1) we vary the bid level B_L from 0.1 to 0.9, that is, the level of bidding price between bidder a_i 's price and seller a_j 's price in

$$b_i^j = p_j + B_L \cdot (p_i - p_j).$$

In (2) we vary the block size, the size of a partial transaction executed in a round. In other words, we represent the block size as B_S such that agent a_i spends $B_S \cdot m_i$ in a round.

Table 1 shows the constants used in our experiments. We repeat the experiment up to 50 trials, where a trial ends with an equilibrium, and obtain mean results. The total number of nodes is 100. Initially, each node has money and savings 500 units each, and has goods between 50 and 150 units at random. Then, the relation (i) determines the price for each node. Table 2 shows the parameters used in our experiments. It has the third column named “standard” which means a constant value if another parameter is varying. For example, the bid level is 0.5 when we vary the transaction size from 0.1 to 0.9. Notice that the number 100 of nodes is sufficient because an experiment of varying number of nodes from 100 to 500 has shown no influence on the reduction of inventory stocks.

Table 1. Constants

Meaning	Value
Number of trials	50
Number of nodes	100
(Money (m_i), Savings (s_i)) per node	(500, 500)
Quantity of goods per node (q_i)	[50, 150]
Marginal propensity to consume (β)	0.5

Table 2. Parameters

Meaning	Value	Standard
Bid level (B_L)	0.1–0.9	0.5
Block size (B_S)	0.1–0.9	0.5

4.1 Influence of Bid Level

Figure 2 shows how the bid level has influence on the reduction of inventory stocks in three networks. All the curves decrease as the bid level grows because the quantity of newly produced goods is small by (ii) in Lemma 1. The bidding conflicts in the path do not frequently occur because almost all nodes have only two links. More (resp. less) conflicts occur in the grid than those in the path (resp. the complete graph). Thus the highest (resp. lowest) bidding price tends to occur in the complete graph (resp. the path). This means that the number of reduced inventory stocks is small in the complete graph and is large in the path.

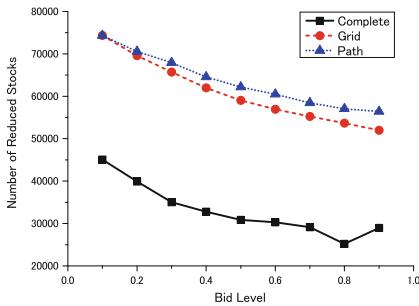


Fig. 2. Varying bid level

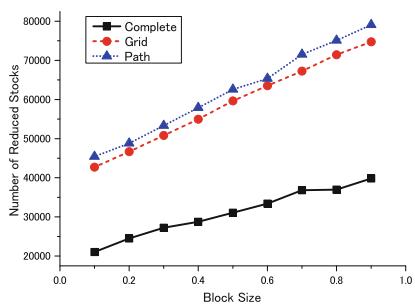


Fig. 3. Varying block size

4.2 Influence of Block Size

Figure 3 shows how the block size has influence on the reduction of inventory stocks in three networks. All the curves increase as the block size grows because the quantity of newly produced goods tends to exceed demand. The same reason as above holds, that is, the number of reduced inventory stocks is small for the dense network and is large for the sparse network, respectively.

5 Conclusion

In this paper we presented a multiplier effect model for the price stabilization in networks. Then we have obtained the following two advantages:

- we can consider a fiscal policy as well as a monetary policy in our multiagent network model, and
- we can investigate some influences of parameters by using our protocol.

First, we described the correctness and some properties of our protocol in Sect. 3. Then, we ran simulation experiments and revealed several features of our protocol in several networks in Sect. 4.

We showed that our protocol has a multiplier effect on the total used money during the stabilization. Furthermore, we showed that the higher bid level has the better influence on the number of reduced inventory stocks. This also holds for dense networks.

Our future work includes developing a practical stabilization model, for example, on-line shopping, consumption tax, and other protocols.

Acknowledgements. The authors would like to thank Dr. Kensaku Kikuta for useful discussion and helpful comments. This work was supported by JSPS KAKENHI Grant Number ((C)17K01281).

References

1. Assenza, T., Gatti, D.D., Grazzini, J.: Emergent dynamics of a macroeconomic agent based model with capital and credit. *J. Econ. Dyn. Control* **50**, 5–28 (2015)
2. Benhabib, J., Jackson, M.O., Bisin, A. (eds.): *Handbook of Social Economics*, vol. 1a. North Holland, Amsterdam (2010)
3. Dolev, S., Kat, R.I., Schiller, E.M.: When consensus meets self-stabilization. *J. Comput. Syst. Sci.* **76**(8), 884–900 (2010)
4. Dolev, S.: *Self-stabilization*. The MIT Press, Cambridge (2000)
5. Even-Dar, E., Kearns, M., Suri, S.: A network formation game for bipartite exchange economies. In: Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms (SODA 2007), pp. 697–706 (2007)
6. Jackson, M.O., Wolinsky, A.: A strategic model of social and economic networks. *J. Econ. Theory* **71**(1), 44–74 (1996)
7. Jackson, M.O., Watts, A.: Social games: matching and the play of finitely repeated games. *Games Econ. Behav.* **70**(1), 170–191 (2010)
8. Kiniwa, J., Kikuta, K.: A network model for price stabilization. In: Proceedings of the 3rd International Conference on Agents and Artificial Intelligence (ICAART), pp. 394–397 (2011)
9. Kiniwa, J., Kikuta, K.: Price stabilization in networks – what is an appropriate model? In: Defago, X., Petit, F., Villain, V. (eds.) *SSS 2011. LNCS*, vol. 6976, pp. 283–295. Springer, Heidelberg (2011)
10. Kiniwa, J., Kikuta, K., Sandoh, H.: A price stabilization model in networks. *J. Oper. Res. Soc. Jpn.* **60**(4), 479–495 (2017)
11. Kiniwa, J., Kikuta, K., Sandoh, H.: Equilibrium bidding protocols for price stabilization in networks. In: 4th International Conference on Behavioral, Economic, and Socio-Cultural Computing, BESC 2017, pp. 1–6 (2017)
12. Kiniwa, J., Kikuta, K., Sandoh, H.: Asynchronous price stabilization model in networks. In: Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICAART), vol. 1, pp. 121–128 (2018)
13. Kranton, R., Minehart, D.: Theory of buyer-seller networks. *Am. Econ. Rev.* **91**(3), 485–508 (2001)
14. Krishna, V.: *Auction Theory*. Academic Press, Cambridge (2002)
15. Lynch, N.A.: *Distributed Algorithms*. Morgan Kaufmann Publishers, Burlington (1996)
16. Mankiw, N.G.: *Macroeconomics*. Worth Publishers, New York (2015)
17. Mankiw, N.G.: *Principles of Economics*. Cengage Learning, Boston (2018)

18. Murota, R., Ono, Y.: Fiscal policy under deflationary gap and long-run stagnation: reinterpretation of Keynesian multipliers. *Econ. Model.* **51**, 596–603 (2015)
19. Riccetti, L., Russo, A., Gallegati, M.: Financialisation and crisis in an agent based macroeconomic model. *Econ. Model.* **52**, 162–172 (2016)
20. Wang, Y., Xu, Y., Liu, L.: Keynesian multiplier versus velocity of money. *Phys. Procedia* **3**, 1707–1712 (2010)



Sector Neutral Portfolios: Long Memory Motifs Persistence in Market Structure Dynamics

Jeremy D. Turiel^{1(✉)} and Tomaso Aste^{1,2,3}

¹ Department of Computer Science, University College London,
Gower St, Bloomsbury, London WC1E 6BT, UK

{jeremy.turiel.18,t.aste}@ucl.ac.uk

² UCL Centre for Blockchain Technologies, University College London,
Gower St, Bloomsbury, London WC1E 6BT, UK

³ Systemic Risk Centre, London School of Economics and Political Science,
Houghton Street, London WC2A 2AE, UK
http://www.cs.ucl.ac.uk/staff/tomaso_aste/
<http://blockchain.cs.ucl.ac.uk/tomaso-aste/>

Abstract. We study soft persistence (existence in subsequent temporal layers of motifs from the initial layer) of motif structures in Triangulated Maximally Filtered Graphs (TMFG) generated from time-varying Kendall correlation matrices computed from stock prices log-returns over rolling windows with exponential smoothing. We observe long-memory processes in these structures in the form of power law decays in the number of persistent motifs. The decays then transition to a plateau regime with a power-law decay with smaller exponent. We demonstrate that identifying persistent motifs allows for forecasting and applications to portfolio diversification. Balanced portfolios are often constructed from the analysis of historic correlations, however not all past correlations are persistently reflected into the future. Sector neutrality has also been a central theme in portfolio diversification and systemic risk. We present an unsupervised technique to identify persistently correlated sets of stocks. These are empirically found to identify sectors driven by strong fundamentals. Applications of these findings are tested in two distinct ways on four different markets, resulting in significant reduction in portfolio volatility. A persistence-based measure for portfolio allocation is proposed and shown to outperform volatility weighting when tested out of sample.

Keywords: Portfolio diversification · Market structure · Networks · Sector diversification

1 Introduction

Portfolio diversification has been a central theme for investors and the financial industry since the beginning of the 20th century, with the first long-lasting academic results produced by Markowitz [1]. Sector diversification, in particular, is

a non-trivial type of diversification for localised systemic risk. This risk arises from persistently correlated groups of stocks which often correspond to industry sectors. Correlations within these groups of stocks are found to be highly persistent in time, this should be accounted for when allocating capital within a portfolio. These persistently correlated groups are often subject to similar regulatory, political and resource shocks. Examples of this are stocks belonging to the Oil and Gas, Healthcare or Pharmaceutical sectors. It will be shown in the present work that (absolute) correlation values and their persistence should be treated differently for the purpose of portfolio diversification as they represent different properties of the system and are only weakly related.

The present work introduces an unsupervised technique to identify groups of stocks which share strong fundamental price drivers. This technique can be of particular impact in less traded markets, where identifying structures with shared fundamental price drivers might require in-depth knowledge of the companies. A persistence-based measure is also proposed to optimise portfolio allocation and tested for out of sample performance against $1/\sigma$ weighting (where σ is the standard deviation of log-returns from which the correlation matrix is constructed).

Correlation networks [2] and network filtering techniques applied to the study of financial assets have recently gained wide attention [3–10]. These methods show that meaningful taxonomy of financial assets is identifiable from these sparse network structures. Filtering through the Minimum Spanning Tree (MST) technique was initially suggested by Mantegna [2], this concept was further extended to planar graphs with the Planar Maximally Filtered Graph (PMFG) [11] and more recently to chordal graphs with predefined motif structure, as the TMFG in [12] and the Maximally Filtered Clique Forest (MCFC) in [4].

Correlations are noisy measures of comovement of financial asset prices, which are often non-stationary within the observation window. Longer windows benefit the measure's stability, as we have more observations to estimate the $N(N - 1)/2$ parameters of the matrix of N assets. However, a longer observation window can come with the disadvantage of weighting more and less recent movements equally with the risk of averaging over multiple non-stationarities. In order to compensate for this effect, we apply the exponential smoothing method as discussed in Pozzi et al. for Kendall correlations [13]. This allows for more stable correlations, while prioritising recent observations. The method applies an exponential weighting to the correlation window, prioritising more recently observed comovements.

The rest of the paper is structured as follows: Sect. 2 describes the methods applied and defines measures which are used throughout the paper. Section 3 describes the results obtained, with Sect. 3.1 introducing long-term memory processes in persistence, Sect. 3.2 analysing market development through its decay exponents, Sect. 3.3 illustrating the coherence of highly persistent motifs with sectors and Sect. 3.4 outlining various results which highlight the importance of this work for portfolio allocation. Section 4 then presents an analysis of the results from Sect. 3 and Sect. 5 concludes the paper with a summary and thoughts for further work.

2 Method

In the present paper we apply the TMFG [12] to filter matrices obtained from Kendall correlations with exponential smoothing, applying the method by Pozzi et al. [13].

2.1 Data

We select the 100 most capitalised stocks from the NYSE, Italy, Germany and Israel's markets (400 in total). Markets range from highly liquid and more developed ones such as the NYSE and Germany to less liquid and stable markets such as Italy and Israel.

We investigate daily closing price data from Bloomberg between 3/01/2014 for the NYSE, Germany and Italy (5/01/2014 for Israel) and 31/12/2018 (inclusive) for the NYSE (28/12/2018 for Germany and Italy, 1/1/2019 for Israel). The data is composed of 1258 observations for the NYSE, 1272 for Italy and Germany and 1225 for Israel.

2.2 TMFG Network Motif Persistence

We look at temporal persistence of tetrahedral and triangular motifs in the TMFGs constructed over rolling windows. TMFG networks can be viewed as trees of tetrahedral (maximal) cliques connected by common triangular faces, these are then triangular cliques with different meaning in the taxonomy, called separators. Not all triangular faces of the tetrahedral cliques are separators and we will refer to those which are not as triangles (these do not include separators in the way we shall refer to them).

Differently from “hard” persistence (survival) of motifs between consecutive layers in the temporal network, which is more common in the literature [14, 15], here we apply a form of “soft” persistence. A motif corresponding to clique \mathcal{X}_c is considered “soft” persistent at time $t + \tau$ if and only if the motif is present at both the initial time t and at $t + \tau$.

Considering the motif sets $\mathcal{X}_C^t = \{\mathcal{X}_i^t\}_{i=1,\dots,C}$ and $\mathcal{X}_C^{t+\tau} = \{\mathcal{X}_i^{t+\tau}\}_{i=1,\dots,C}$, the binary persistence value of motif $c \in C$ at time t and $t + \tau$ is

$$P_m(\mathcal{X}_c^{t,t+\tau}) = (\mathcal{X}_c \in \mathcal{X}_C^t) \wedge (\mathcal{X}_c \in \mathcal{X}_C^{t+\tau}) \quad (1)$$

Where $P_m(\mathcal{X}_c^{t,t+\tau})$ represents the binary persistence value of motif $c \in C$ at times t and $t + \tau$.

2.3 Portfolio Construction

We investigate the decay in the number of persistent motifs between filtered TMFG correlation networks with observation windows progressively shifted by one trading day. We iterate over $t = [0, \dots, 200[$ different starting correlation networks and investigate persistence up to a time shift of $\tau = 900$ days. Hence the

analysis covers a significant portion of temporal layers which do not overlap with the time window of the initial layer. We observe $\langle P_m(\mathcal{X}^\tau) \rangle_{T,C}$ from Eq. 3 to decay with the time shift τ . We then obtain the power law fit for the decay law and identify the two regimes: one with a faster decay followed by one with a slower decay. The transition point is computed by minimising the unweighted average mean squared error (MSE) between the two fits over all possible transition points in time.

The average motif persistence in the plateau regime is defined as

$$\langle P_m(\mathcal{X}_c) \rangle_{T,\tau} = \frac{1}{T} \cdot \frac{1}{\tau - \tau_{plat}} \cdot \sum_{t=0}^T \sum_{\tau=\tau_{plat}}^{\mathcal{T}} P_m(\mathcal{X}_c^{t,t+\tau}) \quad (2)$$

Where τ_{plat} identifies the transition point to the plateau region identified by minimising the Mean Squared Error (MSE), as explained below.

The average persistence for the entire clique set over T starting points at time shift τ is defined as

$$\langle P_m(\mathcal{X}^\tau) \rangle_{T,C} = \frac{1}{T} \cdot \frac{1}{|C|} \cdot \sum_{t=0}^T \sum_{c \in C} P_m(\mathcal{X}_c^{t,t+\tau}) \quad (3)$$

We also compare the decay exponents for multiple random stock selections over different markets to identify whether the steepness of motif decay (edge, closed triad or tetrahedron clique) is indicative of market stability/development stage. We further investigate more liquid markets such as the NYSE from both a quantitative and qualitative point of view. We classify motifs in the plateau by their “soft” persistence and study the sector structure of the most persistent motifs. We also verify that these motifs are not trivially retrieved by maximum correlation edges or motifs in the correlation matrix.

In order to further justify the analysis of motifs over individual edges, we test to reject the assumption that motifs are formed by edges in the network whose existence is not mutually dependent. The assumption would imply that coexistence of edges in motifs is not statistically significant and that motif structures have no extra persistence beyond the individual edges that form them. The hypothesis being tested implies that motif persistence is simply the result of persistence characterising their component edges:

$$P_m(\chi_c^{t,t+\tau}) = P_m(\chi_{c1}^{t,t+\tau}) \cdot P_m(\chi_{c2}^{t,t+\tau}) \cdot P_m(\chi_{c3}^{t,t+\tau}) \quad (4)$$

Where the motif and its edges are defined as $\chi_c^{t,t+\tau} = \{\chi_{c1}^{t,t+\tau}, \chi_{c2}^{t,t+\tau}, \chi_{c3}^{t,t+\tau}\}$.

In order to provide an industry-oriented point of view, we construct a portfolio containing all stocks in the ten most persistent motifs (for each market) and compare its volatility with that of random portfolios.

We conclude by defining the persistence measure $P_m(v_i)$.

$$P_m(v_i) = \sum_{\mathcal{X}_c \in \mathcal{X}_C | i \in \mathcal{X}_c} \langle P_m(\mathcal{X}_c) \rangle_{T,\tau} \quad (5)$$

The measure presented in Eq. 5 is defined for each vertex v_i in the network as the sum over all $\langle P_m(\mathcal{X}_c) \rangle_{T,\tau}$ (average persistence of motif \mathcal{X}_c in the plateau) where vertex v_i belongs to clique \mathcal{X}_c .

$P_m(v_i)$ is defined in Eq. 5 to compare random portfolios weighted by $1/\sigma$ with those weighted by $1/P_m(v_i)$. In Sect. 3.4, we do this for the four different markets, with all results showing meaningful volatility reductions.

3 Results

3.1 Long-Term Memory of Motif Structures

The plot in Fig. 1 shows the power law decay (evident from the linear trend in log-log scale) in $\langle P_m(\mathcal{X}^\tau) \rangle_{T=200,C}$ vs. τ , followed by a plateau region. We also observe that all motif decays have $\tau_{plat} = [\delta t_{window}/2, \delta t_{window}]$, where δt_{window} is the initial window's time span. The window used has $\delta t_{window} = 126$ trading days and a value of $\theta = 46$ for exponential smoothing, as per [13]. The choice of δt_{window} corresponds to roughly 6 months of trading and satisfies $N < \delta t_{window}$, with N the number of assets in the correlation matrix. The correlation matrix is hence well-conditioned and invertible.

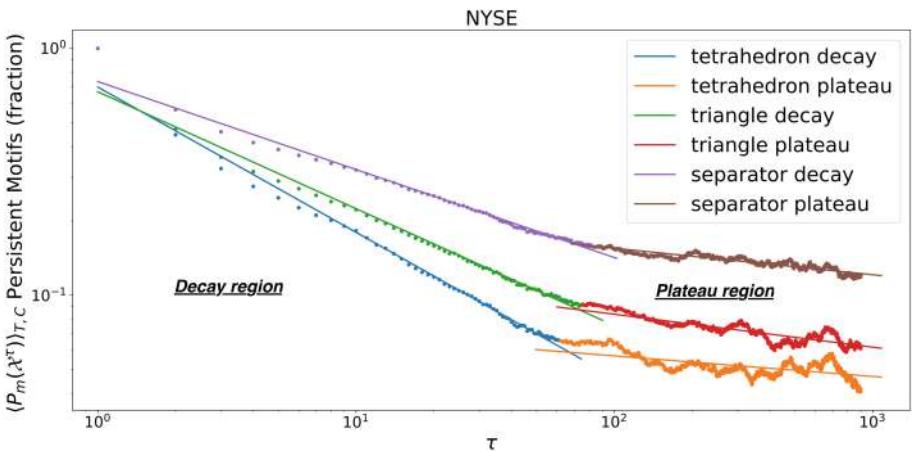


Fig. 1. Decay of tetrahedral clique, triangle and separator $\langle P_m(\mathcal{X}^\tau) \rangle_{T=200,C}$ for the 100 NYSE stocks with highest market capitalisation vs. τ . The two power-law regimes are identified by the minimum MSE sum of the fits.

As per the plot in Fig. 1, there are $N - 3 = 97$ cliques in the starting TMFG networks and $3N - 8 = 292$ face triangles.

In Fig. 1 we notice that the minimum MSE for the two linear fits is achieved at the transition point between the decay phase and the plateau. The transition point can therefore be identified by minimising a standard fit measure, this strengthens the unsupervised nature of our method. The method for minimum MSE search is described in Sect. 2.

3.2 Market Classification via Decay Exponent

We now consider how the decay exponent behaves across markets. Table 1 compares the decay exponents for cliques, triangular motifs and clique separators in the NYSE, German stock market, Italian stock market and Israeli stock market.

Table 1. Exponents for the decay power law regime computed with MSE. The analysis refers to 100 randomly selected stocks amongst the 500 most capitalised, over time intervals $\tau = [0, 900]$ and $t = [0, \dots, 200]$ different initial temporal network layers. For all motif analyses in this work triangles and separators constitute non-overlapping sets, as these represent theoretically and taxonomically different structures and decay characteristics.

Market	Clique	Triangular motif	Clique separator
NYSE	-0.392	-0.493	-0.245
Germany	-0.792	-0.598	-0.381
Italy	-0.785	-0.811	-0.174*
Israel	-1.024	-0.866	-0.728

*Result compromised by regimes not well identified for motif decay in large systems (≈ 100 stocks)

We notice from the results in Table 1 that the NYSE, which is clearly the most developed and liquid stock market, has the lowest decay exponent (in modulus, which corresponds to the slowest decay) for both cliques and triangles. This indicates that its correlations are more stable on a shorter time window, due to a higher signal to noise ratio. Germany and Italy have similar values for clique exponents, with Germany seemingly more stable in terms of triangular motifs. Israel, a younger and less liquid stock market, follows with a faster decay in both cliques and triangular motifs. The ordering of these markets is not clearly identifiable in clique separators as noise in the data does not allow for the two decay regimes discussed in Sect. 3.1 to be correctly identified in all markets (in this case for Italy). Separators have a distinct role and meaning in the graph's taxonomy and further work should allow for a more thorough analysis of those.

In Table 1 the decay exponent is not adjusted by the probability that all edges in the clique must be present in the temporal layer for the clique to exist. We show in Table 2 that, when adjusted by the probability of all its edges existing simultaneously, triangular motifs have a slower decay than individual edges. In order to strengthen the consistence of the phenomenon across buckets of randomly selected stocks, Table 2 corresponds to a different random bucket than Table 1.

We stress that Table 2 falsifies the hypothesis discussed in Sect. 2 that motifs are formed by edges in the network whose existence is not mutually dependent. This is falsified by the consistently lower decay exponent (in modulus) for adjusted persistence of triangular motifs. We can then conclude that motifs are more stable structures across temporal layers of the network, with significant interdependencies in their edges' existence.

Table 2. Exponent for the power law decay regime identified by MSE in different sample markets. The analysis refers to 100 randomly selected stocks amongst the 500 most capitalised, over time intervals $\tau = [0, 900[$ and $t = [0, \dots, 200[$ different initial temporal network layers.

Market	Edge	Triangular motif	Triangular motif**
NYSE	-0.164	-0.398	-0.133
Germany	-0.265	-0.471	-0.157
Italy	-0.144*	-0.458	-0.153
Israel	-0.397	-0.830	-0.277

*Result compromised by regimes not well identified for edge decay in large systems (≈ 100 stocks)

**Motif exponent adjusted by the probability of simultaneous edge persistence in the motif

3.3 Sector Analysis in Persistent Motifs

Figure 2 provides a visualisation of the network components formed by the ten most persistent triangles in the NYSE. We observe that all strongly persistent triangles have elements which belong to the same industry sector. Table 3 shows this for the ten most persistent triangles displayed in Fig. 2.

We notice that most sectors for the motifs in Table 3 share strong fundamental price drivers, which justify the persistent structure in the long term, as per the discussion in Sect. 1. Other motifs are constituted by ETFs with similar underlying assets (Vanguard FTSE ETF, MSCI EAFE ETF, Vanguard FTSE ETF) or the NASDAQ ETF with its main holdings (Amazon and Alphabet). The reason for the existence of these motifs is intuitive and does not affect our analysis, as ETF-related motifs are unlikely to be present in the network formed by a random selection of stocks or by stocks in a portfolio. These motifs are present here as we focus on the 100 most capitalised securities in the NYSE, which include ETFs.

We also investigate whether motif persistence and motif structures can be easily retrieved from the original correlation matrix. The purpose of this is to check that our method is not redundant and trivially replaceable. To test this, we consider the ten most present persistent triangles across the plateau region and check their overlap with the ten most correlated triplets in each unfiltered correlation matrix. We find that no more than one triangle lies in the intersection between the two sets, in each temporal layer. We also check the correlation between motif persistence and the average sum or product (results are equivalent for our purpose) of its individual edges' correlation for all unfiltered correlation layers. We observed both visually (by means of different plots) and statistically (through the Pearson and Kendall correlation measures) that the two measures are only loosely related, if they are at all. This was reflected by plots with no

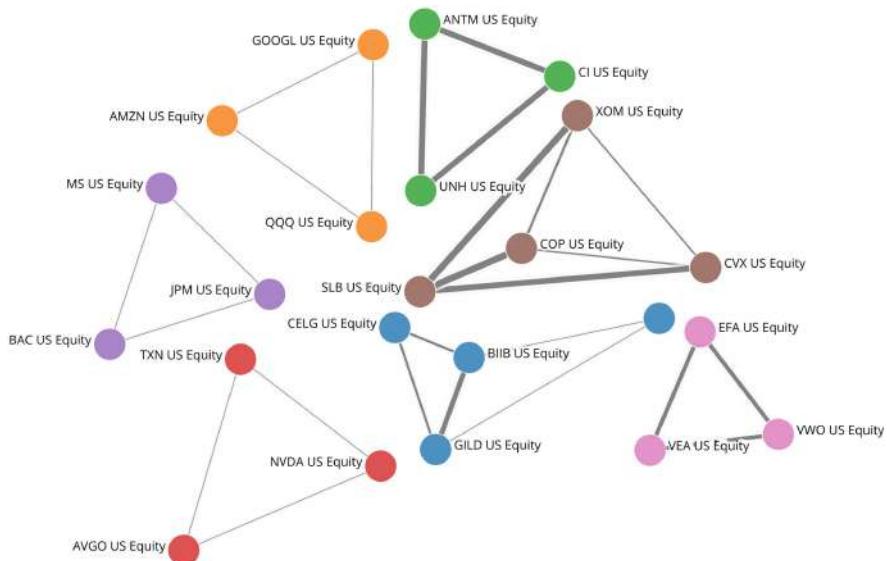


Fig. 2. Network representation of the ten most persistent triangular motifs in the plateau (highest $\langle P_m(\mathcal{X}_c) \rangle_{T,T}$) for the 100 most capitalised stocks in the NYSE.

clear trend or apparent functional form. The variables also presented significant, low correlation values, where the correlation explained no more than 20% of the variance in the persistence values. This result is especially significant for a wide power law distribution as that of persistence values.

Table 3. Motif components and Financial Times sector affiliation for the ten most persistent motifs in the NYSE's 100 most capitalised stocks.

Security 1	Security 2	Security 3	FT sector
Biogen Inc.	Gilead Sciences Inc.	Celgene Corp	Biopharmaceutical
UnitedHealth Group Inc.	Cigna Corp	Anthem Inc.	Health Care
Biogen Inc.	Gilead Sciences Inc.	Amgen Inc.	Biopharma/tech
Bank of America Corp	JPMorgan Chase & Co	Morgan Stanley	Financials-Banks
Vanguard FTSE ETF**	MSCI EAFE ETF	Vanguard FTSE ETF***	Index ETFs
Invesco QQQ Trust*	Amazon.com Inc.	Alphabet Inc.	Tech
ConocoPhillips	Schlumberger NV	Exxon Mobil Corp	Oil & Gas
NVIDIA Corp	Texas Instruments Inc.	Broadcom Inc.	Tech Hardware
Chevron Corp	Schlumberger NV	Exxon Mobil Corp	Oil & Gas
Chevron Corp	ConocoPhillips	Schlumberger NV	Oil & Gas

*ETF on NASDAQ - Top Holdings include Amazon, Facebook, Apple, Alphabet

**Vanguard FTSE Developed Markets Index Fund ETF Shares

***Vanguard FTSE Emerging Markets Index Fund ETF Shares

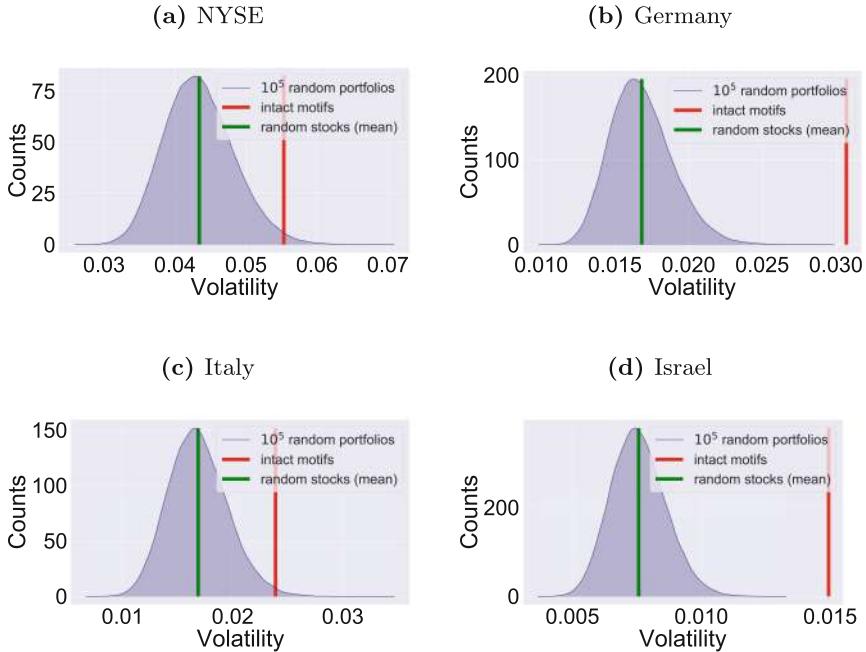


Fig. 3. Portfolio volatility distribution for the 100 most capitalised stocks in the NYSE (a), German stock market (b), Italian stock market (c) and Israeli stock market (d). The reference portfolio (red bar) contains all stocks in the 10 most persistent triangles and distribution portfolios are formed from a random selection of stocks (mean distribution volatility represented by the green bar).

3.4 Long-Only Portfolio Diversification Across Markets

3.4.1 Motif vs. Random Portfolios

We check that a portfolio formed by the 10 most persistent motifs in each market has a highly enhanced out of sample volatility due to its stable correlations.

This is shown in Fig. 3 where we consider the volatility of the motif portfolio and a distribution of volatilities for 10^5 randomly selected portfolios with the same number of stocks.

As expected, we observe the motif portfolio to yield a volatility close to the higher end of the distribution. We should highlight that the volatility of portfolios is evaluated out of sample with respect to the period the persistence was calculated on, making this method not only observational, but also predictive.

3.4.2 Volatility vs. Persistence Weighting - An Industry Standard Comparison

We now deploy our findings to form a measure directly applicable to portfolio optimisation and compare it with the widespread inverse volatility $1/\sigma$ weighting. In Fig. 4 we present out of sample results where we observe a reduction in

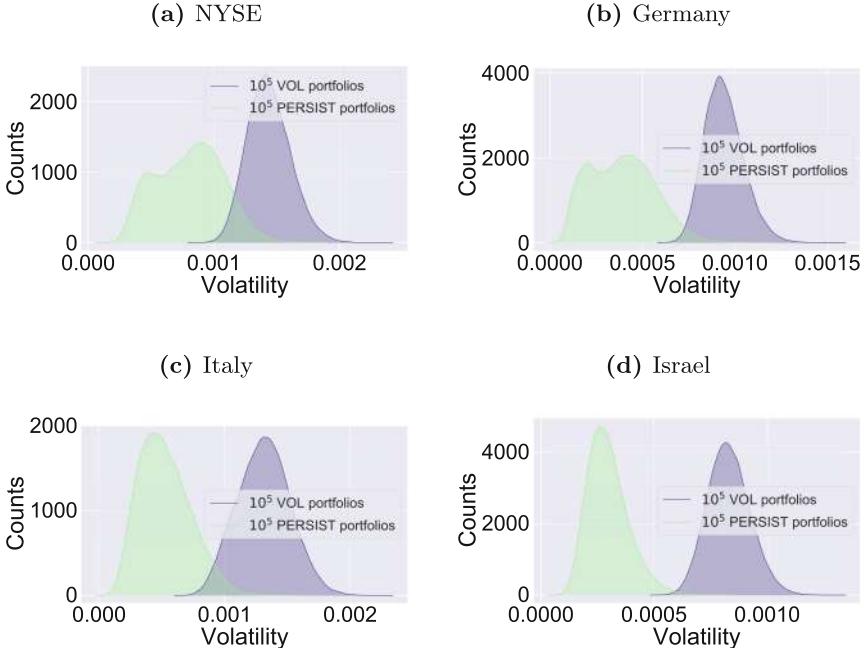


Fig. 4. Portfolio volatility distribution for the 100 most capitalised stocks in the NYSE (a), German stock market (b), Italian stock market (c) and Israeli stock market (d). The VOL portfolios are formed by weighting a random portfolio of assets by $1/\sigma$ while the PERSIST portfolios are formed weighting assets by $1/P_m(v_i)$.

volatility throughout markets, with distributions separated beyond one standard deviation.

This comparison is meaningful beyond that of an industry standard with a novel approach. The volatility weighting is based on individual assets taken in isolation, weakly influenced by the portfolio's composition of assets. The persistence-based weighting is instead strongly based upon the cliques (therefore the other assets) present in the portfolio. Portfolio composition also influences how the network is filtered, providing a second level of the system's influence on the asset's weighting by persistence. This shows how network analysis and complex systems can greatly enhance our understanding of real world systems beyond traditional methods.

4 Analysis

The power law decay identified in Fig. 1, as opposed to an exponential decay, can be interpreted in terms of long memory in the system. This corroborates observations by Bouchaud et al. and Lillo et al. in [16–19], where power law decays in autocorrelation are identified as manifestations of long-memory processes in efficient markets.

The strict ordering of markets based on their decay exponents in Table 1 can be interpreted in terms of more liquid markets being more structured and having less noisy correlations. Suggesting that more efficient and capitalised markets are characterised by structures which are more stable in time. The ordering also leads to the conclusion that more developed markets are characterised by more meaningful underlying structures and cliques, suggesting that systemic risk may represent a greater threat in developed markets. These are interesting observations in relation to the efficient market hypothesis, indicating that more liquid and efficient systems display more stable, autocorrelated and predictable market structures.

The hypothesis of motifs constituting meaningful structures in markets, beyond their individual edges, is strengthened by the results in Table 2. Table 2 tests the independence hypothesis of individual edges in motif formation and shows solid evidence to reject it. We can then state that highly persistent motifs are not a mere consequence of highly persistent individual edges, but also of the correlation in those edges existing concurrently.

Table 3 demonstrates the need to identify persistent motifs. The ten most persistent motifs visualised in Fig. 2 are representative of industry sectors in the NYSE. These sectors are not identified by the motifs with higher edge correlation, which instead are dominated by motifs often due to correlation noise in high volatility stocks. Persistent motifs are hence found to be non-trivial with respect to correlation strength of individual edges. The impact on portfolio diversification of the motifs in Fig. 2 indicates that these structures are highly relevant for diversification in medium to long term investment portfolios. The persistence of these motifs is an intrinsic temporal feature with forecasting power on market structure. As these motifs are not characterised by noticeably strong correlation, a common variance optimisation of the portfolio is unlikely to optimise the weights to sufficiently minimise the effect of these structures. We therefore suggest that filtering of these structures is then perhaps necessary prior to portfolio optimisation, for superior diversification outcomes.

Different results in application to portfolio diversification are presented in Sect. 3.4. We first show how persistent motifs increase the out of sample correlation of a portfolio by comparing portfolios containing all stocks from the ten most persistent motifs in each market with 10^5 random portfolios with the same number of stocks. We observe the motif portfolio volatility to be significantly above both the mean and median of the random portfolios' volatility distribution. This is a first example of how just selecting stocks from the ten most persistent motifs forms a portfolio with higher long term volatility. Clearly for investment purposes we want the opposite (small volatility). The observation from Fig. 3 lays the ground for the construction of portfolios where we optimise asset weights in order to reduce the volatility originating from persistent correlations in motif structures.

This is done in Sect. 3.4 where we propose a simple node-specific measure for portfolio weighting and selection. We show the out of sample volatility distribution of random portfolios with weights optimised as $1/P_m(v_i)$ to be significantly

lower than the distribution of portfolios optimised as $1/\sigma$, a widespread industry standard for portfolio weighting. This result can be explained by the persistence in time being the base of this measure, providing strong out of sample predictive power. Volatility is known to change in the medium to long term for most assets, whilst correlation is also difficult to estimate due to noise in the data and measures. This result greatly enhances the importance and applicability of this work to portfolio optimisation by providing a mapping from persistence-related observations to a direct measure for portfolio optimisation. Future works should investigate a technique to jointly optimise portfolio weighting for both persistence and volatility.

5 Conclusion

In the present work we have investigated four markets with different capitalisation, liquidity, economic and development characteristics. These markets range from the NYSE to the Israeli Stock Exchange. We construct correlation matrices for 100 stocks from Kendall correlations with exponential smoothing [13] and filter them with the TMFG [12], as described in Sect. 2.

We then base our study on market structure in the form of “soft” motif persistence. A two-regime power law decay in the number of persistent motifs with τ emerges. The two regimes can be identified via minimisation of the MSE fit measure, providing an unsupervised optimisation method with no tuning parameters. We find that advanced liquid markets are characterised by longer persistence of structured motifs. We argue that this could have consequences for systemic risk. We discuss long-term memory implications of this decay type and how they allow for forecasting power on market structure. Persistence is studied in order to investigate motif structures and retrieve meaningful sectors in each market. We then show that motif portfolios have significantly higher volatility than random ones. Conclusive results are obtained by defining a node-specific measure for portfolio optimisation and showing that it outperforms volatility-based weighting across markets. This result is of high importance to both practitioners and academics in the context of portfolio optimisation. Future works should investigate a portfolio optimisation technique combining persistence and correlation, based on results and methods from this work. Future works should also investigate applications to portfolio diversification in long-short portfolios and the decay in forecasting and diversification power with time.

References

1. Markowitz, H.: Portfolio selection. *J. Finan.* **7**(1), 77–91 (1952)
2. Rosario, N.M.: Hierarchical structure in financial markets. *Eur. Phys. J. B Condens. Matter Complex Syst.* **11**(1), 193–197 (1999)
3. Marcaccioli, R., Livan, G.: A polya urn approach to information filtering in complex networks. *Nat. Commun.* **10**(1), 745 (2019)
4. Massara, G.P., Aste, T.: Learning clique forests (2019). arXiv preprint [arXiv:1905.02266](https://arxiv.org/abs/1905.02266)

5. Miccichè, S., Mantegna, R.N.: A primer on statistically validated networks (2019). arXiv preprint [arXiv:1902.07074](https://arxiv.org/abs/1902.07074)
6. Musciotto, F., Marotta, L., Miccichè, S., Mantegna, R.N.: Bootstrap validation of links of a minimum spanning tree. *Phys. A Stat. Mech. Appl.* **512**, 1032–1043 (2018)
7. Jovanovic, F., Mantegna, R.N., Schinckus, C.: When financial economics influences physics: the role of econophysics. Available at SSRN 3294548 (2018)
8. Cimini, G., Squartini, T., Garlaschelli, D., Gabrielli, A., Caldarelli, G.: The statistical physics of real-world networks. *Nat. Rev. Phys.* **1**(1), 58 (2019)
9. Kojaku, S., Masuda, N.: Constructing networks by filtering correlation matrices: a null model approach (2019). arXiv preprint [arXiv:1903.10805](https://arxiv.org/abs/1903.10805)
10. Masuda, N., Kojaku, S., Sano, Y.: Configuration model for correlation matrices preserving the node strength. *Phys. Rev. E* **98**(1), 012312 (2018)
11. Tumminello, M., Aste, T., Di Matteo, T., Mantegna, R.N.: A tool for filtering information in complex systems. *Proc. Nat. Acad. Sci.* **102**(30), 10421–10426 (2005)
12. Massara, G.P., Di Matteo, T., Aste, T.: Network filtering for big data: triangulated maximally filtered graph. *J. Complex Netw.* **5**(2), 161–178 (2016)
13. Pozzi, F., Di Matteo, T., Aste, T.: Exponential smoothing weighted correlations. *Eur. Phys. J. B* **85**(6), 175 (2012)
14. Dessì, D., Cirrone, J., Recupero, D.R., Shasha, D.: Supernoder: a tool to discover over-represented modular structures in networks. *BMC Bioinf.* **19**(1), 318 (2018)
15. Musmeci, N., Aste, T., Di Matteo, T.: Risk diversification: a study of persistence with a filtered correlation-network approach (2014). arXiv preprint [arXiv:1410.5621](https://arxiv.org/abs/1410.5621)
16. Bouchaud, J.P., Gefen, Y., Potters, M., Wyart, M.: Fluctuations and response in financial markets: the subtle nature of ‘random’ price changes. *Quant. Finan.* **4**(2), 176–190 (2004)
17. Lillo, F., Farmer, J.D.: The long memory of the efficient market. *Stud. Nonlinear Dyn. Econom.* **8**(3), 1 (2004)
18. Bouchaud, J.P., Farmer, J.D., Lillo, F.: How markets slowly digest changes in supply and demand. In: *Handbook of Financial Markets: Dynamics and Evolution*, pp. 57–160. Elsevier (2009)
19. Di Matteo, T., Aste, T., Dacorogna, M.M.: Long-term memories of developed and emerging markets: using the scaling analysis to characterize their stage of development. *J. Bank. Finan.* **29**(4), 827–851 (2005)



Beyond Fortune 500: Women in a Global Network of Directors

Anna Evtushenko¹ and Michael T. Gastner^{2(✉)}

¹ Cornell University, Ithaca, NY 14853, USA
ae392@cornell.edu

² Yale-NUS College, 16 College Avenue West, #01-220, Singapore 138527, Singapore
michael.gastner@yale-nus.edu.sg

Abstract. In many countries, the representation of women on corporate boards of directors has become a topic of intense political debate. Social networking plays a crucial role in the appointment to a board so that an informed debate requires knowing where women are located in the network of directors. One way to quantify the network is by studying the links created by serving on the same board and by joint appointments on multiple boards. We analyse a network of $\approx 320\,000$ board members of 36 000 companies traded on stock exchanges all over the world, focusing specifically on the position of women in the network. Women only have $\approx 9\text{--}13\%$ of all seats, but they are not marginalised. Applying metrics from social network analysis, we find that their influence is close to that of men. We do not find evidence to support previous claims that women play the role of “queen bees” that exclude other women from similar positions.

Keywords: Interlocking directorates · Social networks · Gender inequality

1 Introduction

Females on boards of directors and board diversity more broadly are the topic of many studies [1, 8, 16]. Research has shown that female board representation is “positively related to accounting returns” [31]. The World Bank [36] estimates that 39% of the worldwide labour force in 2016 are women, but the percentage of women in leadership positions is much lower. Recent reports state that only 24% of senior management positions [17] and 15% of corporate board seats [6] are held by women. The percentage of female CEOs among Fortune 500 firms is even lower (6.4%) [27]. Women’s chances to become a CEO or a board member depend on multiple factors, such as “country wealth, gender egalitarianism and humane orientation” [13]. Nevertheless, female board participation is slowly on the rise globally. Shareholders and governments no longer regard it as a legitimate practice to recruit directors from an exclusively male “old-boys network” [11, 30]. Various countries, for example Israel [23] and Norway [32], have enacted laws

that favour the appointment of women [33]. As a consequence, there are signs that, at least in some European countries, the “glass ceiling” that has kept women out of the boardrooms is beginning to crack [14].

Once appointed, female directors must navigate intricate networks of professional relationships. A concrete manifestation of such a professional network are “interlocking directorates” [25] (i.e. the practice that some directors hold seats on more than one board). A recent survey by Credit Suisse [7] relates “overboarding” (i.e. an excessive number of board seats held by an individual) to the current trend towards increasing the number of female board members. The probability that a woman joins the board has been shown to be negatively correlated with the number of women currently on the board and to increase when a woman departs the board [15]. The underlying assumption is that companies tend to recruit “token women” (i.e. exactly one per board) from a limited pool of female candidates [10, 35]. Some commentators compare women directors with multiple seats to queen bees [37], implying that these women allegedly usurp power at the expense of female competitors. The purpose of this article is to test whether such narratives stand up to quantitative scrutiny.

Board interlocks can be inferred from a bipartite graph where every edge is between one company and one director (Fig. 1) [2]. Each director at one end of an edge sits on the board of the company at the other end of this edge. The study of board interlock networks started already in the 1970s [34], but at that time the role of the director’s gender was not yet in the limelight. Interest in the role of women on boards has intensified in recent years, see for example Ref. [12] for a critical review. Still, relatively little is known about the role that the gender plays for the network formed by interlocking directorates.

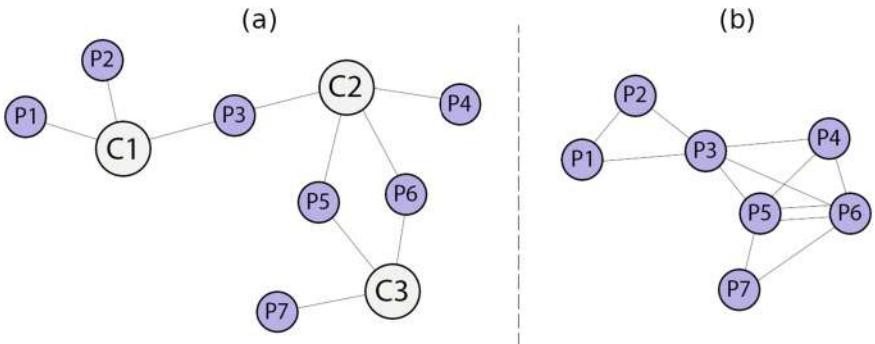


Fig. 1. Network representations of board interlock. (a) Bipartite graph where every edge is between a company (large node) and a person (i.e. a director, small node). (b) In this article, we analyse the social network of directors that results from a one-mode projection of the bipartite graph: directors are connected if and only if they sit together on a board.

A woman’s position in the social network certainly influences her chances to be appointed to a board [3], but only few papers have analysed the real

network [20, 26, 32, 39]. The most comprehensive quantitative network study to date is the PhD thesis by Hawarden [18], which looks at empirical data and builds a modelling framework called “Glass Network” theory. It posits the existence of “glass nets” that prevent women from assuming board seats, but those women who succeeded in crossing a glass net behave like queen bees. We further the exploration of this topic using network analysis applied to a large dataset, which we now describe.

2 Data

The source of our data is the Financial Times [38], an international English-language newspaper specialising in business and economics. Its website is a source of up-to-date information on financial markets and companies traded as equities. The website has data on the performance and structure of roughly 36 000 companies from 54 countries. The resulting data base is, to the best of our knowledge, the largest that has so far been used in the literature on board interlocks. The Financial Times (FT) receives their information from the media company Thomson Reuters, which in principle has data on yet more firms. However, we decided to use the FT data because this subset is more representative of companies that make up the business world.

The specific fields that we obtained for each company were name, unique code, sector, industry (i.e. subsector), country, revenue for the past 12 months, number of employees, date incorporated, and a list of directors, each if available. For each director, we recorded his or her name, gender, and age, each if included in the FT database. People were then identified as the same individual and assigned a unique ID if their three fields matched (for example, the names and genders were the same, and ages were blank) because we assume that the underlying Thomson Reuters database would have identical entries on each individual in all his or her companies. There were a total of 35 927 companies and 321 967 directors. Here we make no distinction between executive and non-executive directors.

In terms of missing data, 273 companies have zero directors listed. Among the fields relevant to the analysis, for 5732 companies (15.95%) we have no information about the country, for 4461 (12.41%) no sector, and for 5020 (13.97%) no industry. 96 751 directors (30.1%) are listed without gender. For 126 092 directors (39.2%), the database contains no information about their age. We show summary statistics of the network and the subgraphs consisting of only male or only female nodes in Table 1.

3 Summary Statistics of Node Attributes

The proportion of female directors in the FT data is 9.43% of all nodes and 13.49% among people with confirmed gender. This percentage is comparable to values stated in previous studies of international data. For example, Dawson et al. [6] found that women hold 14.7% of seats in the CS Gender 3000 data base;

Table 1. Statistics of the full network and the subgraphs consisting of only male or only female nodes. We calculate the average path length with the harmonic mean formula by Newman [28] to handle disconnected components.

Nodes	All	Male	Female
Edges	2809623	1092004	44666
Diameter	24	29	40
Average path length	13.90	22.90	517.79
Density	5.4×10^{-5}	5.7×10^{-5}	9.6×10^{-5}
Components	9404	12393	12355
% of nodes in the largest component	74.7	74.8	79.4
% of edges in the largest component	85.5	85.1	89.6

Deloitte [9] puts this number at 15% for data from nearly 7000 companies. Both of these reports emphasise that there can conceivably be a difference between the proportion of female *directors* and the proportion of female *seats* because of overboarding. On the boards of S&P500 companies, overboarding is more prevalent among women [7]. However, in the more comprehensive FT data, we find that, at the international level, the proportions of seats and directors are similar: the percentage of female seats is 9.73% and thus only 0.29% larger than the percentage of female directors. The underlying reason is that overboarding hardly differs between genders: 14.7% of women and 14.8% of men are multiple directors. Overall, taking into account ungendered nodes, 14.06% of directors are multiple directors. 4.22% of directors are on more than 2 boards, 1.76% on more than 3, and 0.40% on more than 5.

It is interesting to see whether the ratio of female seats to all seats differs by country, sector or industry. We can easily compute these numbers because for each company we know its country, sector, and industry, as well as the number of all directors and the number of female directors.

Figure 2 plots the proportion by sector and then subdivides each sector by industry, with industries following their sectors in descending order. We note that women are more represented in Financials, Consumer Services and Telecommunications than in Technology and Basic Materials. These findings are consistent with the report by Credit Suisse [6].

We find greater discrepancy between our data and Credit Suisse when we split our data by country (Fig. 3) instead of sector or industry. While we agree that Scandinavian countries generally rank highly, we find lower percentages of female seats than those reported by both Credit Suisse and Deloitte [9]. For example, we find the percentage of female seats in Sweden to be 21.52%, whereas Credit Suisse reports 33.6% and Deloitte 31.7%. We believe that the difference is due to our larger sample size. The FT data base includes 465 Swedish companies compared to only 125 in Deloitte's data.

The top-ranked country in our data is Ukraine (22.6%). We have not found previous reports on female directors in Ukraine so that we cannot rule out that its top rank is owed to a relatively small sample size of only 19 Ukrainian companies. Another surprisingly highly ranked country is Thailand (19.5%). Although Deloitte estimates the percentage to be only 11.7%, it is plausible that the true number is higher because Thailand is among the countries with the largest proportion (37%) of women in senior management [17]. Near the bottom of the ranking is Japan (1.2%) where our number is even lower than previous estimates (Credit Suisse: 3.5%, Deloitte: 4.1%).

Similar to the worldwide trend mentioned above, the proportion of female *directors* by country hardly differs from the proportion of female *seats* (i.e. the data shown in Fig. 3). We have inferred the country of a person as the most common country of her or his companies. Based on this assumption, we have calculated the countrywide proportion of female directors. In every country included in the FT database, it differs by less than 0.022% from the proportion of female seats so that the conclusions do not depend on whether we use female seats or female directors as the basis of our analysis.

Another measure for comparing female representation across countries is the proportion of companies with at least one woman on their boards. Worldwide, we find that 50.3% of companies have at least one director who FT identifies as female. Because FT does not include gender information for 30.1% of the directors (see Sect. 2), the true percentage of companies with women on their boards is likely to be higher. Lee et al. [24] estimate 73.5% for the smaller MSCI data base. The country rankings, shown in Fig. 4 (grey bars in the plot), are similar to those for the proportion of female directors by country in Fig. 3. Ukraine drops to the ninth position, but the Scandinavian countries and Thailand maintain their high rankings. Oman, Japan, Pakistan, and Qatar remain at the bottom.

It is instructive to compare the observed percentage of companies with women on their boards with the expected values from a simple probabilistic null model. We assume that the probability of a seat being held by a woman is equal to the observed fraction p of female seats in a given country. In the null model, we assume that the assignment of women to seats is independent of the gender of the other seats. Suppose the size of a board is s . For each of the s seats, we flip a biased coin which shows heads with probability p and tails with probability $1-p$. When the coin shows heads, the seat is, in this model, given to a woman. The probability that the company's board has at least one woman equals $1-(1-p)^s$. If the fraction of boards with s seats is f_s , then the expected fraction of boards with women is $\sum_s f_s [1 - (1-p)^s]$.

The alternative hypothesis is that companies tend to have a single token woman on their boards. In this hypothesis, a woman is only added when there is currently no other woman on the board [15, 35]. With exactly one woman, the board satisfies a minimum criterion of diversity that reduces external pressure for greater female representation without seriously threatening the power of the “old-boys network”. If the token woman hypothesis is true, there would be a higher proportion of boards with exactly one female board member than in the null model.

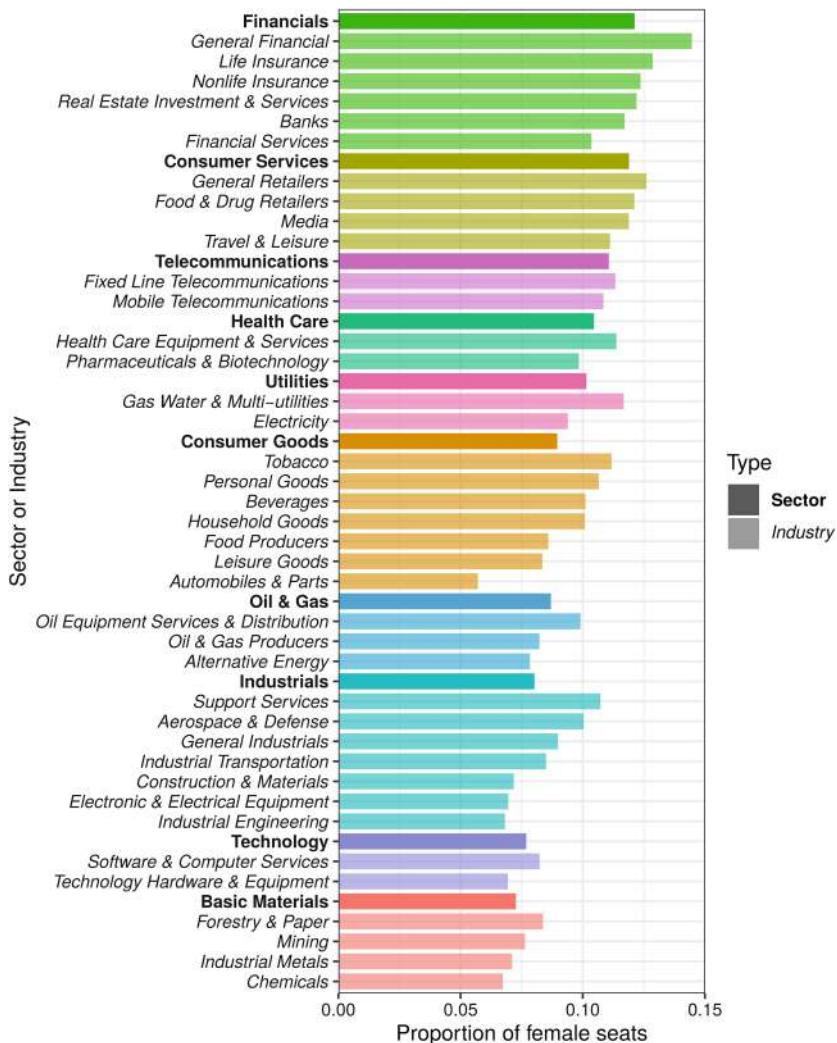


Fig. 2. Proportion of female seats by sector (darker colour) and industry (i.e. subsector, lighter colour).

We calculated the null model's expectation value for the global data and the predicted proportion of single-woman boards for each country. Worldwide, we find that the prediction is higher than the observation (0.604 vs. 0.507). On the level of individual countries, Slovenia is the only case where the prediction is lower than the observation, but even there the predicted proportion is only lower by 0.0013. In all other countries, there are fewer single-woman boards than predicted by the null model (Fig. 4). In some cases (e.g. Iceland or the United States) the difference is substantial. This implies that women are generally more

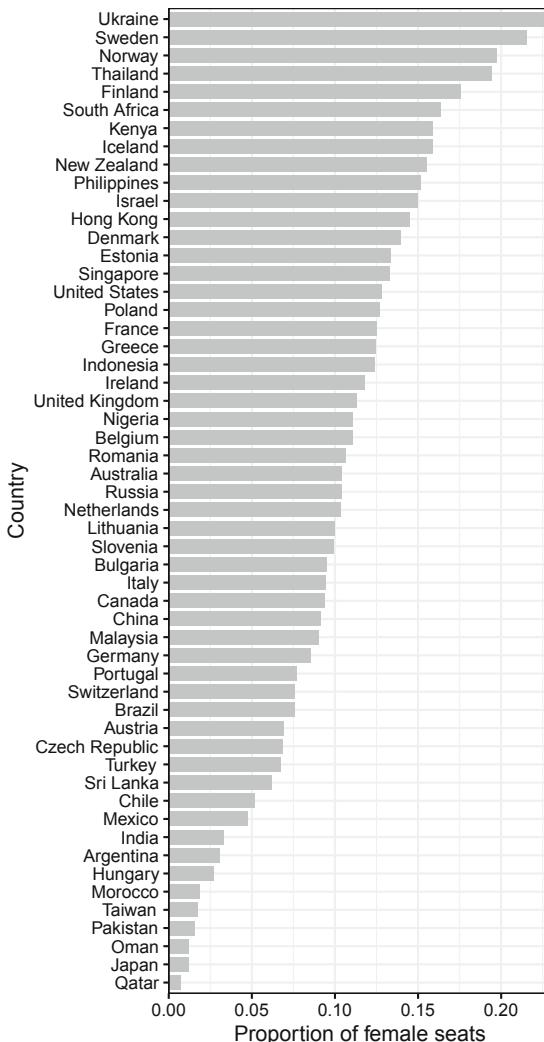


Fig. 3. Proportion of female seats by country.

clustered than expected if they were distributed randomly, contradicting the token woman hypothesis.

Although we find no evidence that women are recruited as tokens, their number has increased in recent years. As a consequence, women directors are, on average, younger than their male colleagues. In Fig. 5, we plot the age distribution by gender for all directors for whom the FT database contains information about both age and gender. The figure makes it clear that the distributions are centered at a different age: the mean is 50.8 years for women and 55.1 for men.

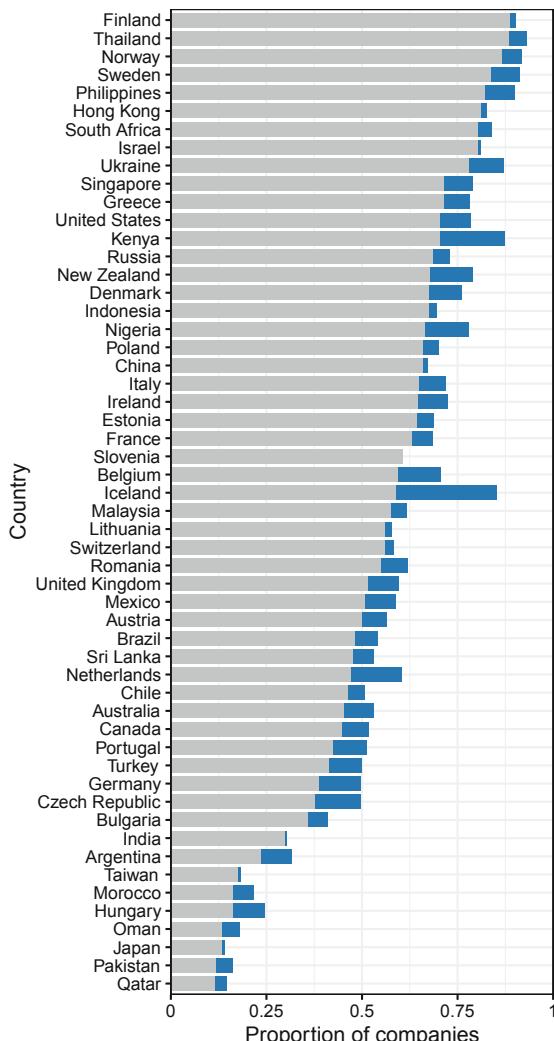


Fig. 4. Observed and predicted proportions of companies with at least one woman on their boards by country. The observed proportion is shown in grey. The predicted proportion is the combination of the grey and the blue bar. The prediction is calculated under the assumption that seats are taken by both genders independently given the observed proportion of female seats in each country (see text). The prediction is higher than the observed proportion in all cases except Slovenia, where the prediction is only slightly lower.

The solid curves in Fig. 5 show that the age distribution is approximately normal for both genders. Strictly speaking, neither the female nor the male distribution passes a χ^2 -test for normality (p -values $< 10^{-9}$) because they are slightly skewed towards higher age. However, the deviations from normality are

sufficiently small to justify using Welch's two-sample t -test. The null hypothesis of an equal mean for men and women is strongly rejected (p -value $< 10^{-15}$). Our result is consistent with earlier studies of French [24] and Singaporean data [40] where female directors were found to be on average 5–10 years younger.

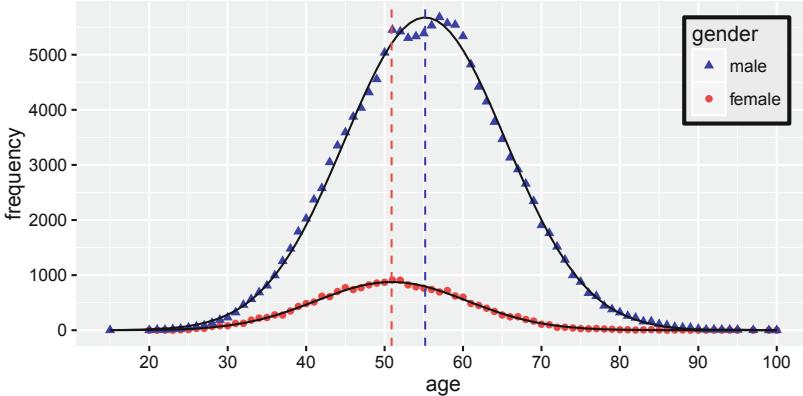


Fig. 5. The age distribution of directors by gender. Both distributions are well approximated by Gaussian functions (black solid curves), but with different means (dashed lines): the mean age of male directors is 55.1 years, that of female directors 50.8 years. The FT data contain information about the age of more than 200 000 directors. Note that 164 directors had their age listed as 1 year. We have removed these data points from our analysis.

4 The Position of Women in the Network

While the attributes discussed in the previous section already give us some insight into gender differences, we can only truly assess the role of women when considering their positions in the network. As we explained in the introduction, our data can be viewed as a bipartite network where edges run between directors and boards (Fig. 1a). In this network, there are 2 809 623 edges connecting 321 967 directors to 35 927 boards. The average board size (of those available) is 11.02. The mean size of a board without women is 8.88, whereas boards with at least one woman have on average 13.10 seats. Some care needs to be taken when interpreting these numbers. Even in our earlier null model, where we assumed that seats are independently taken by men and women, the mean size of a board with a woman is larger than the mean size without a woman. The reason is that, in this model, the probability of at least one woman on a board of size s is $1 - (1 - p)^s$ and thus increases with s . The mean board size conditioned on the presence of at least one woman is

$$\mu_{\text{null}} \equiv E[\text{board size} \mid \text{woman}] = \frac{\sum_{s=1}^{\infty} s f_s [1 - (1 - p)^s]}{\sum_{s=1}^{\infty} f_s [1 - (1 - p)^s]},$$

Table 2. Summary statistics of the network with all nodes (i.e. nodes identified as male, female, and those with missing gender information) and all edges. Larger values are highlighted in bold. “Like degree” is the degree between nodes of the same gender. We calculate the closeness centrality with the harmonic mean formula by Newman [28] to handle disconnected components.

Nodes	All	Male	Female
% of all	100	60.5	9.43
% in the largest component	74.7	74.8	79.4
Maximum degree	1040	1016	1030
Average degree	17.45	16.68	18.74
Average “like degree ”		11.21	2.94
Average degree in largest component	19.97	19.05	20.98
Maximum betweenness centrality	9.864×10^8	9.864×10^8	4.755×10^8
Average betweenness centrality	6.515×10^5	6.556×10^5	8.519×10^5
Average betweenness centrality in largest component	8.709×10^5	8.753×10^5	10.72×10^5
Maximum closeness centrality	0.143	0.143	0.143
Average closeness centrality	0.071	0.072	0.078
Average closeness centrality in largest component	0.096	0.099	0.097
Maximum clustering coefficient	1	1	1
Average clustering coefficient	0.939	0.935	0.936
Average clustering coefficient in largest component	0.9289	0.9243	0.9272

where f_s is, as before, the fraction of boards with s seats. We find $\mu_{\text{null}} = 12.97$, comparable to the observed value 13.10, but statistically significantly smaller (p -value $< 10^{-8}$). This result confirms previous observations that larger boards tend to have a higher probability of recruiting women [4, 5, 29].

Larger boards tend to be in the largest component of the bipartite network. In the one-mode projection that only contains the directors as nodes (Fig. 1b), the largest component consists of 74.7% of the nodes and 85.5% of the edges. Given that women are more likely to be on larger boards, it is not surprising that the proportion of women in the largest component (79.4%) exceeds the fraction of nodes belonging to that component. We confirm with the χ^2 -test proposed by Hawarden and Marsland [19] that the proportion of women in the largest component is significantly higher than that of men (p -value $< 10^{-15}$). We therefore agree with their previous result that, although women are a minority, they are not marginalised by being confined to unconnected, and hence less influential, components.

In terms of degree and betweenness centrality statistics, women are doing marginally better than men (Table 2). The distributions of degree and betweenness centrality by gender are not normal but instead seem to follow power laws. We normalise them by log-transforming the data and restricting our sample to the largest component and nodes with the parameter of interest >0 . The two-sample t -test for degree concludes that the marginal difference between men and women is statistically significant (p -value < 0.0001). The difference in the betweenness centrality is not statistically significant at a significance level of 0.05 (p -value 0.068).

5 Conclusion

In this paper we have analysed a new dataset which allows us to better understand interlocking directorates. In particular, it has allowed us to show the differences of female representation by country and industry. Overall, there are still many fewer women than men on boards, but our analysis contradicts the token woman hypothesis whereby companies recruit exactly one woman to escape accusations of discrimination. A limitation of our dataset is that it only indicates presence or absence of a link, but not its strength, which has been hypothesised to depend on gender [21, 22]. Our binary data, however, do not show evidence that women are marginalised.

Acknowledgements. We would like to thank Adrian Roellin for introducing us to the study of interlocks and to the Financial Times Equities database. M. T. G. was supported by the Singapore Ministry of Education and a Yale-NUS College start-up grant (R-607-263-043-121).

References

1. Adams, R.B.: Women on boards: the superheroes of tomorrow? *Leadersh. Q.* **27**(3), 371–386 (2016)
2. Battiston, S., Catanzaro, M.: Statistical properties of corporate board and director networks. *Eur. Phys. J. B* **38**(2), 345–352 (2004)
3. Burgess, Z., Tharenou, P.: Women board directors: characteristics of the few. *J. Bus. Ethics* **37**(1), 39–49 (2002)
4. Burke, R.J.: Company size, board size and numbers of women corporate directors. In: Burke, R.J., Mattis, M.C. (eds.) *Women on Corporate Boards of Directors*, pp. 157–167. Springer, Netherlands (2000)
5. Carter, D.A., Simkins, B.J., Simpson, W.G.: Corporate governance, board diversity, and firm value. *Financ. Rev.* **38**(1), 33–53 (2003)
6. Dawson, J., Kersley, R., Natella, S.: The CS Gender 3000: the reward for change. Technical report, Credit Suisse Research Institute (2016)
7. Dawson, J., Kersley, R., Vair, B., Preto, M.: Overboarding in the US. Technical report, Credit Suisse ESG Research (2016)
8. Delis, M.D., Gaganis, C., Hasan, I., Pasiouras, F.: The effect of board directors from countries with different genetic diversity levels on corporate performance. *Manag. Sci.* **63**(1), 231–249 (2016)

9. Deloitte: Women in the boardroom: a global perspective. Technical report, Deloitte Global Center for Corporate Governance (2017)
10. Dezső, C.L., Ross, D.G., Uribe, J.: Is there an implicit quota on women in top management? A large-sample statistical analysis. *Strateg. Manag. J.* **37**(1), 98–115 (2016)
11. d’Hoop-Azar, A., Martens, K., Papolis, P., Sancho, E.: Gender parity on boards around the world. Harvard Law School Forum on Corporate Governance and Financial Regulation (2017). <https://corpgov.law.harvard.edu/2017/01/05/gender-parity-on-boards-around-the-world/#more-76896>. Accessed 21 May 2017
12. Eagly, A.H.: When passionate advocates meet research on diversity, does the honest broker stand a chance? *J. Soc. Issues* **72**(1), 199–222 (2016)
13. Elango, B.: When do women reach the top spot? A multilevel study of female CEOs in emerging markets. *Management Decision* (2018, in press)
14. European Commission: Gender balance on corporate boards: Europe is cracking the glass ceiling (2015). http://ec.europa.eu/justice/gender-equality/files/womenonboards/factsheet_women_on_boards_web_2015-10_en.pdf. Accessed 23 May 2017
15. Farrell, K.A., Hersch, P.L.: Additions to corporate boards: the effect of gender. *J. Corp. Financ.* **11**(1–2), 85–106 (2005)
16. Gabaldon, P., De Anca, C., de Mateos Cabo, R., Gimeno, R.: Searching for women on boards: an analysis from the supply and demand perspective. *Corp. Gov.* **24**(3), 371–385 (2016)
17. Grant Thornton: Women in business: turning promise into practice (2016). https://www.grantthornton.global/globalassets/wib_turning_promise_into_practice.pdf. Accessed 20 May 2017
18. Hawarden, R.J.: Women on boards of directors: the origin and structure of gendered small-world and scale-free director glass networks. Ph.D. thesis, Massey University, Palmerston North (2010)
19. Hawarden, R.J., Marsland, S.: Locating women board members in gendered director networks. *Gend. Manag.* **26**(8), 532–549 (2011)
20. Hillman, A.J., Shropshire, C., Cannella, A.A.: Organizational predictors of women on corporate boards. *Acad. Manag. J.* **50**(4), 941–952 (2007)
21. Ibarra, H.: Homophily and differential returns: sex differences in network structure and access in an advertising firm. *Adm. Sci. Q.* **37**(3), 422–447 (1992)
22. Ibarra, H.: Personal networks of women and minorities in management: a conceptual framework. *Acad. Manag. Rev.* **18**(1), 56–87 (1993)
23. Izraeli, D.: The paradox of affirmative action for women directors in Israel. In: Burke, R.J., Mattis, M.C. (eds.) *Women on Corporate Boards of Directors: International Challenges and Opportunities*, pp. 75–96. Springer, Netherlands (2000)
24. Lee, L.E., Marshall, R., Rallis, D., Moscardi, M.: Women on boards (2015). <https://www.msci.com/documents/10199/04b6f646-d638-4878-9c61-4eb91748a82b>. Accessed 26 July 2017
25. Levine, J.H., Roy, W.S.: A study of interlocking directorates: vital concepts of organization. In: Holland, P.W., Leinhardt, S. (eds.) *Perspectives on Social Network Research*, pp. 349–378. Academic Press, New York (1979)
26. Martínez, A.C.: Social network analysis and the illusion of gender neutral organisations. Master’s thesis, Universitat Politècnica de Catalunya (2012)
27. McGregor, J.: The number of women CEOs in the fortune 500 is at an all-time high - of 32. *The Washington Post*, 7 June 2017
28. Newman, M.: Networks: An Introduction. Oxford University Press Inc., New York (2010)

29. Nguyen, H., Faff, R.: Impact of board size and board diversity on firm value: Australian evidence. *Corp. Ownersh. Control* **4**(2), 24–32 (2006)
30. Perrault, E.: Why does board gender diversity matter and how do we get there? The role of shareholder activism in deinstitutionalizing old boys' networks. *J. Bus. Ethics* **128**(1), 149–165 (2015)
31. Post, C., Byron, K.: Women on boards and firm financial performance: a meta-analysis. *Acad. Manag. J.* **58**(5), 1546–1571 (2015)
32. Seierstad, C., Opsahl, T.: For the few not the many? The effects of affirmative action on presence, prominence, and social capital of women directors in Norway. *Scand. J. Manag.* **27**(1), 44–54 (2011)
33. Sojo, V.E., Wood, R.E., Wood, S.A., Wheeler, M.A.: Reporting requirements, targets, and quotas for women in leadership. *Leadersh. Q.* **27**(3), 519–536 (2016)
34. Sonquist, J.A., Koenig, T.: Interlocking directorates in the top U.S. corporations. *Insurg. Sociol.* **5**(3), 196–229 (1975)
35. Strydom, M., Yong, H.H.A.: The token woman. In: 25th Australasian Finance and Banking Conference (2012). <https://doi.org/10.2139/ssrn.2136737>
36. The World Bank: Labor force, female (% of total labor force) (2017). <http://data.worldbank.org/indicator/SL.TLF.TOTL.FE.ZS>. Accessed 20 May 2017
37. Thomson, P., Graham, J.: *A Woman's Place is in the Boardroom*. Palgrave Macmillan, Basingstoke (2005)
38. Thomson Reuters Corporation: Profiles and lists of directors of publicly traded companies (2016). <https://markets.ft.com/data/equities/results>. Accessed 17 Sept 2016
39. Westphal, J.D., Milton, L.P.: How experience and network ties affect the influence of demographic minorities on corporate boards. *Adm. Sci. Q.* **45**(2), 366–398 (2000)
40. Yacob, H., et al.: Women on boards: tackling the issue. Technical report, Singapore's Diversity Action Committee (2016)



Supplier Impersonation Fraud Detection in Business-To-Business Transaction Networks Using Self-Organizing Maps

Rémi Canillas^{1,2(✉)}, Omar Hasan², Laurent Sarrat¹, and Lionel Brunie²

¹ SiS-id, Lyon, France

{remi.canillas,laurent.sarrat}@sisnet.fr

² Liris, INSA Lyon, Lyon, France

{remi.canillas,omar.hasan,lionel.brunie}@insa-lyon.fr

Abstract. Supplier Impersonation Fraud (SIF) is a rising issue for Business to Business companies, as the use of remote and quick digital transactions has made the task of identifying fraudsters more difficult. In this paper, we propose data-driven fraud detection system whose goal is to provide an accurate estimation of transactions' legitimacy by using the knowledge contained in the network of transactions created by the interaction of a company with its supplier. We consider the real dataset collected by SIS-ID for this work.

We propose to use a graph-based approach to design an Anomaly Detection System (ADS) based on a Self-Organizing Map (SOM) allowing us to label a suspicious transaction as either legitimate or fraudulent based on its similarity with frequently occurring transactions. Experiments demonstrate that our approach identifies fraudulent transactions with high success in a real-life dataset.

Keywords: Fraud detection · Graph-based feature engineering · Financial networks · B2B network

1 Introduction

Lately, Supplier Impersonation Fraud (SIF) is on the rise, resulting in the loss of hundreds of thousands of Euros in 2018, and ranked 1st most frequent fraud affecting French companies in the latest survey about cyber-criminality conducted in 2019 by Euler Hermes and DFCG [4]. Supplier impersonation consists in a fraudster impersonating a member of a company providing goods and services to another in order to trigger a payment on an account controlled by the fraudster [1]. More and more companies are using digital tools to process, authorize, or even conduct transactions due to numerous advantages provided by digitalization such as the ability to conduct transactions all over the globe in a timely fashion. However, digital transactions make frauds against companies more effective, firstly due to the difficulty to formally identify and trust remote interlocutors that are sometimes geographically very distant from the company

headquarters, and secondly due to the increased speed of wired transactions, allowing money to be moved between accounts in a very short amount of time, thus hindering the process of recovering it after a fraud.

In this work, we present GraphSIF, a SIF detection system that uses a B2B transactions dataset to construct a network modeling the relationships between client and supplier companies in a B2B ecosystem, and describe how these relationships can be used to derive useful knowledge in order to assert the legitimacy of transactions.

We use the transaction network to create a time-evolving behavior sequence summing up the evolution of the graph through time. We then compare the new graph created by adding a suspicious transaction to the behavior sequence and investigate the potential discrepancy it introduces. If this discrepancy is low then the transaction is considered as legitimate, and if the discrepancy is high then the transaction is considered as likely fraudulent. In order to quantify this discrepancy, a Self-Organizing Map (SOM) is trained on the behavior sequence, and a clustering algorithm is used to quantify the similarity of the tested graph with the ones in the behavior sequence.

The contributions of this paper are the following: a graph-based feature engineering process relying on a bipartite graph constructed from transactions in a B2B context, a classification system that uses Self-Organizing Maps and K-means clustering to investigate the legitimacy of a new transaction, and a comprehensive evaluation of the proposed classification system using data from a real-life B2B ecosystem.

The remaining of the paper is structured as follows: we first present related work in the field of fraud detection, then describe in detail the feature engineering process used to compute the graph used by the SIF detection system. We then describe the classification system we use to label unknown transactions. Finally, we evaluate our SIF detection system.

2 Related Work

Due to the sensitivity of the data linked to supplier fraud detection for victim companies, we have not been able to find any publicly available research work directly related to SIF. However, we can find several systems designed for fraud detection that also use a network-based approach:

In [9] several network-based fraud detection use-cases are introduced, showing examples of successful use of a database graph to detect bank fraud, insurance fraud and e-commerce fraud. However, the authors focus on specific industrial examples without proposing a formalized evaluation of their solution. Akoglu, Tong and Koutra [3] propose a survey on anomaly detection using graphs, notably in the domain of telecommunication fraud detection. An approach close to our own is found in [2] where an egonet (1-step neighborhood graph) is used to derive features describing a node. However, this approach is applied to a static graph that do not evolve through time, contrary to our approach. The analysis of dynamic graphs through the use of windows, as it is the case in our work, is

akin to the ideas developed in [7] and [8] where the graphs are analyzed using a moving window and detecting anomalous connectivity variation [8] or edges p-value variation [7]. To the best of our knowledge there is no previous attempt at combining a window-based analysis as seen in [8] with the feature approach used in [3]. Van Vlasselaer et al. [10] proposes a approach somewhat similar to ours, using graphs to represent interaction between companies in order to detect social security frauds. However, their work focuses on the application of social network algorithms on a large graph of interconnected entities, whereas our work partition the large graph into smaller neighborhood graphs focused on the specific behavior of a single company. In addition, the system proposed in [10] bases itself on the propagation algorithm and Random Logistic Forests to perform their analysis, and do not make use of Self-Organizing Maps. Furthermore, most of these works uses supervised algorithms as they rely on the existence of known legitimate and fraudulent neighboring nodes to conduct their analysis. Our work proposes an unsupervised approach that does not take into account the legitimacy of neighboring nodes to propose a label, but instead uses the topology of a specific ego network.

To the best of our knowledge, our work is the first to use a graph-based approach paired with Self-Organizing Maps in order to address the issue of SIF detection.

3 Problem Description

Let's define two companies C and S that have previously exchanged N historical transactions $\{t_1, t_2, \dots, t_N\}$, all performed on the same account a_l in a sequential manner. A fraudster F wants to attempt a fraud by impersonating S and trigger a payment from C to the account a_f . Let's assume that F simply impersonates an executive in S by hacking into his mail account and advising C that all future payments on outstanding invoice should be wired to a_f . With no means to verify the executive's identity, C complies to the request. This triggers all future transactions t_{N+1}, t_{N+2}, \dots to be performed on a_f instead of a_l . We will use this example to illustrate how the proposed detection system works.

4 Graph-Based Feature Engineering

In order to construct a model of the behavior of a client company using graphs, we first use a sequencing algorithm that aggregates the ordered transactions in bounded windows defining the company's payment behavior in the time frame defined by the window. We then aggregate these transactions into a graph, and we vectorize them by computing the number of patterns occurring in the graph. These vectors create a sequence that we use to train a Self-Organizing Map, resulting in a topography in which similar graphs are regrouped based on the patterns they share. We finally create a test graph containing the transaction to be labeled combined with the most recent transactions, and compute a legitimacy score by quantifying the similarity of the graph created in regards to the ones used to train the SOM. This anomaly score is then discretized to be

turned into a label indicating if the suspicious transaction appears fraudulent or legitimate.

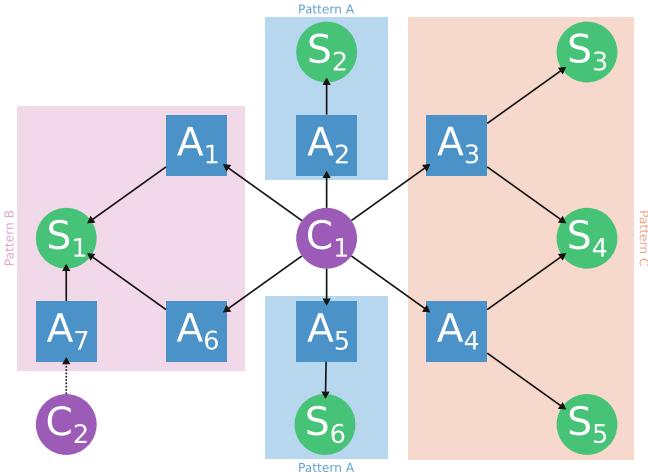


Fig. 1. Transaction graph of client C1.

4.1 Transaction Graphs

In order to use the contextual environment of a transaction in order to assert its legitimacy, we use the relational aspect of the transactions to create a **Transaction Graph** aggregating all the transactions involving a specific agent (any entity involved in a transaction) in a specific period of time. Figure 1 presents an example of a such a graph. This graph is created by first getting all the transactions involving the elements of a specific transaction. Then, a node is created for each element found in this list of transaction. Then, two edges are created between the elements involved in the transaction: one directed edge from the client to the account, and one directed edge from the account to the supplier. This transaction graph is a directed bipartite graph with two types of nodes: companies (either client or supplier), or accounts. The edges of the transaction graph are binary weighted: either an edge exists between a company and an account, and its weight is 1, or no edge exists. This model is very simple and aims to focus the analysis on the relationship shared between entity rather than the semantics of such relationships.

4.2 Featurization

The transaction graph, while providing a useful tool for visualization, is unstructured, meaning that it does not represent as a feature vector, and thus is not usable directly to perform a classification. In order to do so, the need to transform the transaction graph into structured data arises. We propose the following

Table 1. Examples of featurized transactions

Transaction ID	Pattern A	Pattern B	Pattern C	Pattern D
T1	2	1	1	0
T2	3	3	0	1
T3	1	0	0	0

process in order to do so: First, we isolate **Transaction Patterns** from the transaction graph: we define a transaction pattern as the list of connected sub-graphs obtained when the targeted transaction’s client company is removed from the graph. This allows us to create a characterization of the context in which the transaction occurs. We then compute this list of sub-graphs as an histogram where the occurrence of each unique pattern is recorded. We use this histogram as our feature vector in our analysis. Table 1 shows a simple example with three transactions: T1, presented in Fig. 1, T2 and T3. For the rest of the paper, we will use a hash of the canonical code [6] of the graph as labels for transaction patterns: this allows us to provide a unique identifier for each pattern.

4.3 Behavior Sequence

Using the transaction graph comes with an important drawback: the temporal order in which the transactions are processed is lost, hence hindering the dynamic analysis of our target. Moreover, with an increasing number of transactions, it becomes harder to determine the impact of a single transaction on the graph. We overcome these issues by creating **Behavior Sequences**: first, the transactions associated with the agent are sorted based on the date in which they occur, and then a window of specified size is created to group transactions in specific sets. Then, instead of a single, all-containing transaction graph, several smaller graphs are created. These graphs create a sequence that illustrate the temporal behavior of the entity.

4.4 Test Graph

In order to quantify the legitimacy of a specific transaction, a **Test Graph** is created: the transaction is added to the latest window of the behavior sequence, and a transaction graph is derived from this window, thus containing the test transaction. It is important to note that by creating the test graph we shift the focus of our Fraud Detection System from a single transaction to a single graph constructed by aggregating all the transactions from a window.

5 Self-Organizing Map Analysis

One of the issue with comparing a test graph with a behavior sequence is that the number of patterns found in the behavior sequence is highly variable, as it

depends on the window size, the number of transactions, and the complexity of the relationships between the client and its suppliers. In order to provide a consistent analysis of the behavior sequence, we first compute a Self-Organizing map that will regroup similar transaction graphs in the same area of the map. We then use K-Means to regroup together these similar graphs. This concludes the training phase. Then in the testing phase we assign a cluster to the test graph, and calculate the distance on the SOM from the node activated by the test graph to the centroid of the cluster assigned to the test graph.

5.1 Training Phase

In order to train the SOMs used in our system, we use the feature vectors found in the behavior graph of the selected client. More specifically, we feed them to the SOM where all the weights of the nodes have been randomly generated. Each feature vector will activate a single node, that will then update its weights to match the value of the histogram, while its topological neighbors will also be updated (though not as much as the first node). Once all the histograms in the behavior sequence have been fed to the SOM, we then proceed to score the test graph.

5.2 Testing Algorithm

While SOMs provide a way to project a high dimensional feature vector in a topological plane, and regroup similar feature vectors close to each other, they do not provide by themselves a way to formalize the dissimilarity between two given feature vectors. In order to do so, we created an algorithm that aims to compute an anomaly score based on the trained SOM: the K-Nearest Neighbors Distance Algorithm. This algorithm first uses the K-Means clustering algorithm to create clusters from the graphs contained in the behavior sequence. The distance between each node activated by an histogram, and the centroid of the cluster assigned to it, is computed. A cluster is then attributed to the test graph, and the distance between the node activated by the centroid of the cluster assigned to the test graph, and the node activated by the test graph is calculated. This distance is then used to compute the z-score of the test transaction. The intuition behind this algorithm is that the further away from the center of the cluster, the higher the z-score, and thus the more dissimilar to the members of the cluster the test graph is. In order to give a legitimacy label, the z-score is compared with two user-defined thresholds δ_1 and δ_2 representing the severity of the classification system.

6 Datasets

In this section we provide details about the datasets used to train and evaluate our SIF detection system: the History dataset used to train the model, and the Audit dataset used to test the performance of our system.

6.1 History Dataset

In order to build our data-driven fraud detection system, we use a set of B2B transactions provided by the SiS-id platform¹, aggregating the transactions carried out between July 2016 and July 2019 between 5,921 companies. We dubbed this dataset “History”. Table 2 sums up the features available from this dataset. Depending on the transactions’ sources, more data can be available, such as the amount of the transaction or details about the goods or services paid by the transaction, but these pieces of information are not available for every record. This dataset contains more than 2 million transactions.

Table 2. Features describing a transaction between two companies.

Feature	Type	Description
Client	Nominal (ID)	Identification number of the client issuing the transaction
Supplier	Nominal (ID)	Identification number of the supplier receiving the transaction
Account	Nominal (ID)	Identification number of the bank account to which the money is transferred
Date	Continuous (Timestamp)	Timestamp indicating the date when the transaction took place

6.2 Audit Dataset

A second set of transactions is provided by SiS-id. It consists of the list of transactions that were analyzed using the expert system of the company in the past 2 years (July 2017–July 2019). The dataset, called the “Audit” dataset, is composed of 108,102 suspicious transactions submitted by 171 unique client companies. These transactions are attributed a legitimacy label by SiS-id’s fraud detection platform that we use as ground truth for our analysis.

SiS-Id’s expert system is composed of a rule-based engine that compare the transaction with fraudulent or legitimate known cases defined by investigation experts. If the transaction matches one of these cases, then it is assigned the corresponding label (“low” if fraudulent, “high” if legitimate), while if the transaction does not match any cases a “medium” label is assigned to it.

7 Classification Process

In order to evaluate the legitimacy of a transaction, we run GraphSIF with a set of pre-defined window sizes (ranging from 5 transactions to 200 transactions)

¹ <https://my.sis-id.com>.

and $\delta_1 = 0.50$ and $\delta_2 = 0.9$, using 251 test transactions previously labeled by SiS-id's expert system, and trained on 16168 unlabeled historical transactions. Results show that the variation in the distance was not high enough to accurately determine if a transaction graph was significantly different from the ones in the behavior sequence. In order to circumvent this issue, we aggregated the occurrences of each class label for each of the test transaction and reported it in Fig. 2 (upper part, number of occurrences normalized for scale). In this figure we notice more clearly the different labels given by the analysis of the different windows. We define three weights w_h , w_m and w_l for each of the legitimacy labels, in order to “reward” legitimate transaction and “penalize” abnormal ones. Then the weighted average $w = \frac{w_h o(h) + w_m o(m) + w_l o(l)}{w_h + w_m + w_l}$ is computed, where $o(\cdot)$ is the number of occurrences of a label for a specific test transaction. We plot this score (normalized) in the lower part of Fig. 2, computed with $w_h = 20$, $w_m = 5$ and $w_l = -5$.

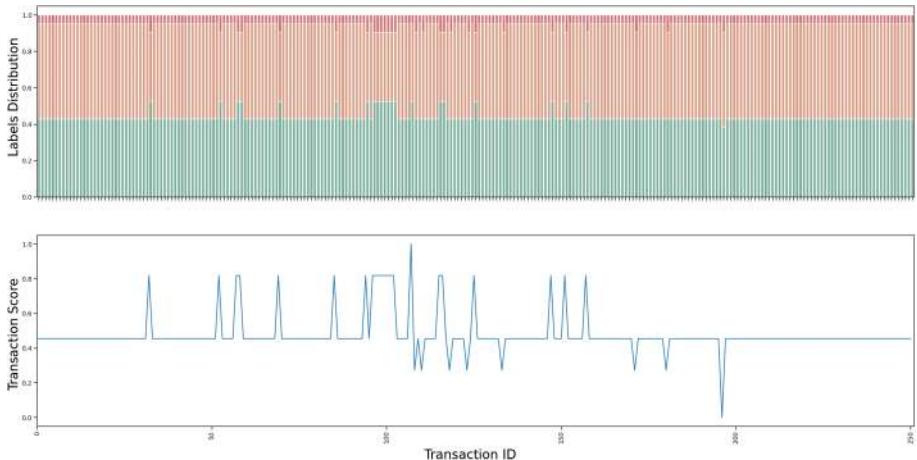


Fig. 2. Score aggregation - Each of the 251 test transactions (X axis) is classified 40 times with window size ranging from 5 to 200 with thresholds $\delta_1 = 0.50$ and $\delta_2 = 0.9$. The uppermost figure shows the distribution of labels for each transaction. The lowermost figure shows the score computed from this distribution with weights $w_h = 20$, $w_m = 5$ and $w_l = -5$.

Finally, the score for each transaction is then compared with δ_1 and δ_2 , as shown in Fig. 3, where the green, orange, and red zones represents the area where transactions are labeled with the “high”, “medium” and “low” legitimacy label respectively. This allows us to label 6 transactions with the “high” legitimacy label, 15 transactions with the “medium” legitimacy label, and 230 transactions with the “low” legitimacy label.

8 Problem Resolution Using GraphSIF

We will now describe how GraphSIF applies to the example defined in Sect. 3, assuming only a single window size w is used. First, a behavior sequence of size $\frac{N}{w}$ is created from the N transactions between C and S . As the N historical transactions are identical, identical histograms will be created. Then, each transaction from t_{N+1} is added to the $\frac{N}{w}^{th}$ behavior sequence's graph, and as a new account (a_f) is added, a new graph will be created. Thus, the feature vector representing this graph will be significantly different from the previous one, and be labeled as fraudulent by GraphSIF.

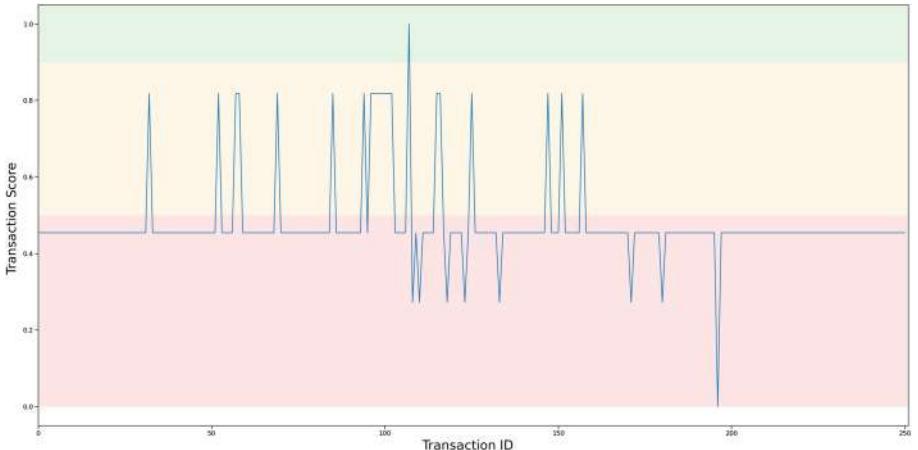


Fig. 3. Score thresholding with $\delta_1 = 0.50$ and $\delta_2 = 0.90$ - Transactions with a score higher than δ_1 are labeled with a “high” legitimacy label. A score between δ_1 and δ_2 is given a “medium” legitimacy label. A score lower than δ_2 is given a “low” legitimacy label.

9 Experimental Results

In this section, we compare the results obtained in the previous section with the ones obtained with SiS-id’s rule-based fraud detection system by first providing an overview of the differences and then focusing on three key metrics for fraud detection: accuracy, efficiency and maintainability. Unfortunately, due to the sensitive nature of the business data, the trained model is not publicly available. However a request might be submitted to the authors in order to obtain an anonymized sample of the dataset.

We first present three Venn diagrams showing the overlap of labels given by the rule engine (Rules) and graph-based detection system (Graph). First, we notice that there are a lot of transactions that has not been given the same label by both of the detection systems. Most of these difference come from the

fact that numerous “high” and “medium” legitimacy labels given by the rule-based engine have been labeled with “low” legitimacy labels by the graph-based analysis. Only 0.01% of the transactions labeled with the “high” legitimacy label by the rule engine are shared with the graph analysis, as shown in Fig. 4a, while the “medium” label is only shared in 0.05% of the transactions (Fig. 4b). However, 83.6% of the “low” legitimacy transaction are shared by the two systems, as shown in Fig. 4c.

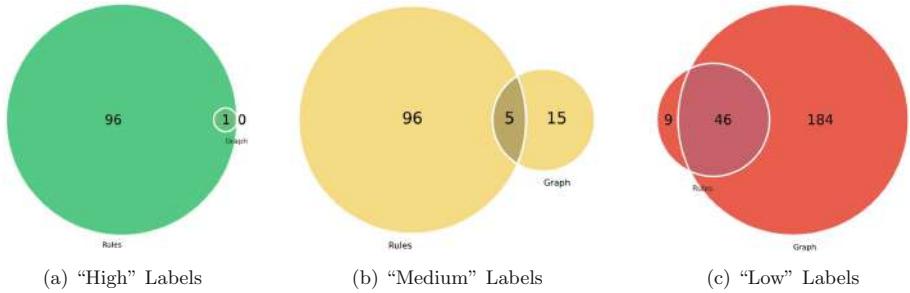


Fig. 4. Overlap of rule-based (“Rules”) and graph-based (“Graph”) analysis classification results for each label.

9.1 Precision

In order to evaluate GraphSIF’s ability to detect a fraudulent transaction, we use a definition of precision slightly different from the one traditionally found in Machine Learning. More precisely, we define the accuracy $P = \frac{d_f}{N_f}$ of a fraud detection system as the number of fraud detected by the system d_f divided by the total number of fraud N_f found in the evaluation dataset. In this work we use the number of transaction labeled with a “low” legitimacy score by SiS-id’s expert system as N_f .

We use Fig. 4c to compute the precision. We have $N_f = 9 + 46 = 55$ the number of transactions given a “low” legitimacy label by the rule engine, and $d_f = 46$ the number of transactions also given the “low” legitimacy label by the graph-based analysis, thus giving $P = \frac{46}{55} = 0.836$.

While our approach leads to a high rate of fraud detection, the number of false positives (legitimate transactions labeled as fraud) is also high. False positives might delay payment and seed distrust between business partners. The elaboration of a cost-sensitive metric is thus an interesting research direction.

Figure 4 shows that our results differs from the rule engine for the “high” and “medium” labels. This might be explained by the fact that the data investigated is very different: while the expert system relies on expert knowledge about the transaction, GraphSIF only considers the relations between companies and accounts in order to perform its analysis.

9.2 Efficiency

The efficiency of a SIF detection system is the time taken by the system to assign a label to a suspicious transaction. This metric is of utmost importance, because the quicker a fraud attempt is detected, the less a system is vulnerable to the fraud. On average, the training time took roughly 33 s using a laptop with 7.4 GB RAM running Ubuntu 18.04, while testing the 251 suspicious transactions took 18 s. Thus the time taken by the graph-based analysis to label one transaction is 71.7 ms for a single window. However, as the final result is derived from the analysis of all the windows, this score must be multiplied by the number of windows size, i.e $100/5 = 20$ in our case, so $E = 20 * 71.7 = 1434$ ms. Thus, there exists a trade-off between efficiency and accuracy, as adding more windows to the analysis will lead to more fine-grained results, but at the cost of increasing linearly the processing time and thus the efficiency. However, it is very likely that GraphSIF computation for each windows size can be performed in a parallel fashion, thus reducing the computation time for testing a transaction.

9.3 Maintainability

The maintainability of SIF detection system represents the time and effort it needs to adapt to a new situation. In the context of GraphSIF, this time roughly corresponds to the time it takes to update the data model when more historical transactions are added to the dataset. Experiments show that the training time for the graph-based system is 33 s on average for each window size. It is slightly higher (35 s) for small window size as more graphs are created from the dataset. This training time corresponds to the maintainability time, and it relatively high, as several steps need to be conducted before the model is complete.

10 Conclusion

In this paper, we introduce GraphSIF, a novel feature-engineering process that creates a feature vector based on the relationship between a client company and the accounts it used to pay its supplier company, providing a new tool to describe the underlying transaction mechanism involved in their interaction.

Several recent papers such as [11] [5] and [7] propose an human interpretation of the patterns uncovered by their approach and how they might suggest illegal behavior. The focus of our work is to emphasize on the variation of behavior, instead of the behavior itself. However the relation between the uncovered patterns and fraud attempts is currently under investigation.

In conclusion, we used the temporal information contained in the transactions of the History dataset to create a behavior sequence composed of the transactions emitted by a client aggregated in several bounded time windows. We showed how to use this behavior sequence to create a data model based on Self-Organizing maps representing the behavior of a client company through time. We then used this data model to infer the legitimacy of new transactions using the K-means

clustering algorithm, along with an aggregation algorithm allowing us to combine the results obtained for different window size in a comprehensive score.

We presented the result of our classification system first on a single client to investigate its performance locally, and showed that the results differs widely from the rule engine. However, GraphSIF shows a very good accuracy locally, and quick training and testing time, and thus can be used in a complementary fashion with the expert-based analysis.

References

1. AIG. Impersonation Fraud Claims Scenarios (2019). <https://www.aig.com/content/dam/aig/america-canada/us/documents/business/management-liability/impersonation-fraud-claims-scenarios-brochure.pdf>. Accessed 06 Nov 2019
2. Akoglu, L., McGlohon, M., Faloutsos, C.: Oddball: spotting anomalies in weighted graphs. In: Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining-Volume Part II, pp. 410–421. Springer-Verlag (2010)
3. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. *Data Min. Knowl. Disc.* **29**(3), 626–688 (2015)
4. Euler-Hermes DFCG. Barometre Euler Hermes-DFCG 2019 (2019). <https://www.eulerhermes.fr/actualites/etude-fraude-2019.html>. Accessed 06 Nov 2019
5. Luc, A.D., Daryl, K.G., Edward, J.T.: Surgical images: Soft tissue Calcinosis cutis (2007)
6. Stephen, G.H., Radcliffe, A.J.: McKay’s canonical graph labeling algorithm. 0000, 99–111 (2012)
7. Mongiovì, M., Bogdanov, P., Ranca, R., Evangelos, E.P., Faloutsos, C., Ambuj, K.S.: NetSpot: spotting significant anomalous regions on dynamic networks. In: Proceedings of the 2013 SIAM International Conference on Data Mining, SDM 2013, pp. 28–36 (2013)
8. Carey, E.P., John, M.C., David, J.M.: Youngser park scan statistics on enron graphs. *Comput. Math. Organ. Theory* **11**(3), 229–247 (2005)
9. Sadowksi, G., Rathle, P.: Fraud detection: discovering connections with graph databases the #1 database for connected data fraud detection: discovering connections using graph databases, January 2017
10. Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B.: GOTCHA! network-based fraud detection for social security fraud. *Manage. Sci.* **63**(9), 3090–3110 (2016)
11. Wachs, J., Kertész, J.: A network approach to cartel detection in public auction markets. *Sci. Rep.* **9**(1), 1–18 (2019)



Network Shapley-Shubik Power Index: Measuring Indirect Influence in Shareholding Networks

Takayuki Mizuno¹, Shohei Doi¹(✉), and Shuhei Kurizaki²

¹ National Institute of Informatics, Tokyo, Japan
sdoi@nii.ac.jp

² Faculty of Political Science and Economics, Waseda University, Tokyo, Japan

Abstract. Extending the Shapley-Shubik power index to networks, we propose a new measure and numerical method to calculate the indirect influence of investors on companies: *Network power index* (NPI). While the original index, reflecting the characteristics of majority vote in a shareholders meeting, measures the direct voting power of a shareholder, NPI captures not only an investor's direct influence over a company but also indirect influence over this company's subsidiary. Since NPI is often incalculable in a large network, we present a new method to numerically compute NPI: *label propagation*. Applying this method to the global shareholding networks in 2016, we find NPIs and raw vote shares dramatically diverge for some investors and this discrepancy suggests the difference in investment strategies between governments and private financial institutions.

Keywords: The Shapley-Shubik power index · Complex networks analysis · Indirect control · Shareholding networks

1 Introduction

To understand the architecture and the dynamics of shareholders networks, it is important to measure the influence of an investor over the decision-making of companies in networks. There are two obstacles to the achievement of this task: one is the characteristics of networks and one is the characteristics of shareholders meetings. First, investors may have multiple paths to influence companies' behavior through global complex networks of shareholdings. For example, stockholders have the influence over the decision-making process of the companies for which they own stocks, but also have the indirect influence on those companies' subsidiaries. Second, ownership and control of corporations has been distinguished in corporate governance [15]. While ownership of a company and reward to shareholders are proportional to the share of stocks they hold, control over a company is not, because decision-making in a shareholders meeting is usually (special) majority vote. Drawing on cooperative game theory, the Shapley-Shubik power index [14] is typically used to measure the power of entities in

collective decision-making bodies like legislature and shareholders meetings [5] (Note that the Banzhaf, or Penrose-Banzhaf power index, is another popular measurements of voting power [2, 6, 13]). However, the Shapley-Shubik power index cannot be simply applied to networks because multiple decision-making bodies are mutually inter-connected in a network. Researchers propose several concepts of indirect influence in networks [3], but they rely on a slightly strict assumption or do not measure one-to-one influence.

In this study, we propose the measurement and the calculation method of indirect influence of shareholders in networks incorporating the aspects of both network and majority vote. The remainder of this paper proceeds as follows. In Sect. 2, we illustrate the problem of indirect influence with an example and briefly review the existing approaches to indirect influence. In Sect. 3, we propose a network version of the Shapley-Shubik power index (*Network power index*: NPI) and our approximation method of this index based on the Monte-Carlo method. In Sect. 4, we apply the new index to global shareholding networks to demonstrate the implications and utilities of NPI. In Sect. 5, we conclude this paper.

2 Indirect Influence in Shareholdings Networks

We describe the problem of indirect influence with the illustrative example of a network of shareholders and companies depicted in Fig. 1. Suppose that shareholders B , C and D respectively own 30%, 30% and 40% of shareholdings of company A and each shareholder B and E has a half of stocks of company C . Let w_{ij} denote the vote share of shareholder i holds on j . The simplest way of measuring indirect influence of i on k is to use the product of share ratios, i.e., $w_{ij}w_{jk}$ [17]. We call this measurement *network share ratio* (NSR). For example, in the abovementioned network, company B 's NSR on A is $w_{BA}+w_{BC}\cdot w_{CA} = 0.3+0.5\cdot 0.3 = 0.45$. NSR reflects the structures of a network but dismiss the property of majority vote. If company B had 51% of shares of C ,

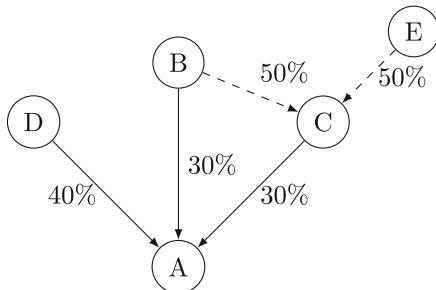


Fig. 1. Example of a network: each node is a company and arrow is shareholding relation with a proportion of shareholdings.

B could solely dominate the decision-making of C and, therefore, B and C could jointly control company A 's behavior. In this case, however, B 's NSR remains almost 0.45 although B completely controls two companies A and C .

The Shapley-Shubik power index is a game-theoretic approach to this non-linear transformation from vote share to the degree of power. To formally define this index, we introduce some notations. Suppose that there are n shareholders on company j and $q \in (0.5, 1]$ of total shares are necessary to pass a bill in a shareholders meeting. Let S denote a subset of shareholders, namely, *coalition*, and W represent a set of *winning coalitions*: the sum of vote shares of members of each winning coalition exceeds q . Note that in defining the Shapley-Shubik power index any coalitions are considered to be a permutation of shareholders. Then, the Shapley-Shubik power index is defined as

$$\phi_i \equiv \sum_{S \in W, S \setminus \{i\} \notin W} \frac{(s-1)!(n-s)!}{n!}. \quad (1)$$

We call shareholder i a *pivot* when i is necessary for S to be a winning coalition, i.e., $S \in W$ and $S \setminus \{i\} \notin W$. Then, the Shapley-Shubik power index, ϕ_i , can be interpreted as the probability that i is a pivot. Consider the Shapley-Shubik power index of B , C and D over A in Fig. 1. None of these three companies, B , C , and D , alone can form a winning coalition in A 's decision-making if decision-making requires 50% of shareholdings. Each needs to form a coalition to be decisive in A 's decision making, which in turn implies that each has the equal chance of being pivotal in forming a winning coalition. That is, the Shapley-Shubik power index for each of these three companies is $\frac{1}{3}$, even though each company has the varying amount of stocks.

This example highlights how the size of shares is inadequate in measuring a shareholder's influence on decision-making power, and how useful the Shapley-Shubik power index is for this purpose. We, in this study, extend this concept to networks by defining an individual *network power index* (NPI) as the probability of investor i being a direct or indirect pivot in the decision-making of company j through networks (for the purpose of comparison, we call the Shapley-Shubik power index an individual *direct power index* (DPI)). We assume that a pivot in the decision-making of each company are sequentially decided from investors, which make a decision independent from others, to downstream companies. That is, initially, a pivot of each company are stochastically determined from investors, then, companies act as if they were their pivot in decision-making of their subsidiaries and pivots of the subsidiaries are determined and so on. If there are companies sharing their pivot in a single decision-making, they vote jointly.

In the example in Fig. 1, because company C is controlled either by B or E with equal probability, NPIs for companies B and E over C is $\frac{1}{2}$ as is the case with their individual DPIs. This observation leads to the fact that, in company A 's decision-making, both companies B and E may each control C . In the former case, company B 's individual NPI on A is 1 because B and C jointly has 60% vote share in total while, in the latter case, B 's influence is just $\frac{1}{3}$ analogous to the previous example of individual DPIs. Thus, company B 's individual NPI on

A is given by $\frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{1}{3} = \frac{2}{3}$. Similarly, we obtain D and E 's individual NPIs as $\frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$. The result is intuitive in that company B has the most powerful influence on A and counterintuitive in that company E has the same power as D although E is connected with A only through C .

There are several concepts of indirect influence based on game theory [3]. Karos–Peters [8] and Mercik–Lobos [12] models define importance of not only investors but also companies in shareholding networks, while our objective is to measure of indirect influence of one investor on one company. Gambarelli–Owen [7] and Crama–Leruth [4, 10, 11] models measures individual indirect influence in a network but relies on a seemingly stringent assumption. They assume that investors vote for or against all bills of companies which investors potentially influence. For example in Fig. 1, company E is supposed to make a decision on behavior of A whether E control or not C . This assumption becomes unrealistic when it comes to global shareholding networks, which consists of millions of companies. In contrast, our approach models the chain of delegation from shareholders to executives like principal-agent model [9, 16] in that a company behave as an agent of a pivot.

3 Label Propagation

The definition of NPI, the probability of being pivotal in other entity's decision-making via networks, is straightforward, but its calculation is not as simple. We, therefore, propose a simulation-based approximation algorithm of NPI, called *label propagation*. The core algorithm is shown in Algorithm 1. The intuition behind this algorithm is as follows. In the first step, for each node j , shareholders i who has influence on j is randomly drawn and the vote shares are summed up sequentially. When the sum of votes reaches q , then i is recorded as a pivot on company j . After determining all pivots for all companies in the first step, the label for j (denoted by L_j^1) is replaced with i (since j now act as an agent of i). From the second step onward, before sampling pivots, the vote share for each shareholder is summed if their pivots are the same.

After T iterations, where $t \in \{0, \dots, T\}$, the empirical distribution of L_j^t is the frequency of the event that company j is under the control of another shareholders i . Let $\mathbf{I}(L_j^t = i)$ be a function which takes one if $L_j^t = i$ and zero otherwise. Then, we approximate the individual NPI for i on j as follows:

$$\hat{p}_{ij}^N \equiv \sum_{t=1}^T \frac{\mathbf{I}(L_j^t = i)}{T} \approx p_{ij}^N, \quad (2)$$

where p_{ij}^N and \hat{p}_{ij}^N are an individual NPI of i on j and its estimator. The approximated individual NPIs in our example of Fig. 1 with 10,000 iterations are $\hat{p}_{BC}^N = 0.4957 \approx \frac{1}{2}$, $\hat{p}_{DC}^N = 0.5043 \approx \frac{1}{2}$, $\hat{p}_{BA}^N = 0.661 \approx \frac{2}{3}$, $\hat{p}_{DA}^N = 0.1697 \approx \frac{1}{6}$ and $\hat{p}_{EA}^N = 0.1642 \approx \frac{1}{6}$, which are sufficiently close to theoretical values.

To avoid indeterminacy problems, we introduce two restrictions to the algorithm. First, to avoid the initial value dependency, for arbitrary $\bar{t} \ll T$, we drop

Algorithm 1. Calculate NPI

```

 $L_j^0 \leftarrow j$ 
 $N_j^0 \leftarrow \{i \in N \mid w_{ij} > 0\}$ 
 $w_{ij}^0 \leftarrow \bar{w}_{ij}$ 
for  $t$  in  $1 : T$  do
    for  $j$  in  $N$  do
         $w_{ij}^t \leftarrow \sum_{L_k^{t-1}=i} w_{kj}^{t-1}$ 
         $N_j^t \leftarrow \{i \in N \mid w_{ij}^t > 0\}$ 
         $n_j^t \leftarrow |N_j^t|$ 
         $S_0 \leftarrow 0$ 
        for  $k$  in  $1 : n_j^t$  do
             $i \sim \text{Multinom}(N_j^t, 1, \frac{1}{n_j^t})$ 
             $S_k \leftarrow S_{k-1} + w_{ij}^t$ 
            if  $S_k > q$  then
                 $L_j \leftarrow i$ 
                break
            else
                 $N_j^t \leftarrow N_j^t \setminus \{i\}$ 
                 $n_j^t \leftarrow n_j^t - 1$ 
            end if
        end for
    end for
     $L_j^t \leftarrow L_j$ 
end for
 $\hat{p}_{ij}^N(\bar{W}) \leftarrow \frac{\sum_{t=1}^T \mathbf{I}\{L_j^t=i\}}{T}$ 

```

the results for the first \bar{t} iterations from the calculation of NPIs in the last. Second, the labels for companies are stochastically restored to the initial one, L_j^0 , with the small probability, $\varepsilon \in (0, 1)$.

4 Analyzing the Influence Structure in the Global Shareholders Networks

We now analyze the structure of influence in the dataset of global shareholders network. Our data set contains the information on the shareholdings of 49 million corporations among 69 million shareholders (including firms and individuals alike) in 2016 obtained from *Bureau van Dijk's Orbis* database [1]. Note that the amount of sales of companies in this dataset aggregated for each country highly correlated with the GDP of them, which suggests the coverage of the dataset across countries are not biased.

Although our method measures an individual NPI of one investor on one company, the aggregated NPI for each investor, $p_i^N = \sum_j p_{ij}^N$, is also useful index to capture the structure of global shareholders networks (an aggregated NSR and DPI are defined similarly). Based on the probabilistic interpretation of the power index, an aggregated NPI of i means the expected value of the number

of companies under direct and/or indirect control of i . In this section, we focus on aggregated NPIs in order to evaluate the overall performance and utilities of NPIs, keeping the micro-level analysis by individual NPIs a future research.

4.1 Validation of Calculation Errors

We first evaluate the robustness of our method since we adopt a numerical approach that uses random numbers. Unfortunately, since we do not know true values of aggregated NPIs, it is impossible to compare estimated and true parameters. Instead, as a second best approach, we calculate aggregated NPIs for each entity at two separate iterations, namely T_1 and T_0 , and define the calculation error for entity i as

$$1 + \Delta_i(T_0, T_1) = \frac{\hat{p}_i^N(T_1)}{\hat{p}_i^N(T_0)}, \quad (3)$$

where $\hat{p}_i^N(T)$ is the aggregated NPI for i in the calculation with T iterations.

Figure 2 shows the result. To obtain the approximate distribution of the calculation errors shown in this figure, we calculate the aggregated NPI and its calculation errors twice: first with 20,000 iterations (i.e., $T_0 = 20000$) and, second with 2,000 (i.e., $T_1 = 2,000$). Then, we stratify the resulting samples into six groups according to the value of $\hat{p}_i^N(T_0)$: $[0, 0.1]$, $[0.1, 1]$, $[1, 10]$, $[10, 100]$, $[100, 1000]$ and $[1000, \infty)$. We obtain several quantiles of calculation errors for each group. When it comes to influential investors whose aggregated NPIs are greater than 100 for 20,000 iterations, the calculation errors fall in the range from around 85% to 115% with the probability of 95%.

Next, we confirm whether the estimated aggregated NPIs converge as the number of iterations increases because if this is the case the calculation errors is ignorable with sufficiently many iterations. Figure 3a and b show the calculation errors for the entities whose aggregated NPIs are between 0.1 and 1 and between 100 and 1,000 with T_1 varying from 200 to 6,000 and T_0 remaining 20,000.

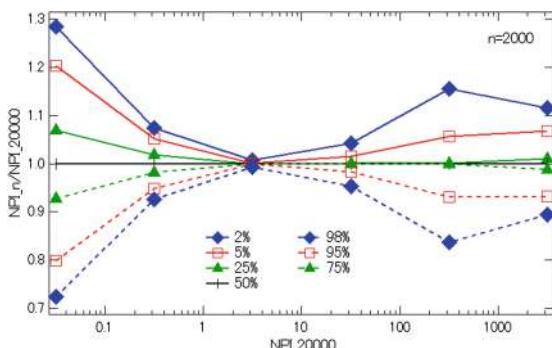


Fig. 2. Distribution of calculation errors where $T_0 = 20000$ and $T_1 = 2000$

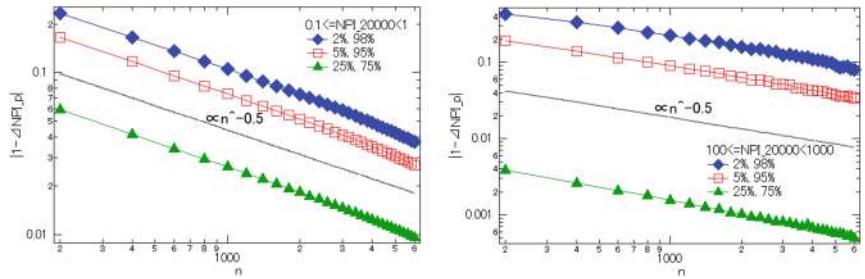


Fig. 3. The number of iterations and calculation errors in Label Propagation

The calculation errors shrink as the number of iterations becomes large. Moreover, the order of the convergence is approximately $\sqrt{T^{-1}}$, suggesting that Central Limit Theorem is at work and, therefore, the calculation errors are expected to converge to zero as $T \rightarrow \infty$.

4.2 Comparing NPIs to NSRs

We evaluate the aggregated NPIs *vis-a-vis* NSR. Figure 4 compares the aggregated NPIs to NSR. Although it looks as though the dots are symmetrically distributed around the 45-degree line, the number of investors whose influence is greater with NSRs (i.e., the entities located below the 45-degree line) is twice as small as the number of investors whose influence is less with their NSRs. Unlike NSRs, which is a simple product of vote shares, NPIs appreciates influence of small fraction of large investors, taking into account the non-linearity due to majority vote.

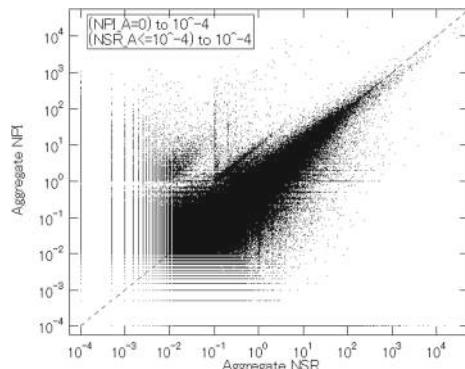


Fig. 4. Comparison of aggregated NPIs with aggregated NSRs

Table 1. Top-10 influential shareholders in terms of weighted aggregated NPIs

NSR	NPI	NPI (\$Billion)	Name
13285.6	13006.6	7392.6	Government of China
38980.7	19150.9	2617.7	Government of Norway
275.0	16833.6	2432.0	Capital Group Co Inc
26459.2	16242.6	2050.7	Wellington Management Group LLP
26.3	10138.3	1166.8	Sun Life Financial Inc
6446.2	6280.2	1049.0	Government of The Russian Federation
5.1	5390.7	989.4	Sumitomo Mitsui Trust Holdings, Inc
4.0	8143.6	951.1	HSBC Custody Nominees (Australia) Limited
79.3	7078.6	686.2	DIMENSIONAL FUND ADVISORS LP
6341.4	5132.3	640.5	Government of Singapore

How reasonable is it to rely on aggregated NPI instead of NSR? Or what is the relative advantage of this index? In order to answer these questions, we focus on highly influential investors and because every company are not equally important, we use aggregated NPIs weighted by the sales of each company under direct and/or indirect control. This weighted index can be interpreted as the expected value of total sales of companies which an investor can potentially control. Table 1 shows the top-10 influential shareholders: aggregated NSRs and NPIs appears in the first and second columns and aggregated NPIs weighted by sales in the third columns. It is straightforward that government and large financial institutions are listed as influential shareholders. A striking feature is that the divergence between (simple) aggregated NPIs and NSRs is heterogeneous, that is, whereas governments (except Norway) has similar values of NPIs and NSRs, financial institutions (except Wellington Management Group LLP) has dramatically large NPIs than NSRs. This finding suggests that financial institutions efficiently obtain the amount of vote shares sufficient to pass or block the bill at will in global networks. Relying only on NSRs may mislead researchers by making them dismiss “hidden influencers”.

5 Conclusion

This paper makes three contributions. First, we propose Network Power Index (NPI) to measure influence each investor i has on specific company j in a networked society and on the society as a whole. NPIs allow us to capture indirect influence i has on j through controlling the influence a third-party entity k has on j . This $i \rightarrow k \rightarrow j$ indirect influence is overlooked in the past application of the Shapley-Shubik power index as it only captures direct influence. Second, we propose an algorithm to numerically calculate NPIs since calculating NPIs entails indeterminacy. Our algorithm, *label propagation*, completes the calculation rapidly as it only calculates one-degree of links in each node and continues

until the calculation exhausts subsequent linkages in a network. Third, applying our method and algorithm to the data on global shareholding networks, we show that the gap between aggregated NPIs and NSRs can be quite substantial, suggesting that identifying a “hidden influencer” in a networked society would require the use of NPI.

Acknowledgements. This work was partially supported by JSPS KAKENHI Grant Numbers 18H03627 and 16H05904.

References

1. Bureau van Dijk’s Orbis. <https://www.bvdinfo.com/en-gb>. Accessed 27 Apr 2019
2. Banzhaf III, J.F.: Weighted voting doesn’t work: a mathematical analysis. *Rutgers L. Rev.* **19**, 317 (1964)
3. Bertini, C., Mercik, J., Stach, I.: Indirect control and power. *Oper. Res. Decisions* **2**, 7–30 (2016)
4. Crama, Y., Leruth, L.: Control and voting power in corporate networks: concepts and computational aspects. *Eur. J. Oper. Res.* **178**(3), 879–893 (2007)
5. Crama, Y., Leruth, L.: Power indices and the measurement of control in corporate structures. *Int. Game Theory Rev.* **15**(03), 1340017 (2013)
6. Dubey, P., Shapley, L.S.: Mathematical properties of the banzhaf power index. *Math. Oper. Res.* **4**(2), 99–131 (1979)
7. Gambarelli, G., Owen, G.: Indirect control of corporations. *Int. J. Game Theory* **23**(4), 287–302 (1994)
8. Karos, D., Peters, H.: Indirect control and power in mutual control structures. *Games Econ. Behav.* **92**, 150–165 (2015)
9. Laffont, J.J., Martimort, D.: *The Theory of Incentives: The Principal-Agentmodel*. Princeton University Press, Princeton (2009)
10. Levy, M.: Control in pyramidal structures. *Corp. Governance Int. Rev.* **17**(1), 77–89 (2009)
11. Levy, M., Szafarz, A.: Cross-ownership: a device for management entrenchment? *Rev. Finance* **21**(4), 1675–1699 (2016)
12. Mercik, J., Lobos, K.: Index of implicit power as a measure of reciprocal ownership. In: *Transactions on Computational Collective Intelligence XXIII*, pp. 128–140. Springer, Heidelberg (2016)
13. Penrose, L.S.: The elementary statistics of majority voting. *J. Roy. Stat. Soc.* **109**(1), 53–57 (1946)
14. Shapley, L.S., Shubik, M.: A method for evaluating the distribution of power in a committee system. *Am. Polit. Sci. Rev.* **48**(3), 787–792 (1954)
15. Shleifer, A., Vishny, R.W.: A survey of corporate governance. *J. Finance* **52**(2), 737–783 (1997)
16. Strøm, K.: Delegation and accountability in parliamentary democracies. *Eur. J. Polit. Res.* **37**(3), 261–290 (2000)
17. Vitali, S., Glattfelder, J.B., Battiston, S.: The network of global corporate control. *PloS One* **6**(10), e25995 (2011)



“Learning Hubs” on the Global Innovation Network

Michael A. Verba^(✉)

School of Economics and Management, Tilburg University, Tilburg, The Netherlands
M.A.Verba@tilburguniversity.edu

<https://research.tilburguniversity.edu/en/persons/michael-verba>

Abstract. In this paper, drawing on techniques from patentometrics, network analysis, and probability theory, we model the global system of innovation as a dynamic network. The sphere of technologically relevant knowledge is conceptualized as a reflexive, directed, link- and node-weighted complex network, with distinct spheres of knowledge (or technology domains) representing network nodes and learning (or knowledge flows) across domains acting as inter-nodal links. The empirical knowledge network is constructed from a sweeping patent database, including records from more than 100 patent-granting authorities over the 22-year period spanning 1991–2012. After establishing the structure of the global innovation network, we simulate its dynamics and study its evolution over time. The modelling exercise reveals technological trends and provides a ranking of technologies in terms of their level of technological dynamism.

Keywords: Innovation · Knowledge network · Network dynamics · Patents · Citations

1 Introduction

In a World moved by technological trends, the question of the relative importance of various technologies arises in policy and management settings. In this study we outline a methodological approach for ranking technology domains in terms of their importance within a broader network of technological learning and progress. We draw on the growing literature on knowledge networks [4, 5, 8, 18, 19] to conceptualize and model the stock of technological knowledge as a network of distinct but interacting technology domains. To operationalize the knowledge network we rely on findings and techniques from patentometrics, and methods from probability theory and network science. Our analysis exploits a unique global database covering patents granted at more than 100 intellectual property authorities during the 1991–2012 time period and including inter-patent citations. The data is a rich source of information on the state of technology available to humanity on the basis of which we can make inferences about the global system of innovation.

Prior research viewing knowledge systems as complex networks has tended to treat these networks as static graphs, and to use the tools of network science

designed for analysis of network topology. Yet, analysis of network dynamics provides several advantages in the context of citation networks. Firstly, citation networks can be meaningfully interpreted as networks of flows because each citation represents transfer of some quantity of knowledge, per unit of time, from one inventor to another. For knowledge networks, a flow-based analysis is truer to the object being studied than a purely topological approach more appropriate for physical infrastructure (such as road) networks, or undirected graphs representing connectivity arrangements (as in telephone circuit or computer networks), where flow volumes may not always be important (as, for example, in network vulnerability analysis). Second, the dynamic modeling approach lets us incorporate two essential features of knowledge networks: link weighting and directionality. Although weighting and directionality can be taken into account in a (static) analysis of network structure, flow-based analysis is particularly well adapted to the study of directed, weighted networks, such as trade [12] and citation [1] networks, and has previously been applied to bibliometric knowledge networks [2,9]. To our knowledge, this study is the first application of a dynamics-on-network analysis to a patent network constructed from a comprehensive database representing technology at the global level.

2 Formal Model

2.1 Topology of the Knowledge Network

At the abstract level, aggregated stock of technologically useful knowledge can be conceived as partitioned into distinct technology domains. New knowledge draws on pre-existing knowledge in own and other fields of technology. The size of the fields of knowledge, their growth rates, and the pattern of knowledge flows between them characterize the knowledge network. In the present section we describe such a network in general terms, by specifying a formal graph-theoretic model.

We model the knowledge network as a node- and edge-weighted, directed, reflexive graph, denoted by $G = \{V, E, h, w\}$, where $V = \{1, \dots, n\}$ is a finite set of vertices representing technology domains and $E = \{1, \dots, m\}$, is a finite set of ordered pairs of vertices in V , where each edge $(i, j) \equiv e_{ij}$, represents directed links between technology domains, with i as the head (source) and j as the tail (target) of the link. Each vertex $v_i \in V$ is associated with a node weight $h_i = h(v_i)$ given by the node-weight function $h : V \rightarrow \mathbb{R}^+$. In like fashion, each edge $e_{ij} \in E$ connecting nodes i and j is associated with a weight value $w_{ij} = w(e_{ij})$ representing the connection strength between nodes, given by the edge-weight function $w : E \rightarrow \mathbb{R}^+$. The adjacency matrix $A(G) = [a_{ij}]$ of graph G is a square $n \times n$ matrix consisting of elements:

$$a_{ij} = \begin{cases} w(e_{ij}) & \text{if } e_{ij} \in E \\ 0 & \text{otherwise} \end{cases}$$

In the adjacency matrix A element a_{ij} contains the weight of the link from the source node i to the target node j . A weight value of 0 is assigned if a link

is absent. Links are drawn in the direction of knowledge flow, which is opposite from the direction of knowledge search indicated by the citation.

To fully represent the knowledge network, graph G has two distinguishing attributes: it is reflexive and node-weighted. Unless stated otherwise, the default assumption in network analysis is to exclude the possibility of loops (or self-links), setting $w_{ii} = 0 \forall e_{ii} \in E$. By contrast, our graph is reflexive. Self-loops are allowed because technological domains can draw on internal knowledge. Hence, values in the diagonal of the adjacency matrix are not restricted to 0, although we do not exclude the theoretical possibility that there may be some technology domains that source knowledge exclusively from other domains. The operative restriction on the weight of self-links is $w_{ii} \geq 0 \forall e_{ii} \in E$.

The second attribute that differentiates our knowledge network from most graphs is that it is node-weighted. Node weights given by the function h represent the size of the domain of knowledge. The vector $H(G) = [h_i]$, defined as:

$$h_i = \begin{cases} h(v_i) & \text{if } v_i \in V \\ 0 & \text{otherwise} \end{cases}$$

contains the sizes of technology domains. The state of the network is completely specified by an $n \times n$ adjacency matrix A and an n -dimensional vector of node sizes H .

2.2 Modeling Network Knowledge Dynamics

Knowledge Flow as a Measure of Domain Importance. In social network analysis node importance has been defined in terms of centrality measures, such as degree centrality, closeness centrality, betweenness centrality, and other network-theoretic measures. These common measures have a connection to flow dynamics on the network. The greater the degree of a node, the more flow can pass through it. Likewise, a central or bridging node can be a more capacious conduit of flow as a result of its topological position. The relationship between network flow and topological measures has been explored previously in [3, 7, 13], with the latter study concluding that “most commonly used centrality measures are not appropriate for most of the flows we are routinely interested in” (p. 55).

Ultimately, it is the contribution of the node to network flow—by virtue of all its structural attributes—that represents its importance. Furthermore, nodal importance can also be influenced by intrinsic attributes, such as node size and magnitude of dynamics on the node itself. Simulating network flow directly is a way to estimate the contribution of each node to flow on the network. In the context of a knowledge network, a flow-based analytic approach lets us identify technologies that are relative “knowledge sinks”, i.e. receive large volumes of knowledge from other technological fields. Modeling knowledge dynamics of the global innovation network provides an empirical way to estimate the learning intensity for the set of technological domains that constitute the network, which in turn gives us a way to identify technological hot-spots.

Modeling Dynamics with Random Walks. A random walk can be used to simulate flow dynamics on a network. Flux simulated by random walks has been used to study certain topological features of networks [13] and for detecting network community structure, as in the Walktrap algorithm of [16] and the Map Equation Framework of [17] and its associated¹ In the present study we draw on and apply the methodology introduced by [17] in the context of community detection, directing it towards modeling knowledge dynamics on the patent citation network.

Knowledge Flow: From Heuristics to Simulation of the Stationary Distribution. The features of the random walk are adapted to reflect the pattern of knowledge dynamics on the empirical knowledge networks constructed from patent citation data. We can think of the random walker as a unit of knowledge circulating through the knowledge network.

Below, we outline the heuristics for the random walk process that will simulate flow on the network. First, flux is greater on heavily weighted links than on lighter links. Thus, the greater the total strength of all in-links connected to a node, the more frequently the random walk will return to the node, the more time the random walker will spend on the node, and the greater will be the node's flow estimate. Technologies that are knowledge-absorbing “learning hubs” will tend to cite more frequently. Knowledge generated on the node also contributes to nodal flux. The more innovations in a technology domain cite other innovations from within the domain, the greater the importance of autocatalytic knowledge dynamics and the longer the random walker will remain on the node once it arrives.

Hence, the conditional probability with which a random walker moves from node i to node j is given by the weight of the directed link connecting i to j , relative to the total weight of all out-links projecting from i to its neighbors:

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}. \quad (1)$$

Since self-links are included in the network structure, in the next step of the random walk the random walker can remain on node i with a probability proportional to the relative weight of the self-link.

Dynamics on the network can be measured with the steady-state distribution of the random walker on the network. With links drawn in the direction of knowledge flow, the random walker will tend to return to comparatively more knowledge-absorbing nodes that serve as learning hubs on the network. Changes in knowledge flow estimates will also document changes in intensity of innovative dynamism on the node.

If the network is irreducible and aperiodic, the flux generated by the random walker on the network will eventually approach a unique stationary distribution.

¹ Methods for community detection based on induced flows directed by the topology of the network are known as “flow models” [1]. A review of the palette of methods available for community detection can be found in [6].

However, if the network has disconnected or weakly connected components, the estimate of the steady-state distribution of the random walker on the network can be sensitive to the identity of the starting node. To overcome this problem, dynamics on complex networks can be simulated with a Markov process that includes a “teleportation parameter” τ which represents the probability with which the random walker jumps from node i to any other node in the network, chosen at random. Teleportation allows the random walker to get “unstuck” from a weakly connected or disconnected node, ensuring a unique steady-state solution.²

The stationary distribution of the random walker on the nodes of the network is given by the system of equations in Eq. (2), where the first term on the right-hand-side represents the probability that the random walker walks over to node i , while the second term represents the probability that the random walker arrives on node i as a result of teleportation.

$$p_i = (1 - \tau)\lambda_i + \tau \cdot h_i \quad (2)$$

In Eq. (2), $1 - \tau$ is the fraction of time our random walker walks between nodes, while τ is the fraction of time he hops to a node. Terms λ_i and h_i are conditional probabilities of arriving on node i by “walking” or through teleportation, respectively.

The probability the random walker steps onto node i is given by the multiple of p_j (the probability the walker is on node j) and p_{ji} (the conditional probability with which the random walker moves from node j to node i), summed over all j , as per Eq. (3):

$$\lambda_i = \sum_j p_j \cdot p_{ji}. \quad (3)$$

Probability p_{ji} can be decomposed further, as function of the weight of the directed link from j to i , relative to the total weight of all out-links projecting from j to neighboring nodes (analogous to the heuristic in Eq. (1)):

$$p_{ji} = \frac{w_{ji}}{\sum_k w_{jk}}. \quad (4)$$

We now have all the elements in place for a complete description of the walking process and turn to set the rules governing teleportation.

Teleportation is more likely to nodes that play an important role on the network. Therefore, the conditional probability of teleporting to a node on the network can be set proportional to the sum of weights of the node’s out-links, as per Eq. (5):

$$h_i = s_i^{out}. \quad (5)$$

The value s_i^{out} is also known as the out-strength of the node; it is defined as:

$$s_i^{out} \equiv \frac{\sum_j w_{ij}}{\sum_{i,j} w_{ji}}. \quad (6)$$

² See [17] for additional discussion of teleportation.

The stationary distribution of knowledge across technology domains is given by the solution to the following system of equations:

$$p_i = (1 - \tau) \sum_j p_j \left(\frac{w_{ji}}{\sum_k w_{jk}} \right) + \tau \left(\frac{\sum_j w_{ij}}{\sum_{i,j} w_{ji}} \right) \quad (7)$$

which is derived by substituting Eq. (4) into Eq. (3), Eq. (6) into Eq. (5) and Eqs. (3) and (5) into Eq. (2). In Eqs. (1)–(7) indices i, j and k are over the full range of knowledge categories in set O . We solve the system in Eq. (7) via Von Mises iteration. The solution provides the steady-state distribution of knowledge flow across the nodes of the global technology network, which we describe and discuss in the next section.

3 The Empirical Network

3.1 Data

In this paper we work with patent data covering the 22-year period spanning 1991–2012, drawn from the April 2013 edition of the PATSTAT database. This proprietary database contains more than 83 million patent records from 105 patent authorities and includes more than 135 million citation records.³ On the whole, the database provides a uniquely comprehensive view of the evolution of technological progress and inter-technological learning.

We define distinct technologies on the basis of the International Patent Classification System (IPC) maintained by the World Intellectual Property Organization (WIPO).⁴ The use of patent classes to track developments within specific technologies is a widely accepted technique in both scholarship and policy.⁵ IPC is particularly attractive for patentometric purposes; its widespread use even by patent authorities with concurrent native systems facilitates consolidation and harmonization of patent documents at a global level. Furthermore, the IPC provides precise definitions of the domains' contents, based on well-defined rules. Finally, the system is curated by experts, which ensures a high degree of consistency and reliability.

3.2 Overview of the Empirical Network

To analyze knowledge dynamics between technologies, we take unique inventions granted within each patent subclass, as well as citations between unique

³ This number includes citations to non-patent literature, which are excluded from our analysis.

⁴ Our database is linked to the January 2006 version of the IPC (also called Version 8), which is divided into 8 sections, 129 classes, 639 subclasses, 7,314 main groups and 61,397 subgroups.

⁵ With [10, 15] and [11], among many examples.

inventions, aggregated at subclass level.⁶ We construct the global knowledge network by calculating the fractional count of the total number of links between IPC subclasses. The timeline covered by the analysis starts in 1991 and ends in 2012.⁷ We then split the timeline for which data are available into two periods of 11 years in duration. Network I captures knowledge dynamics during the 1991–2001 period, while Network II covers the years 2002–2012.

The constructed empirical network pertaining to the 1991–2001 period (Network I) consists of 629 technological sub-classes and 180,218 directed weighted links (including self-links). The network representing knowledge space in the 2002–2012 period (Network II) is only slightly bigger, at 637 nodes, but more dense, with 245,795 links. Networks I and II represent separate cross-sections, or “snapshots”, of the network at two time periods.

4 Results

Empirical simulation of knowledge dynamics on our networks requires specification of τ . For each empirical knowledge network we run the simulation under a number of alternative assumptions for τ , including 0, 0.01, 0.05, 0.1, 0.15, 0.25, 0.35 and 0.5. We find that the importance ranking of technology domains is highly stable. While nodal flow estimates are sensitive to the teleportation parameter, node relative ranking changes very little as a result of changes in τ (except when the change is from $\tau > 0$ to $\tau = 0$). In the remainder of the paper we focus on estimates obtained from simulations performed with the baseline $\tau = 0.15$.

4.1 Ranking Technological Domains by Knowledge Flow

Tables 1 and 2 present the estimated knowledge flow on the global innovation network during the 1991–2001 and 2002–2012 periods, respectively. Technology domains are ranked according to their absorption of knowledge on the network, as measured by nodal flow. The first column of the two tables represents subclass rank. Column (2) represents change in subclass ranking between the two long decades: 1991–2001 ranking minus the rank in 2002–2012. By construction, a

⁶ Because the patent record can contain multiple documents per invention, reflecting different types of legal events, proper handling of the record base is important. Within the same patent office, the multiple events include (but are not limited to) patent application, granting, modification, inclusion of additional claims through a patent of addition, or a divisional patent that splits the claims to an existing invention into separate bundles of rights. Most importantly, there exist applications and grants covering the same invention in multiple jurisdictions. Inclusion of all records in the construction of the knowledge network would lead to distortive duplication. We overcome this potential hazard by extracting the unique invention from the multitude of redundant records.

⁷ Between 1980 and 1990, data are very sparse, with less than 1000 links annually, so we excluded these years from the analysis.

Table 1. Network 1 (1991–2001), ranking of subclasses by knowledge flow

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rank	Δ Rank	IPC Subclass	Abridged Description	Total Patents		
				Flow	(Prop.)	(1, 000 s)
1	0	G06F	Electric digital data processing	0.076	0.031	174
2	0	H01L	Semiconductor devices; electric solid state devices	0.050	0.039	218
3	0	A61B	Diagnosis; surgery; identification	0.043	0.013	74
4	0	A61K	Preparations for medical, dental, or toilet purposes	0.041	0.029	161
5	0	H04L	Transmission of digital information, e.g. Telegraphy	0.035	0.015	82
6	0	A61F	Filters implantable into blood vessels; prostheses	0.030	0.009	49
7	0	H04N	Pictorial communication, e.g. Television	0.023	0.022	120
8	-2	G01N	Determining chemical or physical properties of materials	0.022	0.017	94
9	0	G06Q	Data processing, esp. administrative and financial	0.020	0.003	18
10	-1	A61M	Devices for introducing media into, or onto, the body	0.019	0.007	37
11	-7	G11B	Information storage (based on movement)	0.016	0.017	96
12	0	G02B	Optical elements, systems, or apparatus	0.015	0.012	64
13	-2	C07D	Heterocyclic compounds	0.012	0.013	72
14	-14	C12N	Micro-organisms or enzymes; compositions thereof	0.011	0.008	47
15	-1	H04B	Transmission of information	0.011	0.010	54
16	-3	H04M	Telephonic communication	0.011	0.007	39
17	-7	A61P	Therapeutic chemical compounds or medicinal preparations	0.010	0.011	60
18	-4	B41J	Typewriters; selective printing mechanisms	0.010	0.008	45
19	2	A61N	Electro-, magneto-, radiation- and ultrasound-therapy	0.010	0.003	15
20	-7	B65D	Containers for storage or transport of articles or materials	0.009	0.012	64
21	-18	B29C	Shaping or joining of plastics	0.009	0.010	54
22	8	E21B	Earth or rock drilling	0.008	0.006	31
23	15	G06K	Recognition and presentation of data	0.008	0.005	26
24	-16	C07K	Peptides	0.008	0.005	28
25	12	G11C	Information storage (static)	0.008	0.007	37
26	-36	A61L	Sterilisation and disinfection	0.007	0.003	17
27	-28	G06T	Image data processing or generation	0.007	0.006	31
28	-5	H05K	Printed circuits; manufacture of electrical components	0.007	0.008	42
29	9	H04W	Wireless communication networks	0.007	0.004	24
30	4	B01D	Separation	0.007	0.009	52

Note: Network 1 represents all patents granted during the 1991–2001 period, aggregated into 4-digit patent subclasses (nodes), along with inventor-origin citations between subclasses (links). The empirical knowledge network consists of 629 subclasses. Presented are the top 50 subclasses as measured by knowledge flow. Knowledge flow on the network is simulated with random walks. Column descriptions: (1) Subclass rank in Network 1 (1991–2001). (2) Change in rank: rank in 1991–2001 minus rank in 2002–2012. (3) IPC Version 8 subclass code. (4) Subclass description, based on IPC Version 8 definition, but abridged for compactness. (5) Proportion of network knowledge flowing onto the node representing the subclass. (6) Subclass patents as proportion of all patents granted during the 1991–2001 period. (7) Number of patents granted during the 1991–2001 period assigned to subclass, in thousands (fractional count). An extended version of this table containing all subclasses is available from the author.

Table 2. Network 2 (2002–2012), ranking of subclasses by knowledge flow

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rank	Δ Rank	IPC Subclass	Abridged Description	Total Patents		
				Flow	(Prop.)	(1, 000s)
1	0	G06F	Electric digital data processing	0.107	0.051	272
2	0	H01L	Semiconductor devices; electric solid state devices	0.059	0.052	277
3	0	A61B	Diagnosis; surgery; identification	0.043	0.015	82
4	0	A61K	Preparations for medical, dental, or toilet purposes	0.031	0.024	131
5	0	H04L	Transmission of digital information, e.g. Telegraphy	0.026	0.025	134
6	0	A61F	Filters implantable into blood vessels; prostheses	0.023	0.007	39
7	0	H04N	Pictorial communication, e.g. Television	0.018	0.023	123
8	15	G06K	Recognition and presentation of data	0.017	0.008	43
9	0	G06Q	Data processing, esp. administrative and financial	0.017	0.007	37
10	-2	G01N	Determining chemical or physical properties of materials	0.016	0.018	94
11	-1	A61M	Devices for introducing media into, or onto, the body	0.016	0.006	30
12	0	G02B	Optical elements, systems, or apparatus	0.015	0.013	70
13	12	G11C	Information storage (static)	0.014	0.008	44
14	8	E21B	Earth or rock drilling	0.013	0.006	32
15	-2	C07D	Heterocyclic compounds	0.013	0.010	52
16	-1	H04B	Transmission of information	0.012	0.014	76
17	2	A61N	Electro-, magneto-, radiation- and ultrasound-therapy	0.011	0.003	14
18	-7	G11B	Information storage (based on movement)	0.011	0.013	68
19	-3	H04M	Telephonic communication	0.010	0.007	37
20	9	H04W	Wireless communication networks	0.010	0.011	58
21	23	B32B	Layered products	0.010	0.005	27
22	-4	B41J	Typewriters; selective printing mechanisms	0.009	0.009	46
23	22	H01R	Electrically-conductive connections; coupling devices	0.009	0.009	46
24	-7	A61P	Therapeutic chemical compounds or medicinal preparations	0.009	0.010	56
25	14	G09G	Arrangements or circuits for control of indicating devices	0.009	0.007	38
26	4	B01D	Separation	0.008	0.008	44
27	-7	B65D	Containers for storage or transport of articles or materials	0.008	0.009	49
28	-14	C12N	Micro-organisms or enzymes; compositions thereof	0.008	0.007	39
29	4	G01R	Measuring electric variables; measuring magnetic variables	0.007	0.008	41
30	10	A63B	Apparatus for physical training	0.007	0.004	21

Note: Network 2 represents all patents granted during the 2002–2012 period, aggregated into 4-digit patent subclasses (nodes), along with inventor-origin citations between subclasses (links). The empirical knowledge network consists of 637 subclasses. Presented are the top 50 subclasses as measured by knowledge flow. Knowledge flow on the network is simulated with random walks. Column descriptions: ⁽¹⁾ Subclass rank in Network 1 (2002–2012). ⁽²⁾ Change in rank: rank in 1991–2001 minus rank in 2002–2012. ⁽³⁾ IPC Version 8 subclass code. ⁽⁴⁾ Subclass description, based on IPC Version 8 definition, but abridged for compactness. ⁽⁵⁾ Proportion of network knowledge flowing onto the node representing the subclass. ⁽⁶⁾ Subclass patents as proportion of all patents granted during the 2002–2012 period. ⁽⁷⁾ Number of patents granted during the 2002–2012 period assigned to subclass, in thousands (fractional count). An extended version of this table containing all subclasses is available from the author.

negative integer in column (2) represents a drop in subclass importance over the past two decades, while a positive integer indicates appreciation in ranking. The next two columns provide the IPC code and description for each subclass, based on the original IPC definitions.⁸ Column (5) contains a measure of the flow of knowledge on the node—the focal indicator of this analysis. This contribution is measured as the proportion of time the simulated random walker spends on the node in the steady-state. The proportions sum to 1 when all domains are taken into account. The last two columns are based on the count of patents associated with each IPC subclass. Column (6) presents the patents allied to each subclass as a proportion of all patents granted during the period—a simple measure of relative subclass size. Column (7) contains the total number of patents granted in each subclass. Columns (6) and (7) allow us to evaluate the extent to which knowledge flows, dissociated from intrinsic technology attributes and network position, are a measure of the importance of patent categories to innovation dynamics. In many cases, subclasses that do not rank highly on patent-count-based measures of importance, turn out to be vital to the knowledge network (for example, see subclass A61F in Network II).

4.2 Identifying Primary “Learning Hubs”

Tables 1 and 2 give an overview of the evolution of technology. Drawing a line between primary “learning hubs” and other technologies is somewhat arbitrary if the placement is based solely on the ranking. However, if we add to the concept some measure of persistence, the list is more straightforward to define. The top 7 subclasses in Tables 1 and 2 consistently play an oversized role in knowledge dynamics on the innovation network. These 7 learning-hub technologies each constitute between 2 and 11% of flow on the network. Further, going from the 1991–2001 to the 2002–2012 network, we see that the relative rank of the top 7 subclasses has remained unchanged.

In both decades, among the top 7 learning hubs one finds select technologies focusing on information, computing and telecommunication (ICT)⁹ (“Electric digital data processing” [G06F], “Semiconductor and electric solid state devices” [H01L], “Transmission of digital information” [H04L], “Pictorial communication” [H04N]) and categories dealing with medicine (“Diagnosis, surgery and identification” [A61B], “Preparations for medical, dental or toilet purposes” [A61K], “Filters implantable into blood vessels and prostheses” [A61F]). Altogether, the above 7 subclasses account for 30.7 (29.8) percent of all knowledge flow, and 19.7 (15.8) percent of all patents granted in the 2002–2012 (1991–2001) period.

⁸ Readers are encouraged to consult the official International Patent Classification system for more detail on the content of patent subclasses.

⁹ Although the importance of information and telecommunication technology (ICT) is widely recognized, we see from our results that only specific technological domains within the broader ICT constellation of technologies have a key role in the global innovation system.

4.3 Identifying Technology Trends Through Change in the Contribution of Technology Domains to Knowledge Flow on the Global Innovation Network

For the top technologies we observe some change in knowledge flow, but a constant relative ranking over a long period of time. For other technological domains, however, we see movement in knowledge flow, as well as in relative ranking, that in some cases is quite pronounced. Changes in knowledge dynamics across the network and within individual subclasses give us a unique view on global trends in technology. In the remainder of this section we highlight several key findings that emerge from the results.

Consolidating information from Tables 1 and 2 we highlight technological domains that have experienced profound change in their contribution to dynamics of technological innovation. From the subclasses that were sizeable enough in their learning intensity to feature on the top 50 list either in 1991–2001 or 2002–2012, the following 9 technologies underwent the greatest appreciation in ranking, moving up 20 places or more:

- Signalling or calling systems; order telegraphs; alarms [G08B] (+67)
- Functional features or details of lighting devices [F21V] (+65)
- Taking, projecting or viewing photographs [G03B] (+54)
- Electric heating; electric lighting [H05B] (+28)
- Electrography; electrophotography; magnetography [G03G] (+24)
- Layered products [B32B] (+23)
- Card, board, or roulette games [A63F] (+22)
- Electrically-conductive connections; coupling devices [H01R] (+22)
- Speech analysis or synthesis; speech recognition [G10L] (+20).

During the past two decades, among the technological fields that have lost importance in innovation intensity, moving down in ranking by 20 places or more, are the following 6 subclasses:

- Devices using stimulated emission [H01S] (-20)
- Coin-free or like apparatus [G07F] (-24)
- Image data processing or generation [G06T] (-28)
- Detergent compositions [C11D] (-35)
- Sterilisation and disinfection [A61L] (-36)
- Selecting [H04Q] (-36)

The decline in the innovative role of subclass H01S deserves a special note. To non-specialists outside the laser industry, the above finding may come as a surprise, as indeed it was to this author. Lasers are still examples of cutting-edge technology in popular imagination. But professionals in the photonics industry have known for some time that lasers are a maturing technology with well-established product niches [14]. Our model of knowledge network dynamics picks up on the drop in innovative dynamism in this technological domain.

5 Discussion and Conclusion

Technology managers require information about past development, current position, and likely future trajectory of technologies in their portfolio. By synthesizing techniques from patentometrics, dynamical network models, and probability theory, this study went beyond conventional methods, such as simple or composite indexes used in monitoring technology trends. One of the contributions of the presented methodology is its ability to take into account the interactive and relational nature of innovative activity.

Due to space constraints, the information on technology trends only scratch the surface of the information that can be revealed by application of the implemented methodology. While our measure of knowledge flows on nodes of the global innovation network captures in a summary way all information about the relational structure and dynamics of the network, there exist a number of questions we have barely touched, that could also be answered by application of similar methods. For example focus on dynamics could also reveal interaction between specific components of the network. We place identification and exploration of modular sub-components of the global knowledge network on our future research agenda and in a forthcoming companion study, we explore the evolving community structure of the global innovation network.

Acknowledgements. This work was carried out on the Dutch national e-infrastructure with the support of SURF Foundation. The author is thankful to the Netherlands Organisation for Scientific Research (NWO) for its grant program for high-performance computing and to the staff of SURF. Additionally, this study was partially supported by United Nations University—Maastricht Economic and Social Research Institute on Innovation and Technology (UNU-MERIT). The author is particularly grateful to Jojo Jacob and Mathijs Kattenberg for their help and technical advice during the course of this project. All remaining errors are my own.

References

1. Bohlin, L., Edler, D., Lancichinetti, A., Rosvall, M.: Community detection and visualization of networks with the map equation framework (2014). <http://www.maapequation.org/publications.html>
2. Bohlin, L., Viamontes Esquivel, A., Lancichinetti, A., Rosvall, M.: Robustness of journal rankings by network flows with different amounts of memory. *J. Assoc. Inf. Sci. Technol.* (2015)
3. Borgatti, S.P.: Centrality and network flow. *Soc. Netw.* **27**(1), 55–71 (2005). <http://www.sciencedirect.com/science/article/pii/S0378873304000693>
4. Carnabuci, G., Bruggeman, J.: Knowledge specialization, knowledge brokerage and the uneven growth of technology domains. *Soc. Forces* **88**(2), 607–641 (2009)
5. Feldman, M.P., Kogler, D.F., Rigby, D.L.: rKnowledge: the spatial diffusion and adoption of rDNA methods. *Reg. Stud.* **49**(5), 798–817 (2015). <https://doi.org/10.1080/00343404.2014.980799>
6. Fortunato, S.: Community detection in graphs. *Phys. Rep.* **486**(3–5), 75–174 (2010). <https://doi.org/10.1016/j.physrep.2009.11.002>

7. Freeman, L.C., Borgatti, S.P., White, D.R.: Centrality in valued graphs: a measure of betweenness based on network flow. *Soc. Netw.* **13**(2), 141–154 (1991)
8. Lafond, F.: Knowledge networks. Ph.D. thesis, Maastricht University, December 2015
9. Lambiotte, R., Rosvall, M.: Ranking and clustering of nodes in networks with smart teleportation. *Phys. Rev. E* **85**(5), 056107 (2012)
10. Leydesdorff, L.: Patent classifications as indicators of intellectual organization. *J. Am. Soc. Inf. Sci. Technol.* **59**(10), 1582–1597 (2008). <https://doi.org/10.1002/asi.20814>
11. Leydesdorff, L., Kushnir, D., Rafols, I.: Interactive overlay maps for US patent (USPTO) data based on international patent classification (IPC). *Scientometrics* **98**(3), 1583–1599 (2014). <https://doi.org/10.1007/s11192-012-0923-2>
12. McNerney, J., Fath, B.D., Silverberg, G.: Network structure of inter-industry flows. *Phys. Stat. Mech. Appl.* **392**(24), 6427–6441 (2013). <http://www.sciencedirect.com/science/article/pii/S0378437113006948>
13. Newman, M.E.: A measure of betweenness centrality based on random walks. *Soc. Netw.* **27**(1), 39–54 (2005). <http://www.sciencedirect.com/science/article/pii/S0378873304000681>
14. Overton, G., Anderson, S.G.: Laser marketplace 2009: photonics enters a period of high anxiety (2009). <http://www.laserfocusworld.com/articles/2009/01/laser-marketplace-2009-photonics-enters-a-period-of-high-anxiety.html>. Accessed 1 Jan 2009. <http://www.laserfocusworld.com/articles/2009/01/laser-marketplace-2009-photonics-enters-a-period-of-high-anxiety.html>
15. Paci, R., Sassu, A., Usai, S.: International patenting and national technological specialization. *Technovation* **17**(1), 25–38 (1997). <http://www.sciencedirect.com/science/article/pii/S016649729600065X>
16. Pons, P., Latapy, M.: Computing communities in large networks using random walks. *J. Graph Algorithms Appl.* **10**(2), 191–218 (2006)
17. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. *Proc. Natl. Acad. Sci.* **105**(4), 1118–1123 (2008)
18. Triulzi, G.: Looking for the right path: technology dynamics, inventive strategies and catching-up in the semiconductor industry. Ph.D. thesis, Maastricht University, December 2015
19. Valverde, S., Solé, R.V., Bedau, M.A., Packard, N.: Topology and evolution of technology innovation networks. *Phys. Rev. E* **76**, 056118 (2007). <https://doi.org/10.1103/PhysRevE.76.056118>



Global Transitioning Towards a Green Economy: Analyzing the Evolution of the Green Product Space of the Two Largest World Economies

Seyyedmilad Talebzadehhosseini, Steven R. Scheinert,
and Ivan Garibay^(✉)

Complex Adaptive Systems Laboratory, Department of Industrial Engineering
and Management Systems, University of Central Florida, Orlando,
FL 32816, USA
igaribay@ucf.edu

Abstract. Research on transitioning towards a green economy seeks to enable countries to expand their green production basket more competitively than other nations. Current research argues that countries are capable of producing new green products if these new green products share similar capability requirements (e.g. technology, capital, skills, etc.) with green products that the country already produces and exports with a high Real Comparative Advantage (RCA). Recently, the Green Product Space (GPS) framework was developed based on the hypothesis that countries diversify their green production baskets following the path-dependent process, meaning that a country's green production basket development should be based on its current accumulation of green capabilities. In this paper, we conduct a comprehensive analysis to identify patterns of green growth between 2007 and 2017. The results of our analysis indicate that countries which successfully advanced and diversified their green production basket not only followed the path-dependent process, but also follow the non-path dependent process we term "high-energy jumps".

Keywords: Green Product Space · Green economy · Green products · Revealed Comparative Advantage · Global transition · United States of America · China

1 Introduction

The importance of countries' transitions to green economies has increased as the world's population increased, as natural resource use increased to meet population needs [6]. Due to this increase of natural resources demand, countries are faced with several challenges such as climate change and increased environmental risk [1, 15]. Thus, it is important for countries to move towards an economy with low resource requirements and low carbon outputs, that is, a green economy [10]. Aligning a country's production capabilities with its green development agenda is important for policy makers [17], as the better the plan to expand green production, the more

diversified the green products basket will be, which will enable the country to align its production capabilities with its green development agenda [3].

Research on the direction that countries take to expand their green production basket is based on the concept of path-dependency developed by [17], which states that a country's green production basket diversifies based on the current capabilities it has, such as technologies, capital, institutions, and population skills [2, 13, 16]. According to this hypothesis, countries with high green production capabilities are able to grow their green economy faster than countries with less capabilities, as they are already capable of producing different green products (a more diversified green production basket). Thus, this research enables policy makers to identify green products that countries' already have with the highest potential for growth, which therefore should be promoted through targeted policy actions.

Research based on the path-dependent hypothesis is important for developing countries, as their ability to enhance their green production basket is much less than developed countries [17]. However, recent research on the evolution of countries' production capabilities by [3] showed that countries not only advance their production basket following the path-dependent process, but that countries also followed the non-path dependent process for introducing new products. The non-path dependent process in green economic growth is one of innovation, in essence, the diversification of green products as a result of investments in new environmental technologies, structural changes, and a country's current green production capabilities [3, 4]. The Green Product Space (GPS) developed by [17] predicted a countries' green growth based on the path-dependent process, as it is based upon the idea that countries expand their green production basket to green products which a country is already capable of producing. In this research, the GPS for China and the United States was developed, following the method proposed by [17], to identify their patterns of green growth between 2007–2017, as they are the two largest economies in the world and two of the leaders in green economic growth. This is done using green product historical world trade data between 2007 to 2017. All this was done to answer the following question:

1. How have the United States and China, as two of the leaders in green growth, diversified their green production basket to produce more green products from 2007 to 2017?

This paper is related to the recent contribution of [17], which examined countries that expanded their green production baskets following the path-dependent process, and [3]'s research, which proposed a method to analyze the pattern of economic growth in countries.

2 Related Work

Environmental problems such as climate change and increasing environmental risks have become an important issue for many governments in the world; therefore, it is important for them to consider transitioning to cleaner production, as this will decrease the countries CO₂ emissions and environmental risks [8, 9]. [17] used the concept of

economic complexity [13] and constructed the GPS network to show that countries' green growth is highly dependent on the green products that they are already capable of producing and exporting. Additionally, they showed that the higher the green growth countries have, the lower CO₂ emissions and environmental risks they have. [3] used the concept of economic complexity [13] and explored the pattern of economic growth in all countries, and in doing so proposed a method to test whether countries follow path-dependent process to expand their production baskets or not. Their results confirmed that countries followed the path-dependent process to expand their production baskets, and at the same time that for considerable amount of new products, countries followed the non-path dependent process, which leads them to have a higher rate of economic growth. [6] proposed a method which draws on the concept of economic complexity [13], and showed that the green products with highest potential for growth are the ones that are closest to the products with a high Revealed Comparative Advantage (RCA). The concepts of GPS, and RCA will be explained in detail in Sects. 2.1 to 2.3 as they are the main concepts used to identify the United States and China's pattern of green growth.

2.1 Revealed Comparative Advantage and Product Proximity

The concept of RCA is defined as a country's ability to produce products relative to its trading partners [7]. According to this definition, a country, c , can consider itself to have an RCA on product i if the following condition is met [11]:

$$RCA_{ci} = \frac{\frac{X_{ci}}{\sum_j X_{cj}}}{\frac{\sum_c X_{ci}}{\sum_{cj} X_{cj}}} \geq 1 \quad (1)$$

where X_{ci} shows the country's (c) export value for product i , $\sum_j X_{cj}$ shows the country's total export value for all products j , $\sum_c X_{ci}$ shows total export value of product i in the world, and $\sum_{cj} X_{cj}$ shows the total export values of all products for all countries in the world. In fact, if the result of the ratio for country's (c) product, i , stands above one, it means the country has an RCA on product (i) and is determined to be competitive in both producing and exporting this product, relative to the world average [10]. The concept of product proximity was introduced by [12, 14]. According to their definition, product proximity represents how similar two products are, i.e., if a country is a competitive producer of apples, then it has the capabilities (e.g. technology, capital, skills, infrastructures) to produce peaches, but it may not have the capability to produce car engines. In essence, apples and peaches have a high proximity since they share similar required capabilities, while apples and car engines have a low proximity since they do not share similar required capabilities [6]. Given that a country competitively exports product j , the proximity between i and j represents the lowest number of possible pairwise conditions probabilities that the country exports i competitively [6] and can be calculated as

$$\varphi_{i,j} = \min\{P(x_i/x_j), P(x_j/x_i)\} \quad (2)$$

where $\varphi_{i,j}$ is the proximity value between products i and j , and $P(x_i/x_j)$ can be defined as the probability that i is exported competitively, given that j is also competitively exported [6]. Similarly, $P(x_j/x_i)$ can be defined as the probability that i is exported competitively, given that j is also competitively exported [6]. RCA and product proximity should be understood by policy makers so they can make fully informed decisions to achieve their goals in economic growth for their country.

2.2 Product Space

The concepts of Product Space (PS) and path-dependent economic growth was introduced by [13] as a network that connects 775 products that have been traded between all countries in the world, with growth following the connections in this network. According to the PS network, the 775 products are connected based on their proximity values, that is, if two products share similar required capabilities, they have a high proximity value, and if they share less capabilities, they have a low proximity value. According to [13], most developed countries are capable of producing products that stand in the high density areas of the PS network, which shows their high ability to produce different types of products, while developing countries normally stand in the sides of the product space, which shows that they need more advancement to enter the highest density area of the network. As argued by [13], the development of new products in a country follow a path-dependent process; this path is determined by the current capabilities a country has for production. That is, countries determine what products to diversify to based on the RCA of currently produced products. For instance, if a country has an $RCA \geq 1$ for product i , it shows that the country has the capital, technology, and infrastructure to produce a new product, j , that has a high proximity with product i .

2.3 Green Product Space

Similar to [13, 17] used the concept of PS to develop a GPS network for 293 green products that were traded between 1995 and 2014 for all countries. The main hypothesis of GPS is that countries follow the path-dependent process to enhance their green production basket, that is, countries use their existing green production capabilities to produce new green products and diversify their green production basket. In addition, [17] constructed the green adjacent possible index based on the path-dependence hypothesis to identify the green products that countries have potential to produce in order to enhance their green production basket. Finally, the authors constructed a measure, the green complexity potential, to predict a country's future green product competitiveness, and stated that the relation between the green adjacent possible index and green complexity potential can determine the direction that countries will take to grow their green economy [17].

3 Data

As mentioned earlier, a comprehensive list of green products is constructed in this paper to develop a GPS network. In order to construct a list of green products, a combined list of environmental goods developed by the Organization for Economic Cooperation and Development (OCED) is used [18]¹. The 246 green products are based on the 6-digit Harmonized System (HS). The data for these green products is obtained from the Observatory of Economic Complexity (OEC) [19]. The data includes the export value of each green product that each country exported between 2007 and 2015, the Product Complexity Index (PCI) of each green product, and the green product's name. The export value of each green product for each year is based on US Dollars, and the green product's PCI shows the green product's complexity. The higher the PCI, the more advanced the capabilities needed to produce the green product. To calculate the RCA and proximity value for the listed green products, the green product's export value is used.

4 Methods

The methods used in this paper are divided in two sections. First, we construct the GPS, and second, we track the patterns of growth over time in this green product space to identify patterns of green growth in the United States and China.

4.1 Constructing Green Product Space

Similar to [17], the GPS for 246 green products is developed. However, the GPS network in this paper is developed based on the data from 2007 to 2017, which is more recent. In order to develop a green product space, several steps must be taken:

- From the obtained data, we extract the export value of each green product for each country.
- The RCA value for each green product is calculated using Eq. (1) to identify the green products that each country produces and exports with a comparative advantage.
- After calculating the RCA for the green products, a matrix M_{cp} is developed, where c stands for countries and p stands for green products. The element of the matrix is either one or zero, that is, if the green product's RCA is above one, the element is one, otherwise, it is zero.
- Then, the proximity between all green products is calculated using Eq. (2).
- In the last step, a matrix is developed in which the elements of the matrix are the proximity values calculated in previous step. The proximity values show how two

¹ List of green products can be found here: https://www.oecd-ilibrary.org/trade/the-stringency-of-environmental-regulations-and-trade-in-environmental-goods_5jxrjn7xsnmq-en;jsessionid=T-vbFJB-__7idgVjPRHssK3F.ip-10-240-5-20.

different green products are related to each other. The higher the proximity value is, the more similar the capabilities required for production are for the two green products.

Following these steps enabled us develop a Green Product Space (GPS) for all 246 green products. The GPS is a network that maps all 246 green products based on the similarity of their required capabilities. The nodes in the GPS network are the 246 green products, and the links are the proximity values that connect the nodes.

4.2 Testing the Pattern of Green Growth Based on Developed Green Product Space

In order to understand the pattern of green growth in the GPS, the following steps are followed [3, 20]:

- To begin the analysis, what a new green product is must be defined. Thus we define a new green product as a product that a country was not a competitive producer of ($RCA < 0.5$) in 2007, but became a competitive producer of the product ($RCA > 1$) in 2017. A RCA of <0.5 is considered the threshold for undeveloped green products, as considering other threshold values did not provide a significant difference in the final results.
- The time period of 2007 to 2017 is used to determine new green products for each country. We considered nine time intervals within this span: [2007–2009], [2008–2010], [2009–2011], [2010–2012], [2011–2013], [2012–2014], [2013–2015], [2014–2016], and [2015–2017], to explore a country's green diversification in detail.
- A list of new green products is developed for each country in each time interval (K_c).
- The RCA for all green products is calculated using Eq. (1), and for each base year, e.g. 2007, 2008, 2009, etc. the products with $RCA > 1$ are listed (U_{ct_0}).
- The proximity values between all pairs of green products is calculated for each base year according to Eq. 2.
- In the last step, a proximity matrix for all green products is constructed based on

$$D_{ic} = \left\{ \begin{array}{ll} d_{ic}(\varphi_{i,j}) = \max(\varphi_{i,j}) & \text{when } j \in U_{ct_0}, i \in K_c \\ \text{no value} & \text{if } j \notin U_{ct_0} \end{array} \right\} \quad (3)$$

where $d_{ic}(\varphi_{i,j}) = \max(\varphi_{i,j})$ shows the proximity of new green products to the most related green products ($RCA > 1$) at each base year [3], U_{ct_0} represents the list of green products with $RCA > 1$ as described in step 4, and K_c is the list of new green products within a certain time interval. The maximum proximity is considered in order to determine whether new green products are developed based on products with $RCA > 1$ at the base year. After implementing these analyses, a statistical analysis is used to show how China and the United States expanded their green production basket. To this end, similar to [3], this analysis was implemented using the non-parametric statistical approach, specifically the Kernel Density Estimation (KDE). KDE uses all data points

to estimate the shape of a dataset [5]. For the purpose of this research, and in order to understand how new green products (K_c) are added to the United States and China green production basket, the kernel smoothed density estimation function is used for any level of proximity (relatedness) values. The function is as follows:

$$\bar{K}(d) = \frac{1}{\left(\sum_{i=1}^M \sum_{t=2007}^{2017} I_{it}\right)h} \sum_{i=1}^M \sum_{t=2007}^{2017} I_{it} f\left(\frac{d - d_{it}}{h}\right)$$

where "...densities [are] calculated non-parametrically using a Gaussian Kernel function with bandwidth h set according to Silverman's optimal rule of thumb" [3]. In the above equation, $\sum_{i=1}^M \sum_{t=2007}^{2015} I_{it}$ equals the total number of new green products in each time interval, and d_{it} is calculated using the equation in step six. In order to test [17]'s path-dependent hypothesis, we follow the steps below:

1. The above equation is used to estimate the shape (distribution) of the relatedness (proximity) values in step six. This represents the actual relatedness value.
2. The Monte Carlo simulation is implemented, which takes "...1,000 random draws [equal to] the actual number of new [green products]" [3] in each time interval to generate a simulated relatedness value. These calculations allow us to compare the relatedness values of new green products calculated in the first step with the relatedness values of randomly generated new green products.
3. The distribution of relatedness values (actual relatedness) for the new green products is then compared with the simulated relatedness values (counterfactual relatedness), and will result in three possible scenarios [3, 20]:
 - If the distribution of the actual relatedness value stands fully to the right side of the counterfactual relatedness distribution, the actual relatedness values are higher than the simulated relatedness values, which means that new green products are developed based on the green products that a country is already capable of producing. This shows that the new green products are not randomly added to the country's green production basket; the path-dependence hypothesis is confirmed for all green products.
 - If the distribution of actual relatedness stands below the counterfactual relatedness distribution the non-path dependent hypothesis can be confirmed for all green products.
 - If the distribution of actual relatedness stands below and partially to the right side of the counterfactual relatedness distribution, then the path-dependent hypothesis can only be confirmed for the new green products with relatedness values that are higher than the randomly generated values, that is, new green products added to country's new green production basket followed the path-dependent and the non-path dependent process.

Finally, the number of new environmental related technologies that countries developed between 2007 and 2017 are compared with the previously determined

number of new green products to identify the relation between innovating new environmental-related technologies and a country's new green product development.

5 Results

The results are discussed in two sections: first, the developed GPS will be discussed in detail, and second, the patterns of green growth will be explained to explore how the United States and China diversified their green production basket as two of the leaders in green growth.

5.1 Green Product Space

Figure 1 depicts the GPS of 246 green products for all countries in 2017. The nodes are the 246 green products that are connected based on their proximity values (links).

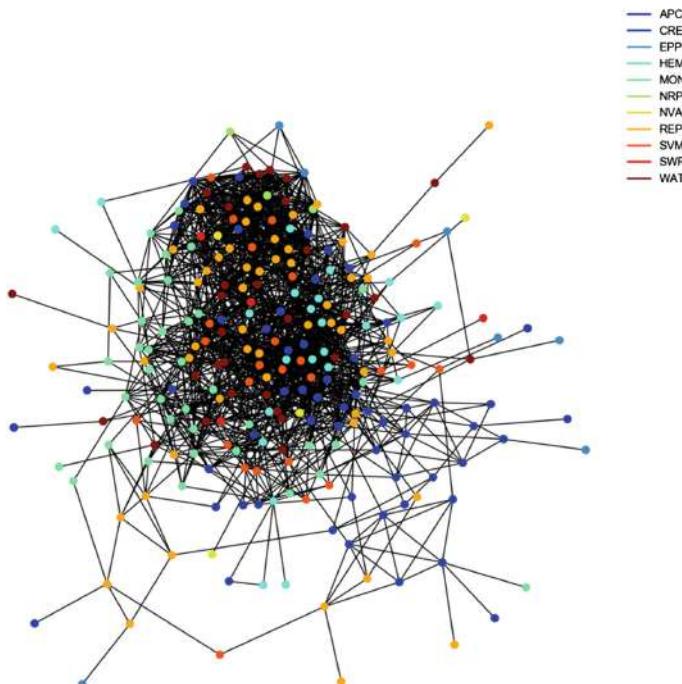


Fig. 1. The Green Product Space Network of 246 green products traded between all countries in 2017; the nodes' colors represent the different categories of green products, and are connected based on their proximity values.

The colors in the network represent the different categories of green products: the Air Pollution Products (APC), Environmental Monitoring, Analysis and Assessment Equipment (MON), Renewable Energy Plant (REP), Management of Solid and

Hazardous Waste and Recycling Systems (SWM) products, and Waste Water Management and Potable Water Treatment (WAT) products. According to the GPS network, green products that require more capabilities to produce are located in the high density area of the network, while green products with less required capabilities are located in the periphery of the network. This means that if a country's export basket stands in the high density area of the network, a country is well positioned to advance its green production basket since it has the capabilities to produce a wide range of products. For example, the GPS networks of the United States and China are visualized in Fig. 2 to represent their GPS and identify how they diversified their green production basket.

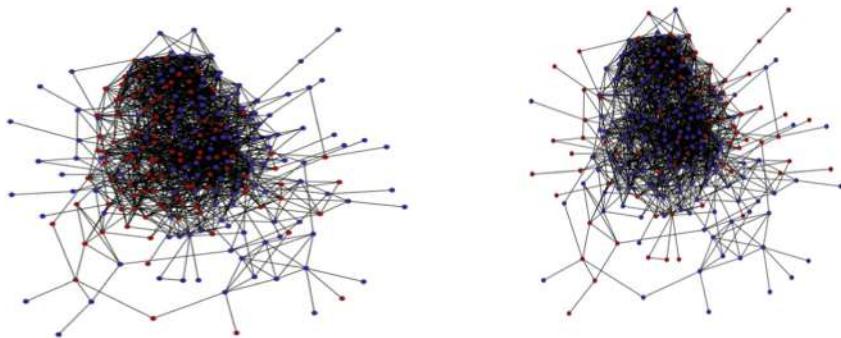


Fig. 2. The number of green products (red nodes) with high RCA ($RCA > 1$) that the United States (left) and China (right) were capable of producing in 2017; the United States' green production basket is more diversified than China, that is, the United States is better positioned to expand its green production basket than China

The red nodes in Fig. 2 show the green products that the United States and China produce with comparative advantage ($RCA > 1$). It can be seen that the United States has a more diversified green production basket (red nodes) in 2017 compared to China. However, we are considering how these countries developed their green production baskets, and thus how they became within the top leaders of producing green products. The next section discuss this in detail.

5.2 Pattern of Green Growth

In this section, the patterns of green growth in the United States and China are analyzed to understand how they expanded their green production basket, specifically, if they followed the path-dependent process or the non-path-dependent process. As depicted in Fig. 3, the actual relatedness distribution of new green product values (blue distribution) is similar to the counterfactual distribution of simulated green product values (orange distribution), that is, the United States' pattern of green growth does not follow [17]'s prediction of the path-dependent process for almost all green products, as only a few actual relatedness values are above the counterfactual relatedness values.

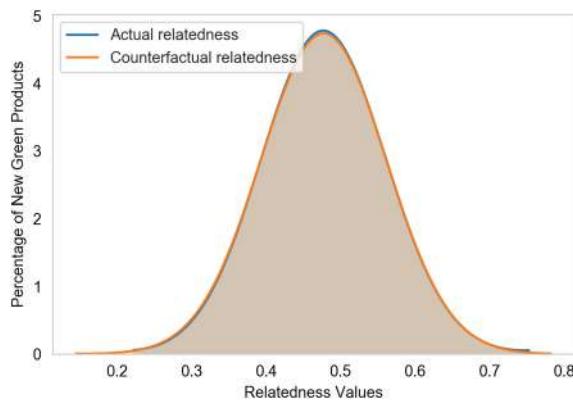


Fig. 3. Comparison of the United States' actual and counterfactual kernel distribution of relatedness between 2007 and 2017

This figure confirms that the existing green production capabilities in the United States did not have an effect on most of the growth of its green production basket. Similar to Figs. 3 and 4 depicts China's counterfactual relatedness distribution which stands slightly to the right side of the actual relatedness distribution, confirming that for most green products, the existing green production capabilities in China did not have an effect on the growth of its green production basket. Thus, China expanded its green production by jumping in its GPS, and followed the non-path dependence process as a result of structural changes in its green production capabilities for many of its green products.

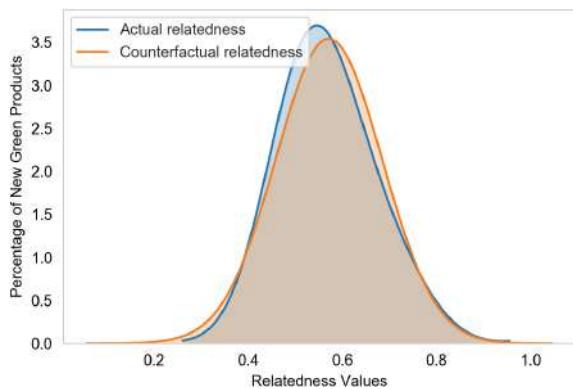


Fig. 4. Comparison of China's actual and counterfactual kernel distribution of relatedness between 2007 and 2017

According to these findings, the United States focused on innovating environmental related technologies to enhance its green production basket rather than depending on its

existing green production basket. China also focused on innovating new environmentally related technologies rather than expanding its previous capabilities to advance its green production basket, as most of the values in Fig. 4. were within the counterfactual distribution; for the relatedness values between 0.43 to 0.62, and 0.78 to 1, the actual relatedness values stand slightly above the counterfactual values. Thus, like the United States, China expanded its green production baskets by jumping in its GPS, and followed the non-path dependent process for many of its products. However, it should be noted that the United States followed the non-path dependent process for more of its products than China. We term this phenomena “high-energy jumping” since we conjecture that heavy investments are needed to achieve these jumps in the GPS.

6 Conclusion

This paper used the methodology proposed by [17] to construct the GPS for 246 green products. As mentioned, [17] constructed the GPS and discussed how countries expand their green production basket following the path-dependent process, that is, countries produce new green products based on their existing production capabilities. According to the literature, [3] proposed a method to test [13]’s prediction of path dependency in PS, and showed that countries follow both the path-dependent process and the non-path dependent process to expand their production baskets. However, this paper used [3]’s method to test the prediction of path-dependence proposed by [17] and tested if the United States and China followed the path-dependent process to expand their production baskets. The results of this research showed that the United States and China did not follow the path-dependent process to enhance their green production basket for almost all green products. In addition, we showed that China and the United States jumped in their GPS in the period of 2007 to 2017 to advance and diversify their green production basket. It can be concluded that the United States and China focused on innovating new environmental technologies, rather than expanding their green production basket based around their existing green production capabilities for this time period. We term the jumps as a result of this process “high-energy jumping”.

References

1. Alfredsson, E., Bengtsson, M., Brown, H.S., Isenhour, C., Lorek, S., Stevis, D., Vergragt, P.: Why achieving the Paris Agreement requires reduced overall consumption and production. *Sustain. Sci. Pract. Policy* **14**(1), 1–5 (2018)
2. Boschma, R., Iammarino, S.: Related variety, trade linkages, and regional growth in Italy. *Econ. Geogr.* **85**(3), 289–311 (2009)
3. Coniglio, N.D., Vurchio, D., Cantore, N., Clara, M.: On the Evolution of Comparative Advantage: Path-Dependent Versus Path-Defying Changes. *SERIES Working Papers N. 01/2018* (2018). <https://ssrn.com/abstract=3136471>
4. Dosi, G., Trancherob, M.: The Role of Comparative Advantage, Endowments and Technology in Structural Transformation (2019)
5. Duranton, G., Overman, H.G.: Testing for localization using micro-geographic data. *Rev. Econ. Stud.* **72**(4), 1077–1106 (2005)

6. Fraccascia, L., Giannoccaro, I., Albino, V.: Green product development: what does the country product space imply? *J. Cleaner Prod.* **170**, 1076–1088 (2018)
7. French, S.: Revealed comparative advantage: what is it good for? *J. Int. Econ.* **106**, 83–103 (2017)
8. Fresner, J.: Cleaner production as a means for effective environmental management. *J. Cleaner Prod.* **6**(3–4), 171–179 (1998)
9. Ghisellini, P., Cialani, C., Ulgiati, S.: A review on circular economy: the expected transition to a balanced interplay of environmental and economic systems. *J. Cleaner Prod.* **114**, 11–32 (2016)
10. Hamwey, R., Pacini, H., Assunção, L.: Mapping green product spaces of nations. *J. Environ. Dev.* **22**(2), 155–168 (2013)
11. Hausmann, R., Hidalgo, C.A., Bustos, S., Coscia, M., Simoes, A., Yildirim, M.A.: The Atlas of Economic Complexity: Mapping Paths to Prosperity. MIT Press, Cambridge (2014)
12. Hausmann, R., Klinger, B.: The structure of the product space and the evolution of comparative advantage (No. 146). Center for International Development at Harvard University (2006)
13. Hidalgo, C.A., Klinger, B., Barabási, A.L., Hausmann, R.: The product space conditions the development of nations. *Science* **317**(5837), 482–487 (2007)
14. Hidalgo, C., Hausmann, R.: The building blocks of economic complexity. *Proc. Natl. Acad. Sci.* **106**, 10570–10575 (2009)
15. Islam, M.N., van Amstel, A. (eds.): Bangladesh I: Climate Change Impacts, Mitigation and Adaptation in Developing Countries. Springer International Publishing, Cham (2018)
16. Laursen, K.: Revealed comparative advantage and the alternatives as measures of international specialization. *Eurasian Bus. Rev.* **5**(1), 99–115 (2015)
17. Mealy, P., Teytelboym, A.: Economic Complexity and the Green Economy (2017). <https://ssrn.com/abstract=3111644>
18. Sauvage, J.: The Stringency of Environmental Regulations and Trade in Environmental Goods. OECD Trade and Environment Working Papers, No. 2014/03. OECD Publishing, Paris (2014)
19. Simoes, A.J.G., Hidalgo, C.A.: The economic complexity observatory: an analytical tool for understanding the dynamics of economic development. In: Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence (2011)
20. Talebzadehhosseini, S., Scheinert, S.R., Rajabi, A., Sacidi, M., Garibay, I.: The pattern of economies green growth: the role of path dependency in green economy expansion. *arXiv: 1906.05269* (2019)



Performance of a Multi-layer Commodity Flow Network in the United States Under Disturbance

Susana Garcia¹, Sarah Rajtmajer^{2,3(✉)}, Caitlin Grady^{1,3},
Paniz Mohammadpour¹, and Alfonso Mejia¹

¹ Department of Civil and Environmental Engineering,
The Pennsylvania State University, University Park, PA 16802, USA
{cgrady,aim127}@psu.edu

² College of Information Sciences and Technology,
The Pennsylvania State University, University Park, USA
smr48@psu.edu

³ The Rock Ethics Institute, The Pennsylvania State University,
University Park, USA

Abstract. Network-based analyses have furthered global understanding of supply chain and commodity trade networks between countries. Much of the previous work in this area has focused on analyzing economic sectors separately, aggregating sectors neglecting inter-sectoral connectivity, or representing trade at a national scale losing intrastate connections and impact. We further previous work by constructing and analyzing the intrastate input-output multi-layer network of the United States commodity and service sectors. We subject the network to perturbations and find that the government services sector represents the most influential sector as it generates the most impact when shocked. Taking this sector as exemplar, we showcase how impact varies differentially across regions, and how impact compares to other measures of resilience.

Keywords: Multi-layer · Supply chain · Input-output · Shock

1 Introduction

The global economy is increasingly reliant on long-distance complex transfers of goods and services. Perturbations to these networks have the potential to impact various sectors across political and regional boundaries in ways that are not fully understood. Previous work suggests that economic shock outcomes strongly depend on the nature of the shock and country size [1,2]. Thus there is a need to build greater understanding around emergent behaviors of trade networks in response to multiple scales of impacts to enable development of predictive tools and policy instruments that can mitigate fluctuations that lead to systematic crises.

Many previous efforts utilizing networks to understand trade flow risk have focused on country to country interactions. Network theory has been used to investigate the topological structure and evolution of aggregated versions of international commodity networks [3–5] and intersectoral connectivity of regions through input-output linkages [1,6]. Input-output linkages provide the ability to examine the interdependence of economic sectors and regions as well as their dependence on primary inputs [1,6]. Studies have shown that input-output linkages can have a significant effect on the frequency and depth of economic perturbations [9]. In addition to considering input-output linkages, how these networks are constructed has been shown to play a role on structures and network evolution. Building upon previous work [6], Alves et al., [7] represented the World Input-Output Database as a single-layer, multiplex, and multi-layer network where they found that when investigating strength and entropy, the multiplex and multi-layer network construction enabled higher resolution understanding of heterogeneity in intra- and cross-industry transactions. Moreover, they have also extended the notion of nestedness in multi-layer trade networks showcasing how nested structures differ significantly from multi-layer null models across international trade [8].

While these efforts have built greater understanding around network properties and responses to perturbations at a country to country scale, there remains a need to analyze subnational, interregional, and intersectoral dependencies. These dependencies shape the process of diffusion of perturbations and may be more directly relevant to policy or predictive tool instruments that can mitigate or soften implications of large shocks.

In this study we characterize the topology and upstream diffusion of shocks in the US subnational multiregional input-output (MRIO) network. We employ metrics from complex network theory and a diffusion model that captures the complexity of economic connectivity. The main objectives of this study are: (i) to characterize the connectivity of the US subnational MRIO network, (ii) and to assess the impact caused by national and regional shocks on the demand of a specific economic sector. Previous efforts in this area have focused on studying economic sectors separately, aggregated sectors neglecting inter-sectoral connectivity, or represented trade at a national scale losing intrastate connections and impact, thus this paper contributes to the literature through developing and analyzing an economic multi-layer network at a subnational scale. The subnational MRIO is based on the regionalized input-output tables developed by Garcia [10], which rely on open source US national input-output tables [11] and commodity trade data [12].

2 Methods

We develop a multi-layer network representation of the commodity flows within the United States, based on a multiregional Input-Output model [10]. The MRIO network consists of 4,838 nodes, representing 118 geographical regions within the United States and 41 commodity groups traded among regions. That is, each

node is representative of a region/commodity pair. The commodity groups represent 41 sectors covering agricultural commodities, manufactured food, industrial commodities and services (Table 1). The regions were created from the original Freight Analysis Framework (FAF) zones [12] but we reduced the number of regions from 132 domestic regions in the FAF database to 118 regions by combining cities with their entire Metropolitan Statistical Areas (MSAs). The regions account for 70 MSAs and 48 State remainders, that is - the area that falls outside of the MSAs which accounts for the remaining land within the State. A detailed list and visualization of these regions can be found through Garcia's previous work [10].

Table 1. Commodity groups represented by the input-output multi-layer network

Commodity groups			
Live animals	Metallic ores	Paper products	Furniture
Cereal grains	Coal	Printed products	Misc. manufactured
Fruits, vegetables	Crude, gasoline, etc.	Textiles, leather	Utilities services
Animal feed	Coal (other)	Nonmetal minerals	Wholesale services
Meat	Basic chemicals	Base metals	Retail services
Milled grain	Pharmaceuticals	Base metals	Transport services
Manufactured food	Fertilizers	Machinery	Food services
Alcoholic beverages	Chemical prods.	Electronics	Government services
Tobacco products	Plastics and rubber	Vehicles	
Building stones etc.	Logs	Transport equpt.	
Nonmetallic minerals	Wood products	Precision insrmts.	

2.1 Network Representation and Characterization

We rely on the standard input-output (IO) model to obtain the weights of the weighted, directed network. The model starts with the basic input-output relationship:

$$x = Ax + y \quad (1)$$

where x , A , and y represent the domestic output, the technical coefficients, and final demand of commodities, respectively. This leads to a well-known input-output relationship:

$$x = (I - A)^{-1}y = Ly = \text{Total Requirements} \quad (2)$$

where L is the Leontief inverse matrix which provides the requirements of the system to satisfy one unit of final demand of each product through whole supply chains. The total requirements from the system to satisfy the final demand y is referred as the Total Requirements matrix.

We used the Total Requirements matrix as the weighted adjacency matrix to build the network. In the multi-layer representation, each commodity group is

represented by a layer in the network (41 layers). Intra-layer connections represent transactions between regions that originate and terminate in the same economic sector. While, inter-layer connections represent transactions that occur amongst regions across different economic sectors. Weighted, directed links between regions represent monetary flows in US dollars. The elements of the weighted adjacency matrix can be defined as $w_{r,s}^{i,j}$ to represent the monetary flow from economic sector i in region r to economic sector j located in region s .

From this multi-layer configuration, we can aggregate weights to obtain a mono-layer configuration as:

$$w_{r,s}^{Mono} = \sum_i \sum_j w_{r,s}^{Multi[i,j]}. \quad (3)$$

That is, the mono-layer configuration represents total flow, irrespective of commodity, amongst the 118 geographical regions described. Following, we characterize the disaggregated (multi-layer) and aggregated (mono-layer) networks using standard network metrics, to gain insight into the structure of these flows.

2.2 Network Perturbation

To better understand the complex connectivity of the disaggregated economic system, we simulated impacts on production caused by a decrease in the demand for a given commodity. From Eq. 2:

$$\Delta Total Requirements = L \Delta y \quad (4)$$

We applied a shock to the demand of government services at each of the 118 regions and quantified the number of impacted regions and sectors. In this model, the sectoral production (Total Requirements) linearly depends on the demand from all sectors and regions in the economy. Applying a shock on the final demand of a sector allows us to study the connectivity without affecting the technological coefficients of the system (i.e. the share of contribution of each commodity to the inputs in an economic sector).

Due to the linear dependence between the demand and the production in this model, we assess the spread of the shock by quantifying the number of regions and sectors impacted with the application of an arbitrary shock. Different magnitudes of shock would result in the same number of regions and sectors impacted.

3 Results

3.1 Characterization of the Network

As described, the global, disaggregated (multi-layer) network consists of 4838 nodes, namely the 4838 region/commodity pairs representing the pairwise combination of 118 regions and 41 commodities. The network is in general very dense,

with 20,958,897 directed edges present in the network, of a total 23,406,244 possible edges (density 0.8954). The high density of this network indicate widespread dependencies amongst economic sectors (commodities and services) and regions. While only 0.8% of the edges are intra-regional flows, $w_{r,s|r=s}^{i,j}$, these flows represent 57% of the monetary flows. This finding reflects the high reliance of regions on locally produced commodities and services. While similar results have been found for international trade networks, we separate the flows of the multi-layer network between intra-layer and inter-layer flows and found that 11% of the monetary flows occur at the same layer $w_{r,s}^{i,j|i=j}$ (i.e., two regions exchanging a commodity within the same economic sector), while 89% of the monetary flows occur between two different layers $w_{r,s}^{i,j|i \neq j}$. These strong dependencies between different economic sectors suggest that a shock on a certain economic sector has the potential to influence the network flows beyond what is captured through international analyses. The disaggregated network follows a power law degree distribution, consistent with many other real-world systems. The sum total commodity flow value over the network (network strength) is 1.904×10^7 monetary units, measured in millions of U.S. Dollars. Globally, the disaggregated network contains one very large, strongly connected component composed of 4519 nodes (region/commodity pairs); the entire network is weakly connected.

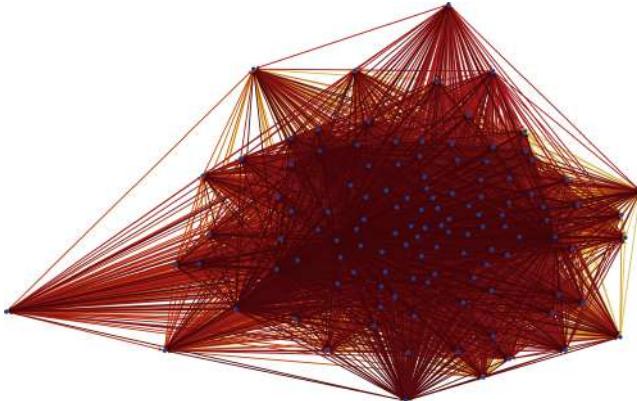


Fig. 1. The regionally aggregated (mono-layer) commodity flows network on 118 nodes. Edges are colored to indicate edge weights, with yellow and dark red representing least (the directed edge from “Rest of NE” to “Rest of HI”, weight 0.1457) and greatest (the self-loop on Los Angeles, weight 819,507), respectively.

Mono-layer Network Representation. The regionally aggregated (mono-layer) network represents the aggregate flow (over all commodities) amongst our 118 regions (see Eq. 3). The aggregated network has complete graph structure, with all 13,924 edges present. This confirms fully-connected regional dependency, as we expect. Accordingly, the average clustering coefficient of the system is negligible at 0.00046. Figure 1 depicts the directed, weighted regionally aggregated

Table 2. Regions with highest in- and out-degree in the regionally aggregated network, rounded to the nearest integer.

Highest regional degree rankings			
Region	In-degree	Region	Out-degree
Los Angeles	1,190,447	Los Angeles	1,090,701
New York City	1,136,992	New York City	1,032,680
Chicago	788,379	Houston	842,459
Dallas	715,103	Chicago	811,688
Houston	554,177	Dallas	648,699
“Rest of TX”	471,857	“Rest of TX”	542,964
Atlanta	423,894	San Francisco	389,965
San Francisco	411,736	“Rest of PA”	378,927
“Rest of PA”	393,765	Boston	374,638
Philadelphia	384,080	Philadelphia	372,422

(mono-layer) network. A listing of regions with highest in- and out-degree is provided in Table 2. Of note, and consistent with what we observe in the multi-layer formulation, the highest-weighted edges in the mono-layer network are intra-regional flows. In fact, the 100 highest-weighted edges in the mono-layer network are intra-regional, together totaling 1.071×10^7 of 1.904×10^7 units (56%) of network flow. This finding is consistent with our observations in the multi-layer case. We analyzed the presence of communities in the mono-layer network through a modularity optimization algorithm and found that the network presents four communities that are geographically constrained (i.e., communities represent the four geographical mega-regions: North Atlantic, Midwest, West, South Atlantic, and Southwest). Such communities suggest that shocks would spread more easily to nearby locations than to distant ones.

3.2 Network Perturbations Across Sectors

We explored the impacts associated with the shock applied to the national and regional demand of a specific economic sector. This analysis highlighted the importance of the government services sector as the sector that generates the most impact after perturbation. At a national level, the government services sector directly requires input from more than 90% of the 41 economic sectors analyzed and uses about 15% the aggregated output of the national economy as an input for its operations. In investigating the connectivity of the government services sector we found the most influential first-order suppliers to be services and commodities including as manufactured food, meat products, printed products, gasoline, and transportation equipment (Fig. 2). From a consumer perspective, services such as this government sector are not always considered, yet this work suggests this multi-layer input-output network analysis has the potential

to unlock greater understanding around interdependent sectors. For example, the indirect dependencies show that the government sector depends indirectly on upstream sectors such as coal, plastics and rubber, chemical products, and textiles.

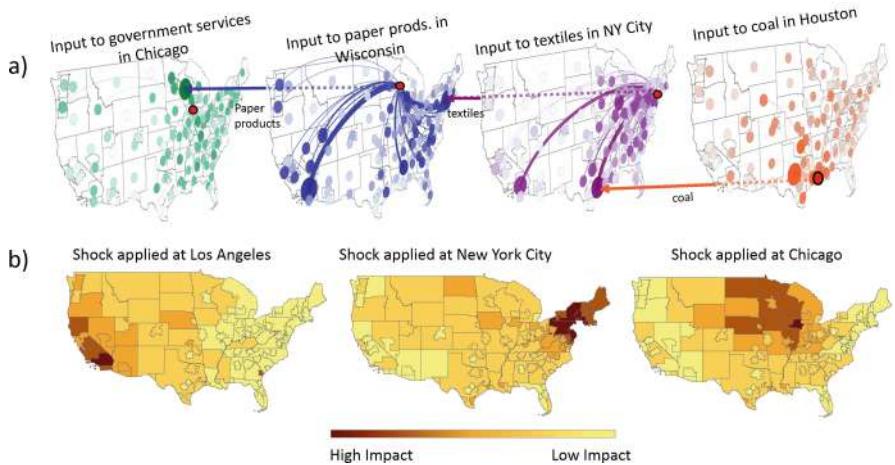


Fig. 2. Distant inter-regional and inter-sectoral connections possible along the supply chain of government services (Panel A) and the localized impact associated with a regional shock on government services applied to the cities of Los Angeles, New York City, and Chicago (Panel B)

3.3 Impact to Government Services Sector

To further unpack the impact to government services, Fig. 2 shows the distant connections possible along the supply-chain of government services (Panel A). Even though the network is highly connected and complex supply-chains allow a region to indirectly demand inputs from distant regions, the shock applied on the demand of government services in specific locations results in an aggregated impact that is more pronounced in nearby locations than in distant ones. Figure 2 illustrates the regional impact associated with the shock applied to three cities: Los Angeles, New York City, and Chicago (Panel B).

Figure 3 displays the size of the impact associated with the shock applied at each of the 118 regions (Panel A). The size of the impacts is measured by the number of regions and sectors impacted. A shock applied at the regions with largest population such as Los Angeles, New York City, and Chicago generate the largest aggregated national impact. They also impact the largest number of regions and economic sectors. Figure 3 demonstrates the impact caused at each region by applying a nationwide shock on the demand of government services (Panel B). The impact is measured by how many regions generate an impact on the region in question and the number of sectors impacted by the shock.

A general trend is shown where the regions with largest population such as Los Angeles, New York, City, and Chicago are impacted by a lower number of regions. Also, the regions with the largest number of sectors impacted are the regions impacted by the lowest number of demanding regions.

The size of the impacts associated with shocks applied at regions or nationwide can be interpreted as a centrality metric of the MRIO network. The case of Atlanta, as seen in Fig. 3 (Panel B), indicates how a region can be impacted by few demanding regions; however, it is the region where the highest number of sectors that could be impacted simultaneously. On the other hand, smaller regions, in terms of population and aggregated contribution the government services sector, are impacted by the largest number of regions for a small number of sectors. The remainder of the state of Utah could be regarded as a highly central region supplying substantial inputs to various regions.

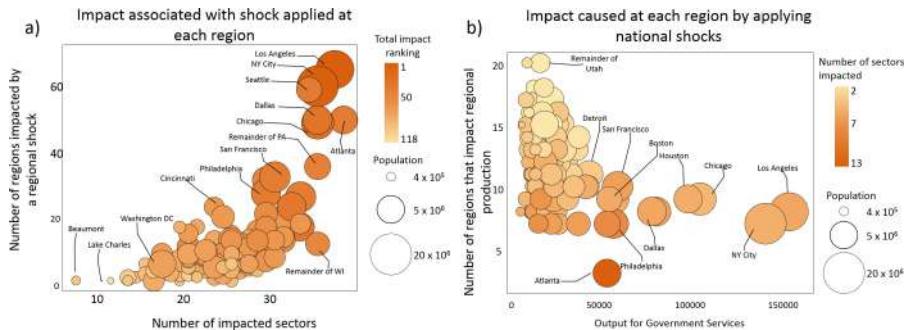


Fig. 3. Regional analysis of impact caused by shocks. Panel A indicates the impact associated with shock applied at each region. Panel B shows the impact caused at each region by applying national shocks.

4 Discussion

Limitations. Several limitations exist that could be further refined for future study. For one, all of the results within this network analysis are based on the value of a commodity through monetary units of analysis (U.S. Dollar). While this is often used in economics studies, monetary units serve as a coarse estimate of value and would likely not be the most important factor when designing policy interventions. Other units of measure such as environmental extensions in water or carbon, for example, have provided useful understandings of sustainability implications for supply chain networks [10, 19, 20]. Another challenge stems from the aggregation of commodity flow data to an annual timescale which limits extension of perturbations that occur at a short timescale resolution. Future work could also further understanding between national input-output multi-layer networks and sub-national multi-layer networks.

Ethical Implications. An under-explored area of shock and risk-related network analyses is the ethical implications of unequal distributions of risk. Numerous theorists across fields of geography, business, and philosophy have reflected on ethical considerations of work, labor, and equity across various economic constructs [13–15]. In the field of supply chain management, the ethical implications of risk often refer to ethical business conduct issues [16,17]. We, however, conceptualize ethical implications beyond its current scope and see room for exploration of unequal distribution of shock and risk impacts. Using the Urban Adaptation Assessment database [18], we explored the relationship between our impact findings and regional adaptability. Adaptability, in this context, consists of key measures of risk and readiness. The Urban Adaptation Assessment defines risk as a measure of a city's vulnerability to climate change using information on exposure, sensitivity, and adaptive capacity while readiness incorporates measures of economic condition, governance, and social capacity [18]. In these cases, a higher measure of risk is negative while a higher measure of readiness is positive. As an example, Chicago and the state remainder of Pennsylvania ("Rest of PA") have two of the highest hazard risks (Fig. 4, Panel A). If taken to represent a higher likelihood of climate related challenges this may imply that while the overall value of impact is lower, it is more likely to occur thus representing a higher risk compared to Houston and Atlanta which share similar impacts but have lower hazard risk indices. Interestingly, "Rest of PA" shows up amongst the highest indegree nodes on the regional net. This first order attempt at coupling and investigating unequal distribution of implications from perturbations showcases value in future pursuits that bridge the disciplines of complex network science and ethics.

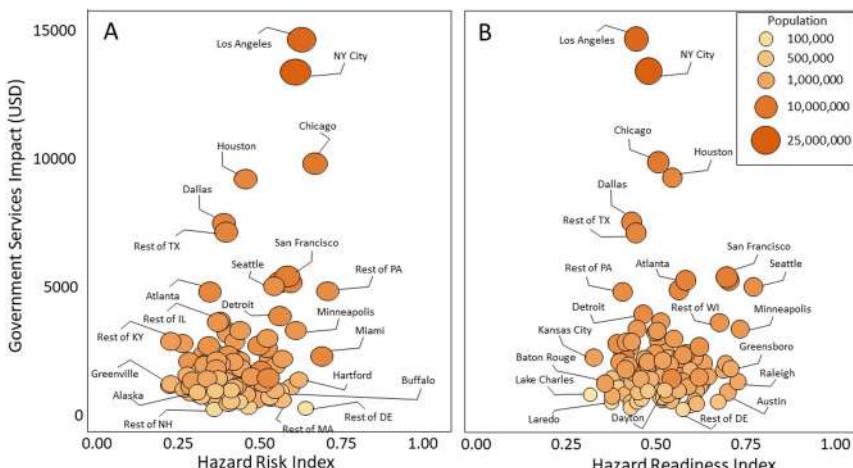


Fig. 4. Risk and readiness index. Panel A indicates the relation between government services impact and Hazard Risk Index. Panel B shows the relationship between government services impact and Hazard Readiness Index.

Summary. We study the impacts of shock propagation through a multiregional input-output commodity flow model organized through a multi-layer network. This paper explored a multi-layer trade flow network with detailed sub-country level interactions furthering previous literature on trade shock analysis. Overall, the main findings from this study are the following: the complex connectivity in the US subnational input-output economic network can be described in terms of complex networks. The network presents characteristics that are shared with other spatially embedded networks. Despite its dense connectivity, the network is likely heavily influenced by distance. The effect of distance in this spatially embedded network can be confirmed by the presence of network communities that are geographically constrained. Such communities suggest that shocks would spread more easily to nearby locations than to distant ones. Finally, the impacts generated after applying a shock on the demand of government services provide insights on the importance of regions for the spread of impacts.

Acknowledgements. Authors would like to acknowledge Venkat Ashish Kumar Simhachalam for assistance with Fig. 1 and Tasnuva Mahjabin for assistance with Fig. 4. Drs. Rajtmajer and Grady gratefully acknowledge seed funding from the Rock Ethics Institute at The Pennsylvania State University.

References

1. Contreras, M.G.A., Fagiolo, G.: Propagation of economic shocks in input-output networks: a cross-country analysis. [ArXiv:1401.4704](https://arxiv.org/abs/1401.4704) (2014)
2. Lee, K.M., Goh, K.-I.: Strength of weak layers in cascading failures on multiplex networks: case of the international trade network. *Sci. Rep.* **6**(1), 26346 (2016)
3. Bhattacharya, K., Mukherjee, G., Saramäki, J., Kaski, K., Manna, S.S.: The international trade network: weighted network analysis and modelling. *J. Stat. Mech. Theory Exp.* **02**, P02002 (2008)
4. Fagiolo, G., Reyes, J., Schiavo, S.: The evolution of the world trade web: a weighted-network analysis. *J. Evol. Econ.* **20**(4), 479–514 (2010)
5. Sartori, M., Schiavo, S.: Connected we stand: a network perspective on trade and global food security. *Food Policy* **57**, 114–127 (2015)
6. Cerina, F., Zhu, Z., Chessa, A., Riccaboni, M.: World input-output network. *PLoS ONE* **10**(7), e0134025 (2015)
7. Alves, L.A., Mangioni, G., Rodrigues, F., Panzarasa, P., Moreno, Y.: Unfolding the complexity of the global value chain: strength and entropy in the single-layer, multiplex, and multi-layer international trade networks. *Entropy* **20**(12), 909 (2018)
8. Alves, L.A., Mangioni, G., Cingolani, I., Rodrigues, F.A., Panzarasa, P., Moreno, Y.: The nested structural organization of the worldwide trade multi-layer network. *Sci. Rep.* **9**(1), 2866 (2019). <https://doi.org/10.1038/s41598-019-39340-w>
9. Acemoglu, D., Ozdaglar, A., Tahbaz-Salehi, A.: The Network Origins of Large Economic Downturns. National Bureau of Economic Research, Cambridge, MA (2013)
10. Garcia, S.: Connectivity in the U.S. hydro-economic network: towards consistent environmental accounting of national consumption. Ph.D. thesis, Penn State University (2019)

11. Bureau of Economic Analysis: Input-Output Accounts Data (2018). <https://www.bea.gov/industry/input-output-accounts-data>
12. Federal Highway Administration and Bureau of Transportation Statistics: Freight Analysis Framework Version 4 (2015). <https://faf.ornl.gov/fafweb/Default.aspx>
13. Popke, J.: Geography and ethics: non-representational encounters, collective responsibility and economic difference. *Prog. Hum. Geogr.* **33**(1), 81–90 (2009)
14. Maloni, M.J., Brown, M.E.: Corporate social responsibility in the supply chain: an application in the food industry. *J. Bus. Ethics* **68**(1), 35–52 (2006)
15. Drake, M.J., Schlachter, J.T.: A virtue-ethics analysis of supply chain collaboration. *J. Bus. Ethics* **82**(4), 851–864 (2008)
16. Hofmann, H., Busse, C., Bode, C., Henke, M.: Sustainability-related supply chain risks: conceptualization and management. *Bus. Strategy Environ.* **23**(3), 160–172 (2014). <https://doi.org/10.1002/bse.1778>
17. Tang, O., Nurmaya Musa, S.: Identifying risk issues and research advancements in supply chain risk management. *Int. J. Prod. Econ.* **133**(1), 25–34 (2011). <https://doi.org/10.1016/j.ijpe.2010.06.013>
18. Notre Dame Global Adaptation Initiative. Urban Adaptation Assessment (2019). <https://gain-uaa.nd.edu/?referrer=gain.nd.edu>
19. Dalin, C., Konar, M., Hanasaki, N., Rinaldo, A., Rodriguez-Iturbe, I.: Evolution of the global virtual water trade network. *PNAS* **109**(16), 5989–5994 (2012)
20. Sartori, M., Schiavo, S., Fracasso, A., Riccaboni, M.: Modeling the future evolution of the virtual water trade network: a combination of network and gravity models. *Adv. Water Resour.* (2017). <https://doi.org/10.1016/j.advwatres.2017.05.005>



Empirical Analysis of a Global Capital-Ownership Network

Sammy Khalife^(✉), Jesse Read, and Michalis Vazirgiannis

LIX, CNRS, Ecole Polytechnique, Institut Polytechnique de Paris,
91128 Palaiseau, France
{khalife,jread,mvazirg}@lix.polytechnique.fr

Abstract. Ownership relationships between legal entities can be represented as a large directed and weighted graph. This paper provides a methodology and an empirical analysis of such network, composed of millions of nodes and edges. To do so, we employ a variety of metrics from graph analytics and algorithms from influence maximization (IM). For reasons of confidentiality, our empirical analysis is carried out on aggregation at country and sector level, analysing in details the case of France. Our results offer new type of intuitions and metrics in this area by highlighting the existence of strong communities of capitalistic property. Finally, we discuss influence maximization methods as means to evaluate an entity impact in the socialistic graph.

Keywords: Complex networks · Legal entities · Capitalistic graphs · Centrality measures · Graph degeneracy · Influence maximization

1 Introduction

A legal entity is a juridic term to designate individual, company, or organization that has legal rights and obligations. The nature and clarity of the information describing legal entities is important for compliance standards. Several type of interactions can exist between these entities, for instance payments or capitalistic property. Graph analysis has been proposed for the interbank market [3] and interbank payment flows [12]. A study on systemic risk in the interbank network, where payment interactions are treated as a complex network was described in [8]. Recent work attempted to describe the actual topologies observed in the financial system [7].

In recent years, graph representations have become ubiquitous in several domains (social networks, protein and gene regulatory networks, and textual documents, among other areas) inciting considerable progress during the last decades in understanding the structure and operation of complex networks [1]. In this work, we propose a methodology based on recent progress in this field to analyze a worldwide capitalistic ownership graph, and provide results on industrial data. After a standard analysis using centrality measures, we propose a influence maximization algorithm in order to target specific nodes that have

high importance in the ownership graph. To the best of our knowledge, this has not been proposed in recent work.

2 Ownership Network and Methodology

We denote a capitalistic graph $G = (V, E)$. A vertex $i \in V$ represents a legal entity, an oriented edge e_{ij} with weight w_{ij} represents capitalistic property: an edge from i to j means that i owns capital of j in proportion w_{ij} . \mathcal{N}_i^+ and \mathcal{N}_i^- are respectively the out-neighbors and in-neighbors of vertex i (i.e., set of nodes connected to and from the vertex). $|I|$ is the cardinal of a finite set I . For mathematical convenience, we impose the following conventions:

$$\begin{aligned} \forall e \in E, w_e &\in [0, 1] \\ \forall i \in V, \sum_{j \in \mathcal{N}_i^-} w_{ji} &\leq 1 \end{aligned} \tag{1}$$

Following standard juridic terms, there are two types of entities in the graph. The first type of entity is called Natural person (or physical person), and is an individual human being. The second one is called Juridical person, and represents incorporated organizations including corporations, government agencies; or non-governmental organizations. A legal person is composed of natural persons, but has a distinct juridic identity.

We suppose that there exist no inner link between the subgraph of natural people, even though natural person can have influence over others (we suppose this link is not explicit so it is not considered as an edge in the graph). Therefore, it is possible to model this ownership graph as a bipartite graph in Fig. 1.

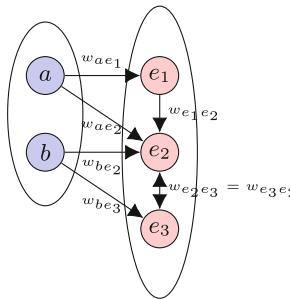


Fig. 1. Bipartite directed weighted graph between natural person (blue) and organizations (red)

We are interested in studying the influence of legal entities, so the distinction between natural entities and organizations raises an important discussion for the evaluation of such influence. A first approximation is to consider that the links

between person and organizations are negligible in terms of global influence. In this case, we are left with the study of the subgraph of the organizations. A second way to operate is to aggregate information from the edges between natural entities and organizations which have been removed, for example by creating additional links between organizations, or adding node features. In view of the instance at our disposal, and in particular the fact that the data is poor with regard to personal attributes (few details), we decided to consider the first situation, that is to say to keep only the links between organizations.

Unlike market payment graphs which are very dynamic, the capitalistic graph evolves across a relatively long time scale and significant changes do not occur frequently (except rare events, such as the onset of an economic crisis). Therefore, we consider the capitalistic graph constant for our analysis: we will not analyze its dynamic aspect, although this may be considered in future work. We provide a description of the network instance at our disposal in Sect. 3. In the next Sects. 2.1 to 2.5, we present the four components of our analysis in order to describe the topology of the graph, and to measure influence of entities:

- Degree distribution and connected components
- Centrality measures and K-cores
- Rooted influence graph
- Aggregation by attribute
- Influence maximization

In Sect. 3, we provide the results of the analysis, with respect to location and sector.

2.1 Degree Distribution and Connected Components

In the following A is the adjacency matrix of the network of the graph $G = (V, E)$.

$$A_{ij} = \begin{cases} w_{ij} & \text{if } e = (i, j) \in E \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The degree distribution is a fundamental and important measure in complex networks, because it allows to classify graphs according to the type of distribution (e.g power law graphs [11]). In the following, when not specified, the degree of a vertex we will refer to its degree of the corresponding undirected graph, in and out degrees of a vertex i are respectively the number of its in and out neighbors (i.e number of non-zero coefficients on the column and row i of A). A component $W \subseteq V$ of G is said connected if for each i and j in W , there exists a path from i to j .

2.2 Degeneracy and Centrality Measures

A graph is said to be k -degenerate if and only if every subgraph has a vertex of degree at most k . The degeneracy of a graph is defined as the smallest value of

k for which it is k -degenerate. As such, it is a measure of sparsity. In order to enrich this notion, let us define the notion of k -core for an undirected graph. A k -core of an undirected graph is defined as the maximal subgraph $C_k \subset G$ such that each node has degree at least k in this subgraph:

$$C_k = \{i \in V \mid |C_k \cap \mathcal{N}_i| \geq k\} \quad (3)$$

If C_k is not empty, then G has degeneracy at least k since there exists a subgraph that contains nodes of degree k . Therefore, the degeneracy of G is the largest k for which C_k is not empty. The core number of a vertex is the maximum k for which he belongs to C_k . The notion of k -core extends to directed graphs, with in and out cores (i.e., \mathcal{N}_i becomes \mathcal{N}_i^- and \mathcal{N}_i^+ respectively). There exist linear-time algorithms to compute k -cores [2].

2.3 Rooted Influence Graph

For each entity, we define a subgraph, the rooted influence graph (RIG), similarly to the rooted citation graph (RCG) in collaboration analysis [6]. The RIG of a vertex i is the subgraph of G induced by the set of vertices that contain i and all the vertices which can be reached by a directed path. That is, $j \in \text{RIG}(i)$ if and only if there is a directed path from vertex i to vertex j . The resulting directed acyclic graph (DAG) contains all the entities that are directly or implicitly influenced by the entity i . Based on this definition, it is natural to consider out-degree of the root, average degree in the RIG, and core number of the root in its RIG to measure influence of an entity.

2.4 Aggregation by Attributes

The ownership graph also contain entity attributes (location (country or region) and a description of the activity (sector), cf. Sect. 3.1 for a precise description). Here, we present a method to analyze the graph of entities by attributes. Let \mathcal{A} be the set of values for a given attribute of the entities. For $(a, b) \in \mathcal{A}^2$, let G_a (resp. G_b) be the set of entities having attribute a (resp. b). We define a new graph $G_{\mathcal{A}} = (\mathcal{A}, E_{\mathcal{A}})$ between attribute values in the following way:

$$w_{ab} = \frac{1}{|G_b|} \sum_{j \in G_b} \sum_{i \in G_a \cap \mathcal{N}^-(j)} w_{ij} \quad (4)$$

In other words, $G_{\mathcal{A}}$ provides a kind of “meta-graph” based on the pairwise relationship between the values of \mathcal{A} . For example, a graph where $\mathcal{A} = \{a, b\}$ defines two countries, will be a graph of two nodes, inheriting the connectivity in the form of an aggregation w_{ab} on up to two directed edges $E_{\mathcal{A}}$. This definition of Eq. (4) insures the following mathematical conveniences, as also in Eq. (1):

$$\forall(a, b) \in \mathcal{A}^2$$

$$0 \leq w_{ab} \quad (5)$$

$$w_{ab} \leq \frac{1}{|G_b|} \sum_{j \in G_b} \sum_{i \in G_a \cap \mathcal{N}^-(j)} w_{ij}$$

$$w_{ab} \leq 1 \quad (6)$$

and

$$\sum_{a \in \mathcal{A}} w_{ab} = \frac{1}{|G_b|} \sum_{j \in G_b} \sum_{a \in \mathcal{A}} \sum_{i \in G_a \cap \mathcal{N}^-(j)} w_{ij} = \frac{1}{|G_b|} \sum_{j \in G_b} \sum_{i \in \mathcal{N}^-(j)} w_{ij}$$

gives:

$$\sum_{a \in \mathcal{A}} w_{ab} \leq \frac{1}{|G_b|} \sum_{j \in G_b} 1 \leq 1 \quad (7)$$

Moreover, the choice of Eq.(4) is a good candidate for the influence of a attribute a over an attribute b (e.g influence of Germany over France), since it corresponds to a quantity representing the percentage or total capital owned by a over b . Therefore, The meta-graph $G_{\mathcal{A}}$ allows us to analyze interactions between attributes (i.e countries or sectors). However the limitation of this analysis lies in the assumption that entities have the same capital: we will come back to this limitation in Sect. 3.

2.5 Influence Maximization

In network and graph theory, influence maximization (IM) is the problem of maximizing influence with regards to seed nodes using a diffusion model. It has been extensively studied recently due to its potential commercial value. An example of application of influence maximization is viral marketing [5], where an organization wants to spread the adoption of a product from selected adopters. Influence maximization is also the corner stone in other important applications such as network monitoring, rumor control, and social recommendation. For more details, we refer to a recent survey [9]. In this work, we will use one of the Monte-Carlo based method to approximate a solution of the problem, Influence maximization based on Martingales [13].

3 Results

3.1 Description, Degree Distribution and Connected Components

Orbis is a database composed of about one hundred million entities, developed by a specialized group based in Bureau Van Dijk's Brussels office, aggregating several sources of data [4]. The database also specifies indirect ownership though

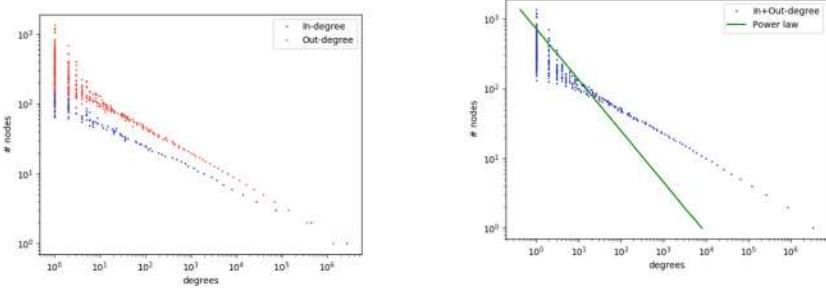
a path of auxiliary entities, but this is not strictly capital ownership, and therefore we do not take into account these relationships in our study. After pre-processing, the graph of non isolated entities is composed of 39,398,321 nodes and 80,874,728 edges. Following our discussion in the first section, we consider the subgraph of juridical entities (organizations) for our entity influence analysis. The subgraph of non-isolated organizations is composed of 6,516,332 nodes and 6,670,813 edges, with 1,429,853 components; additional numbers are in Table 1. An important metric in graph mining is the *density*, representing how well connected (i.e. abundance of edges) on average are the nodes of the graph. The density of a directed graph is a real number in [0, 1], maximized for cliques and minimized for a graph of isolated nodes. For the graph at hand the density is $D = \frac{|E|}{|V|(|V|-1)} = 1.571e - 07$ which means that the graph is very sparse. The degree distribution is in Fig. 2, a typical power law $\log X_i = a \log i + b$, where X_i is the number of nodes having (in and out) degree i . Numerical simulations give $a = -1.36447$ $b = 8.97957$, with R -squared value 0.71332, and suggests that the sparsity of the graph deteriorates the quality of the regression. We can see that the degrees vary importantly in the graph. More specifically there are few entities with up to more than a million capitalistic relations whereas the vast majority have less than one hundred ones. Also it is interesting that generally the outward edges per node supersede the incoming ones (i.e. entities acquire more than they are acquired).

Table 1. Capitalistic ownership graph instance (*Orbis*). (1) All edges, (2) Edges with unknown weights removed

Subgraph	Nodes		Edges	
	(1)	(2)	(1)	(2)
Organizations	81576517	81576517	6818574	4242843
Non isolated organizations	6518718	4789294	6818574	4242843
Total (individuals and organizations)	105426819	105426819	81111480	49777255

The distributions of entities by sector and country are displayed in Table 2a and b. We see that the distribution is not uniform and is not correlated with the sizes of each country. There are relatively few entities in the United States (US) compared to some European or South American countries. This is because the data source has little knowledge of US entities. One possibility for refining the study in relation to this country would be to supplement *Orbis* with data from other sources (outside the scope of this work).

The subgraph of organizations contains a very large component (i.e. a set of entities where each pair of nodes is connected via a path) of 1,442,704 nodes and 1,816,874 edges. The other components are very small, smaller than one thousand nodes. In the next subsections we provide the analytics on the largest component of juridical entities of 1,442,704 nodes following Sect. 2.



(a) Degree distributions on the subgraph of organizations

(b) Degree distribution. Power law
 $\log X_i = a \log i + b$, $a = -1.36$ $b = 8.99$,
 $R^2 = 0.71$

Fig. 2. Degree distribution**Table 2.** Top 10 attributes

(a) Top 10 sectors in total

Sector	Number
Unknown	9691495
Restaurants and mobile food service	2478398
Construction of buildings	1584314
Hairdressing and other beauty treatment	1567390
Retail sale of clothing in specialized stores	1515681
Rental/operating of own/leased real estate	1437334
Retail sale of food, beverages or tobacco	1351262
Freight transport by road	1207457
Other retail sale of new goods	1190241
Business and management consultancy	1168437

(b) Top 10 countries in total

Country	Number
Brazil	19550646
China	9865149
Italy	4985420
United Kingdom	4030892
France	3740322
Russian Federation	3258544
Germany	2631038
Australia	2554195
Netherlands	2498228
Colombia	1924520

3.2 Degeneracy and Aggregation by Attribute

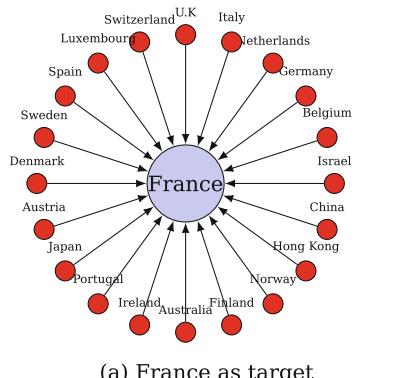
In order to provide an example of location analysis of such network, we consider the case of France. We computed the aggregated sites whose capital is most possessed by French entities (top 20 countries ranked by edge weight). Conversely, we computed the top 20 aggregate locations that own capital of French entities. We reported results in Fig. 3. It is not surprising to see that the entities most held by the French entities are those located in their former colonies (i.e. Algeria, Togo, Congo, Côte d'Ivoire, Benin, ...). Conversely, the French entities are mostly owned by those located in economically strong countries (Germany, China, U.S.A, Japan, ...) as well as countries close to each other geographically or sharing a common history (Belgium, Italy, Morocco, ...). We also reported top-5 neighbors sorted by weight in descending order (Table 3c and d). Top-in weights are much lower than the top-out weights, which represents this means that France has an outflow of capital ownership.

The original graph of entities restricted to organizations, the large component of 1,442,704 million nodes has a degeneracy value equal to 18 composed of 45 entities. The 18-core (C_{18}) is a very dense community of entities, where each

of them is being owned (resp. owns) the capital of (resp. by) at least 18 other entities in total. Recall that the other components are much smaller (i.e few hundreds of entities) and sparse, so that degeneracy in these components is not very informative for data mining. Figure 4 depicts the densest k -core ($k = 44$) on the meta-graph of countries, and represents the most connected subgraph of countries of capitalistic ownership. We observe an important proportion of countries in Europe (France, Germany, United Kingdom, Italy, Spain, Portugal, ...) and other economically strong countries (United States, China, United Arab Emirates ...). We also see that some “tax havens” are present (Cayman Islands, Malta, Cyprus, ...) despite their relative low number of entities. Also, United States of America are in the densest k -core of the meta-graph of countries, even though under-represented in the dataset (less than 850000 of entities, cf. Table 2b).

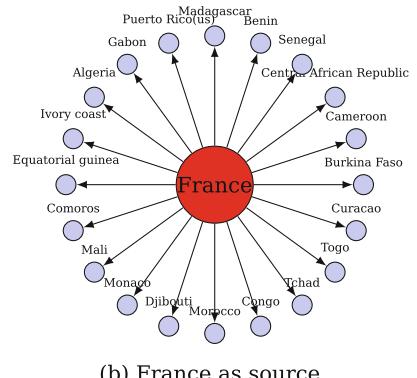
3.3 Influence Analysis

Recall that, in the context of this paper, influence is *the capacity to have an effect on the development or decisions of an entity*. In this subsection we present two different types of methods to measure it. Then, we compare the results obtained using the same diffusion model.



(c) Top 5 corresponding meta-graph weights (France as target)

Country to France	Weight in %
Belgium	0.88
Germany	0.77
Netherlands	0.51
Italy	0.50
United Kingdom	0.50



(d) Top 5 Corresponding meta-graph weights (France as source)

France to Country	Weight in %
Cameroon	39.07
Central African republic	33.33
Senegal	33.12
Benin	26.09
Madagascar	25.21

Fig. 3. France as source and target of capital ownership

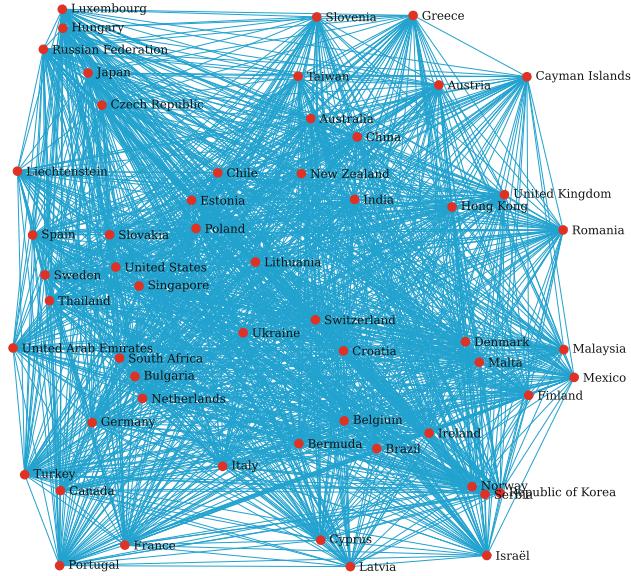


Fig. 4. Densest k-core on the meta-graph of countries ($k = 44$). It reveals a dense community of economic exchanges.

Coreness and Rooted Influence Graph (RIG): We used two coreness-based methods to measure influence indirectly. First, sorting nodes based on their coreness number in the graph. Second, following Sect. 2.3, we compute the coreness of each node in its RIG, and sort them by decreasing value. Then, we keep the top- k nodes and consider the distribution of attributes within these. The results with $k = 10000$ are in Table 3a for location analysis and in Table 4a for sector analysis.

Table 3. Top 10 influential countries using coreness on the RIG ((a), left) and influence maximization on the IM ((b), right) on the total network of entities

(a) RIG

Country	Number
Italy	1085
Germany	995
Ukraine	880
France	866
India	600
Japan	555
Australia	425
Spain	396
United Kingdom	334
Russian Federation	328

(b) IM with $k = 10000$

Country	Number
Germany	1322
China	1161
Australia	1010
France	642
Italy	628
United Kingdom	530
Austria	381
Norway	337
Spain	328
Ukraine	296

Table 4. Top 10 influential sectors

(a) RIG		(b) IM with $k = 10000$	
Sector	Number	Sector	Number
Activities of holding companies	1079	Unknown	1181
Other monetary intermediation	659	Activities of holding companies	520
Activities of head offices	636	Rental/operating of own/leased real estate	354
Rental/operating of own/leased real estate	344	Other business support service activities n.e.c.	298
Other financial service activities	312	Activities of head offices	291
Other activities auxiliary to financial services	249	Buying and selling of own real estate	214
Business and other management consultancy	244	Business and other management consultancy	212
Unknown	243	Other activities auxiliary to financial services	180
Fund management activities	193	Development of building projects	155
Other business support service activities n.e.c.	185	Production of electricity	153

Influence Maximization: The problem and method used are described in Sect. 2.5. The idea is to use the influence maximization (IM) paradigm in order to measure entity influence in the capitalistic graph. Here, we ran the simulations using influence maximization with martingales, with $\epsilon = 0.1$ and a seed set of 10 000 nodes. The output is a seed set of 10000 nodes which is an approximation of the IM solution whose quality will be discussed in Sect. 3.4. We display the distributions within this seed set in Table 3b for location analysis, and Table 4b for sector analysis.

3.4 Discussion

The influence measurement following methods presented in Sect. 2 (RIG and IMM) gave significant difference in the distributions of attributes. In order to estimate the quality of their respective solutions, we compared the number of infected nodes using our diffusion model (independent cascades), with entities obtained in the top k -cores of the graph. Let τ be the ratio of the seed set size and the number of vertices of the graph. For practical software reasons (NDlib library [10]), we considered larger seed sets ($\tau \in \{5\%, 10\%, 15\%, 20\%\}$) of the considered graph where $\tau = 5\%$ corresponds to 75136 nodes). The results are shown in Table 5. For all seed set sizes, we have obtained better results using RIG coreness than with IMM. Influence based on coreness on the initial graph yields better solutions for $\tau \in \{5\%, 10\%, 15\%\}$ seed set sizes, and IMM performs a bit better than top- k coreness for $\tau = 20\%$. This is also in accordance with experiments of [6] suggesting that coreness can yield excellent influence measurement in our context of interest of this work.

Table 5. Comparison with the IC model for several seed set sizes (50 simulations)

Method	$\sigma(S) \pm std$			
	$\tau = 5\%$	$\tau = 10\%$	$\tau = 15\%$	$\tau = 20\%$
Top- k coreness	285250.8 \pm 1304.8	351887.4 \pm 928.8	418440.6 \pm 755.8	474402.2 \pm 454.3
IMM [13]	161485.4 \pm 1211.2	298874.9 \pm 774.3	419304.6 \pm 1055.5	521251.7 \pm 1033.4
Top- k RIG coreness	610138.7 \pm 549.8	701871.5 \pm 706.7	799244.1 \pm 661.1	903330.4 \pm 369.1

4 Conclusion

We proposed a methodology to analyze a worldwide capitalistic ownership network, and reported results on instance of 2018, composed of millions of nodes and edges. This analysis includes standard graph measures and influence maximization. The statistics with regards to location and sector yield interesting results concerning the influence of entities in the capital ownership economy. To the best of our knowledge, structure and influence analysis of a large capitalistic ownership network has not been published yet. Therefore, the results could enrich or supplement economic analyses. Influence measures using k -cores and yield a better influence score for several seed set sizes (seed set size ratio $\tau \in \{5\%, 10\%, 15\%, 20\%\}$) using the independent cascade diffusion model. Therefore, this work also suggests that several means depending on the size of the seed set should be considered to measure influence in large and unknown networks.

We also discussed the method used and its limitations. Due to the sparsity of data for some countries, particularly the United States, some results of this study should be taken with caution. However, the results provide interesting insight and information on the influence of the entities and are consistent with the current global economic situation. A possible extension of this investigation would be to enrich the capital ownership dataset with additional entities or capital information to provide a more precise quantitative measure of influence, and to consider studying the evolution of this network over time (monthly or yearly).

References

- Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex networks. *Nature* **406**(6794), 378 (2000)
- Batagelj, V., Zaversnik, M.: An $O(m)$ algorithm for cores decomposition of networks. arXiv preprint [arXiv:cs/0310049](https://arxiv.org/abs/cs/0310049) (2003)
- Boss, M., Elsinger, H., Summer, M., Thurner, S.: Network topology of the interbank market. *Quant. Financ.* **4**(6), 677–684 (2004)
- Dijk, B.V.: Source: Orbis, bureau van dijk
- Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 57–66. ACM (2001)
- Giatsidis, C., Nikolentzos, G., Zhang, C., Tang, J., Vazirgiannis, M.: Rooted citation graphs density metrics for research papers influence evaluation. *J. Informetr.* **13**(2), 757–768 (2019)
- Inaoka, H., Ninomiya, T., Taniguchi, K., Shimizu, T., Takayasu, H., et al.: Fractal network derived from banking transaction—an analysis of network structures formed by financial institutions. *Bank of Japan Working Paper* **4** (2004)
- Lenzu, S., Tedeschi, G.: Systemic risk on different interbank network topologies. *Physica A Stat. Mech. Appl.* **391**(18), 4331–4341 (2012)
- Li, Y., Fan, J., Wang, Y., Tan, K.L.: Influence maximization on social graphs: a survey. *IEEE Trans. Knowl. Data Eng.* **30**(10), 1852–1872 (2018)

10. Rossetti, G., Milli, L., Rinzivillo, S., Sirbu, A., Pedreschi, D., Giannotti, F.: NDlib: a Python library to model and analyze diffusion processes over complex networks. *Int. J. Data Sci. Anal.* **5**(1), 61–79 (2018)
11. Siganos, G., Faloutsos, M., Faloutsos, P., Faloutsos, C.: Power laws and the as-level internet topology. *IEEE/ACM Trans. Networking* **11**(4), 514–524 (2003)
12. Soramäki, K., Bech, M.L., Arnold, J., Glass, R.J., Beyeler, W.E.: The topology of interbank payment flows. *Physica A Stat. Mech. Appl.* **379**(1), 317–333 (2007)
13. Tang, Y., Shi, Y., Xiao, X.: Influence maximization in near-linear time: a martingale approach. In: Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pp. 1539–1554. ACM (2015)

Multilayer Networks



Patterns of Multiplex Layer Entanglement Across Real and Synthetic Networks

Blaž Škrlj^{1,2} and Benjamin Renoust^{3(✉)}

¹ Jožef Stefan Institute, Jamova 38, Ljubljana, Slovenia
blaz.skrlj@ijs.si

² Jožef Stefan International Postgraduate School, Jamova 38, Ljubljana, Slovenia

³ Osaka University, Institute for Datability Science, Osaka, Japan
renoust@ids.osaka-u.ac.jp

Abstract. Real world complex networks often exhibit multiplex structure, connecting entities from different aspects of physical systems such as social, transportation and biological networks. Little is known about general properties of such networks across disciplines. In this work, we first investigate how consistent are connectivity patterns across 35 real world multiplex networks. We demonstrate that entanglement homogeneity and intensity, two measures of layer consistency, indicate apparent differences between social and biological networks. We also investigate trade, co-authorship and transport networks. We show that real networks can be separated in the joint space of homogeneity and intensity, demonstrating the usefulness of the two measures for categorization of real multiplex networks. Finally, we design a multiplex network generator, where similar patterns (as observed in real networks), are emerging over the analysis of 11,905 synthetic multiplex networks with various topological properties.

Keywords: Multiplex networks · Edge entanglement · Network topology · Network generator

1 Introduction

Real-world networks commonly consist of different types of entities, all connected into a single system. The abstraction of *multiplex* networks offers a structure, capable of capturing the key parts of such systems, such as connectivity patterns. Multiplex networks emerge, and were studied in biology, social sciences, finance, logistics and more. They are both theoretically interesting, as well as practically useful [1]. Recently, the notions of multiplex community detection and centralities have been a lively research area, indicating many insights can be obtained by studying such rich structures directly, without simplification [6, 13, 27] (e.g., aggregation into a single node type). Multiplex networks offer the opportunity to simultaneously explore multiple aspects of the same system [16], and are as

such indispensable for the study of *e.g.*, biological or social networks, where entities can be naturally observed with respect to different aspects (*e.g.*, an user on Twitter, Facebook and Snapchat is the same physical person, yet can be studied with respect to individual social networks where it is present).

The ideas, that influenced this work the most are discussed next. Since the structure of a multilayer corresponds to its layers and aspects [15], the analysis of the organization of layers is key to understanding the properties of a multiplex network [24]. The analysis of the overlapping edges between layers, namely *edge entanglement* [25] studies how the different layers of a multiplex network intertwine to form a coherent whole.

Even though the ideas related to description of multiplex networks are being actively developed [22, 31], we believe little effort is focused on evaluation of such measures at larger scales, across multiple disciplines and contexts. This work was also inspired by multilayer flow analysis [10], where distinct structures, describing parts of networks emerged. The contributions of this work are multiple, and are described next:

- We present an efficient implementation of multiplex homogeneity and intensity, the two measures used in this work [25].
- Both measures, along with normalized homogeneity, are computed for the first time on 35 real-world multiplex networks.
- We demonstrate a distinct relationship between homogeneity and intensity, showing the two measures can separate between different types of multiplex networks.
- We present a multiplex network generator that produced networks with various degrees of intensity and homogeneity. We generated 11,905 synthetic networks, where patterns, similar to the ones in the real networks emerged.

2 Multiplex Networks

A multiplex network can be defined as a sequence $M = \{G_l\}_{l \in L} = \{(V_l, E_l)\}_{l \in L}$ where $E_l \subseteq N_l \times V_l$ is a set of edges in one network $l \in L$ of the sequence [15]. Multiplex networks are commonly understood as layers comprised of interactions, where each layer corresponds to a specific aspect of the system, and nodes represent *the same* entity across all layers. We represent a multiplex network as a structure $M = (V_M, E_M)$, where V_M is the set of nodes and E_M the set of all edges (in all layers).

For example, a biological system can be studied at the protein, RNA or gene level [29], and similarly, social networks can be studied by taking into account a person's presence on multiple platforms [21]. For computational purposes, such networks are commonly represented in the form of supra-adjacency matrices, where block-diagonal structure, connecting the same node across individual layers emerges [9]. Algorithms can operate on such matrices directly and thus exploit such additional information representing multiple aspects. Such approaches are useful when node-level information is considered.

Algorithms for analysis of multiplex networks can also operate on sparse, adjacency structure of the multiplex network directly, yet need to take into account that a given node is present in multiple layers. Such representation is suitable for this work, as we are focused primarily on how edges co-occur across *layers*. Hence, this work focuses primarily on the relations between the *layers* of a given multiplex network. We next discuss the two measures we consider throughout this work.

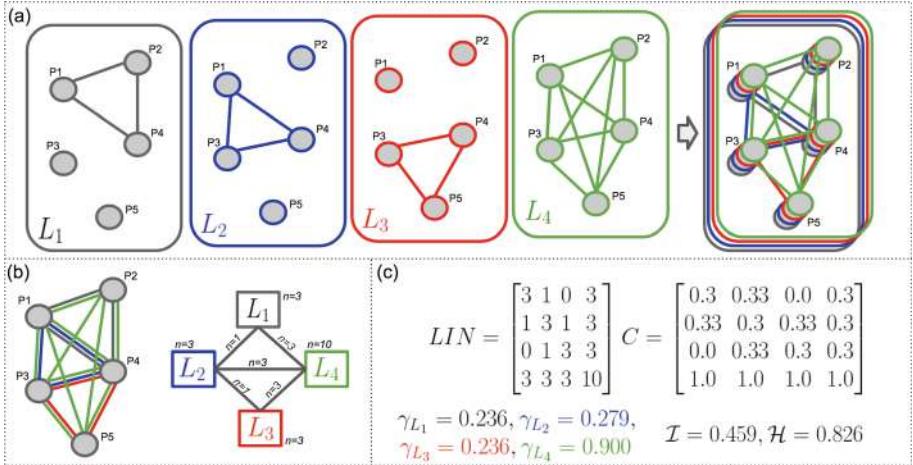


Fig. 1. A toy example of layer entanglement computation: (a) separated layers considered in a multilayer network; (b) constructing the layer interaction network from the example; (c) measuring entanglement from the example.

3 Multiplex Entanglement and Intensity

We briefly discuss the entanglement measures definitions from previous work [25].

3.1 Layer Interaction Network

Recall our multiplex network $M = (V_M, E_M) = \{G_l\}_{l \in L}$. Such a network really distinguishes itself from classical graphs through the use of different layers to connect nodes. These layers may have different patterns and may overlap together. There may even exist latent dependencies among these layers. To investigate this matter, each layer could be abstracted to one single node and form a new graph, the Layer Interaction Network (*LIN*) [25]. Visualizing the *LIN* is a key component for multiplex network visualization such as in Detangler [24].

In the *LIN*, $LIN = (L, F)$, each node corresponds to a layer $l, l', l'', \dots \in L$ of the multiplex network M , and each edge $f \in F$ captures when two layers overlap through edges. More formally, there exist an edge $f = (l, l')$ whenever there exists at least two nodes $u, v \in V_M$ such that there exists at least one edge connecting

these two nodes on each layer $e_M = (u, v) \in l$ and $e'_M = (u, v) \in l'$. The LIN can be interpreted as an edge-layer co-occurrence graph, and the weight of an edge $f = (l, l')$, denoted as $n_{l,l'}$ equals the number of times layers l, l' co-occur. By extension, $n_{l,l}$ is the number of edges on layer l . This process is illustrated in Fig. 1.

3.2 Layer Entanglement

The analysis of edge entanglement is inspired by the analysis of relation content in social networks [3]. The idea is to study the redundancy between relation content, each forming in our formalism a different layer. The edge entanglement measures the “influence” of a layer in its neighborhood.

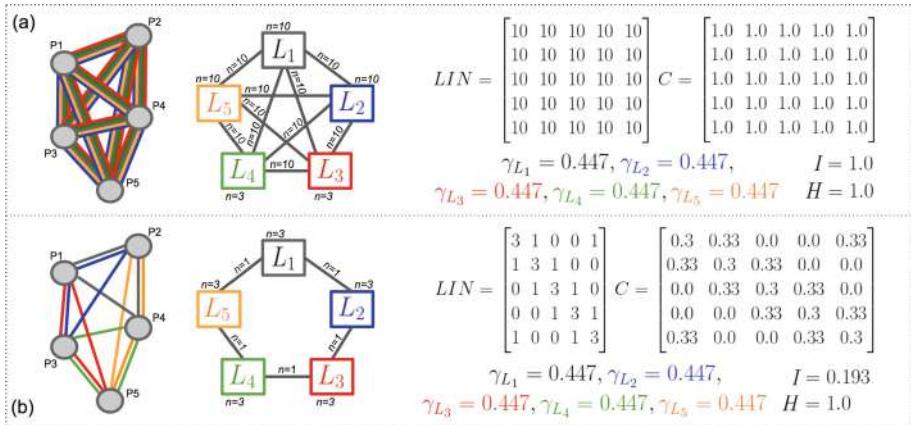


Fig. 2. Two very different cases of maximum homogeneity $H = 1$, the multiplex network and the LIN are shown, with matrices and entanglement measures. (a) all layers are saturating all edges, so we have maximum intensity $I = 1$; (b) layers are well balanced, but we may have a lot more interactions possible.

This measure is recursively defined: the entanglement γ_l of a layer l is defined upon the entanglement of the layers it is entangled with. Similarly to the eigen centrality [32], this translates into the recursive equation:

$$\gamma_l \cdot \lambda = \sum_{l' \in T} \frac{n_{ll'}}{n_l} \gamma_{l'}.$$

The entanglement of a layer γ_l can be retrieved from a vector γ which corresponds to the right eigenvector (associated to the maximum eigenvalue λ) of the layer overlap frequency matrix with corresponding overlap, defined as:

$$C = (c_{ll'}), \quad \text{where } c_{ll'} = \frac{n_{ll'}}{n_l}.$$

this metric was initially discussed in [25]), and is constructed using the weights in the LIN (see Figs. 1 and 2).

3.3 Entanglement Intensity and Homogeneity

The layer entanglement γ_l measures the share of layer l overlapping with other layers, so that nodes of M are connected. The more a group of layers interacts together, the more the nodes they connect will be cohesive in view of *these* layers, hence the more $\gamma_l \forall l \in L$ values will be similar (their share of entanglement will be similar). This is captured by the *entanglement homogeneity* [25] which is then defined as the following cosine similarity:

Algorithm 1: Multiplex network generator.

```

Parameters : Number of nodes  $v$ , number of layers  $k$ , dropout  $d$ 
Result: A multiplex network  $M$ 
1  $M \leftarrow$  emptyMultiplexObject;
2 for node in  $[1 \dots v]$  do
3   | numberOfLayers  $\leftarrow$  randomNumber( $k$ ) ;     $\triangleright$  Layer presence is random.
4   | layerNodes  $\leftarrow$  assignNodeToLayers(node, numberOfLayers);
5   | update( $M$ , layerNodes);                       $\triangleright$  Update global network.
6 end
7 for layer  $l_i$  with corresponding node set  $V_{l_i}$  do
8   | nodeClique  $\leftarrow$  generator of node pairs from  $V_{l_i}$ ;
9   | finalLayer  $\leftarrow$  sampleWithProbability(nodeClique,  $1 - d$ );  $\triangleright$  Sample via  $d$ .
10  | update( $M$ , finalLayer);                       $\triangleright$  Update global network.
11 end
12 return  $M$ ;

```

$$H = \frac{\langle \mathbf{1}_L, \boldsymbol{\gamma} \rangle}{\|\mathbf{1}_L\| \|\boldsymbol{\gamma}\|} \in [0, 1].$$

Optimal homogeneity is not necessarily reached only when all nodes are connected through all layers, but also when all nodes are connected in a very balanced manner between all layers (see Fig. 2). Homogeneity thus permits various *symmetries* in a given LIN.

When a maximum overlap is reached through all layers in the network, the frequencies in the matrix C (of size $|L| \times |L|$) are saturated with $C_{i,j} = 1$. This gives us a theoretical limit to measure the amount of layer overlap through the *entanglement intensity* [25], defined as:

$$I = \lambda / |L|.$$

4 A Multiplex Network Generator

In this section, we describe an algorithm for generation of multiplex networks based on the following observations. Let $M = (V_M, E_M)$ represent a multiplex network with L layers. Each node is associated to a random number of layers $\{l_1, l_2, \dots, l_i\} \subseteq L$. Now for each layer $l_i \in L$ there is a set of nodes $V_{l_i} \subseteq V_M$

which form a potential set of edges of size $|E_{l_i}| = \frac{1}{2}|V_{l_i}|(|V_{l_i}| - 1)$. We introduce the probability p of an edge to be created between any pair of node on a layer so we may avoid cliques to form on each layer. For algorithmic reasons, this probability is implemented as an edge dropout d such as $p = 1 - d$, and randomly prune edges from a potential clique. Thus, the higher the d , the sparser the network. Intuitively, the more similar a given random multiplex is to a clique over each layer, the higher its intensity. The purpose of this generator is to offer a simple testbed for further exploration, as well as additional evidence of the relation between homogeneity and intensity on many random, synthetic networks. The Algorithm 1 represents the proposed procedure.

The generator first randomly assigns the same node index to randomly many layers (lines 1–6). Once assigned, the layers are processed by applying the dropout on $\binom{|V_{l_i}|}{2}$ possible edges in layer l_i . The global multiplex is updated during this process (lines 7–12). Note that in line 8, the whole clique is virtually generated. This step is not necessary, as commonly only a small number of edges need to be sampled from all possible edge combinations. The implementation thus uses a generator with lazy evaluation, avoiding potential combinatorial explosion with a large number of nodes (very large networks).

4.1 Some Theoretical Properties of the Generator

In this section we show two properties of the proposed generator. We denote $v = |V_M|$ the parameter setting the number of nodes of the network, $k = |L|$ the parameter setting the number of edge layers in the network, and d the edge dropout.

Proposition 1 (Number of edges). *Let $\phi \in \mathbb{N}^+$ represent the number of possible edges. Then $\phi \leq k \cdot \binom{v}{2}$.*

Proof. Note that in multiplex layers, each layer can have at most v nodes. Assuming they form a clique, each layer is thus comprised of $\binom{v}{2}$ nodes. As there are k layers, there can be at most $k \cdot \binom{v}{2}$ edges — a clique of v nodes in each layer. As each layer is during generation subject to dropout, which is neglected, when set to 0 (no edges are erased), we refer to this bound as $\phi \leq k \cdot \binom{v}{2}$. \square

Corollary 1 (Time complexity). *In lower limit, $d \rightarrow 0$, thus a full clique needs to be constructed, assuming each node is projected across all layers. The complexity w.r.t. the number of layers and edges is: $\mathcal{O}(k \cdot \binom{v}{2}) = \mathcal{O}(|E_M|)$.*

Note that even though, theoretically, the proposed generator generates a clique and then samples from it, current, lazy implementation only *generates* the edges needed to satisfy a given d percentage. In practice, only when $d \approx 0$, the generator needs larger portions of space (and time). As such, fully connected networks do not represent real systems, we were able to generate networks with tens of thousands of nodes using this approach.

5 Empirical Evaluation

In this section we discuss the empirical evaluation of the two considered measures across a series of real world networks.

All considered networks are summarized in Table 1¹. All considered networks are static. We computed the metrics for all connected components. The entanglement algorithm was integrated into the Py3plex library [26]². For each network,

Table 1. Real multiplex networks and their properties. The ID in the second column corresponds to Fig. 3(c).

Dataset	ID	Type	Nodes	Edges	Number of layers	Mean degree	CC
arXiv-Netscience [10]	6	Coauthorship	26796	59026	13	4.41	3660
PierreAuger [10]	22	Coauthorship	965	7153	16	14.82	131
Arabidopsis [28]	39	Genetic	8765	18655	7	4.26	387
Bos [28]	3	Genetic	369	322	4	1.75	82
Candida [28]	23, 24, 25	Genetic	418	398	7	1.90	50
Celegans [28]	32	Genetic	4557	8182	6	3.59	193
DanioRerio [28]	26, 27	Genetic	180	188	5	2.09	45
Drosophila [28]	31	Genetic	11970	43367	7	7.25	346
Gallus [28]	16	Genetic	367	389	6	2.12	54
HepatitisCVirus [28]	33	Genetic	129	137	3	2.12	4
Homo Sapiens [28]	30	Genetic	36194	170899	7	9.44	785
HumanHerpes4 [28]	29	Genetic	261	259	4	1.98	21
HumanHIV1 [28]	5	Genetic	1195	1355	5	2.27	13
Oryctolagus [28]	7	Genetic	151	144	3	1.91	21
Plasmodium [28]	9	Genetic	1206	2522	3	4.18	27
Rattus [28]	40	Genetic	3263	4268	6	2.62	296
SacchCere [28]	2	Genetic	27994	282755	7	20.20	432
SacchPomb [28]	1	Genetic	10178	63677	7	12.51	286
Xenopus [28]	37, 38	Genetic	582	620	5	2.13	109
YeastLandscape [8]	34	Genetic	17770	8473997	4	953.74	4
CElegans [5]	20	Neuronal	791	5863	3	14.82	6
Cannes2013 [22]	8	Social	659951	991854	3	3.01	48375
CKM-Physicians-Innovation [7]	19	Social	674	1551	3	4.60	12
CS-Aarhus [19]	36	Social	224	620	5	5.54	13
Kapferer-Tailor-Shop [14]	35	Social	150	1018	4	13.57	5
Krackhardt-High-Tech [17]	13	Social	63	312	3	9.90	3
Lazega-Law-Firm [18]	18	Social	211	2571	3	24.37	3
MLKing2013 [22]	14	Social	392542	396671	3	2.02	36041
MoscowAthletics2013 [22]	17	Social	133619	210250	3	3.15	6323
ObamaInIsrael2013 [22]	21	Social	3457453	4061960	3	2.35	651141
Padgett-Florence-Families [23]	28	Social	26	35	2	2.69	2
Vickers-Chan-7thGraders [30]	0	Social	87	740	3	17.01	3
FAO [11]	15	Trade	41713	318346	364	15.26	571
EUAir [4]	4	Transport	2034	3588	37	3.53	41
London [12]	11, 12	Transport	399	441	3	2.21	3

¹ The networks are hosted at <https://comunelab.fbk.eu/data.php>.

² https://github.com/SkBlaz/Py3plex/blob/master/examples/example_entanglement.py. The generator is accessible at https://github.com/SkBlaz/Py3plex/blob/master/py3plex/core/random_generators.py.

we computed homogeneity and intensity. For the generation of synthetic networks, we used the following hyperparameter ranges:

- $v \in \{10, 25, 50, 100, 250, 500, 1000, 2500\}$
- $k \in \{3, 4, 5, 6, 7, 8, 9, 10\}$
- $d \in \{0.001, 0.9, 0.01\}$

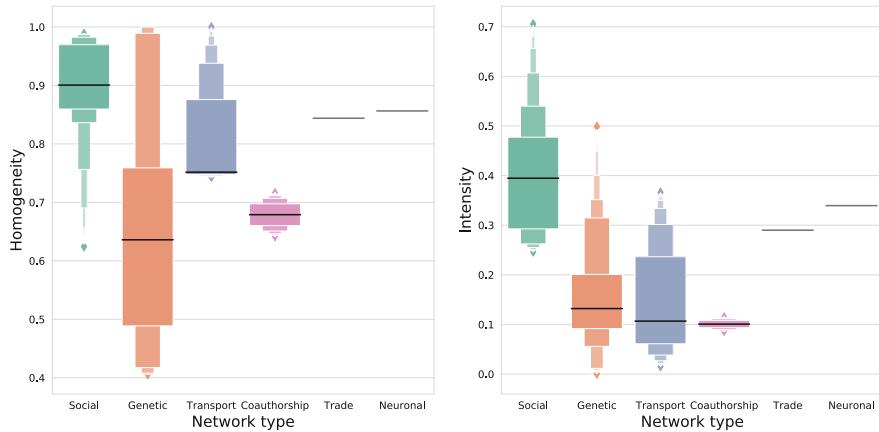
(a) Real networks: homogeneity H (b) Real networks: intensity I (c) Real networks: $H \times I$

Fig. 3. Results on real networks. Labels in (c) map to Table 1 (ID). Gray dots represent synthetic samples.

6 Results

In this section we present the results of empirical evaluation. For readability purposes, we visualize individual results as distributions of a given score across network types. We first show entanglement metrics on real networks in Fig. 3. We next present the results on the generated networks in Fig. 4.

Two main observations are apparent when studying the results on real networks. First, the difference between social and genetic (biological) multiplex networks becomes obvious when both entanglement intensity, as well as homogeneity are considered. We further visualize the two most apparent distributions, *i.e.*, the intensity and homogeneity of social *vs.* genetic networks in Fig. 5.

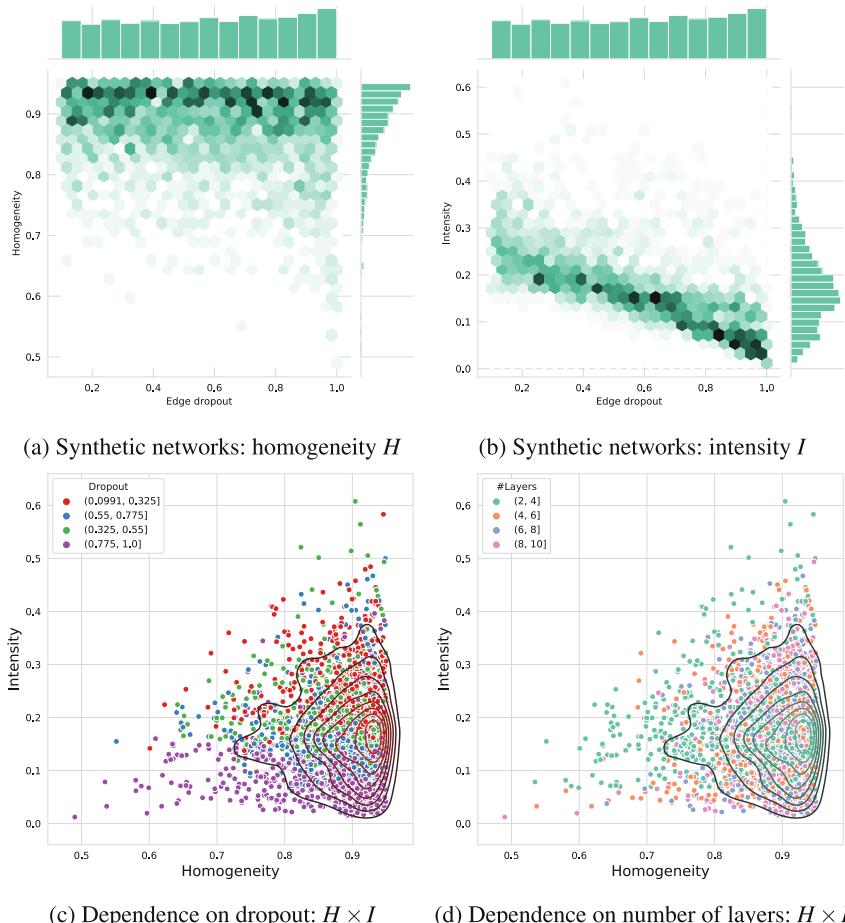


Fig. 4. Results on 11,905 synthetic networks. The homogeneity (a) and intensity (b) dependence on dropout and number of layers parameter results in heavy-tailed distributions of the two measures (c, d).

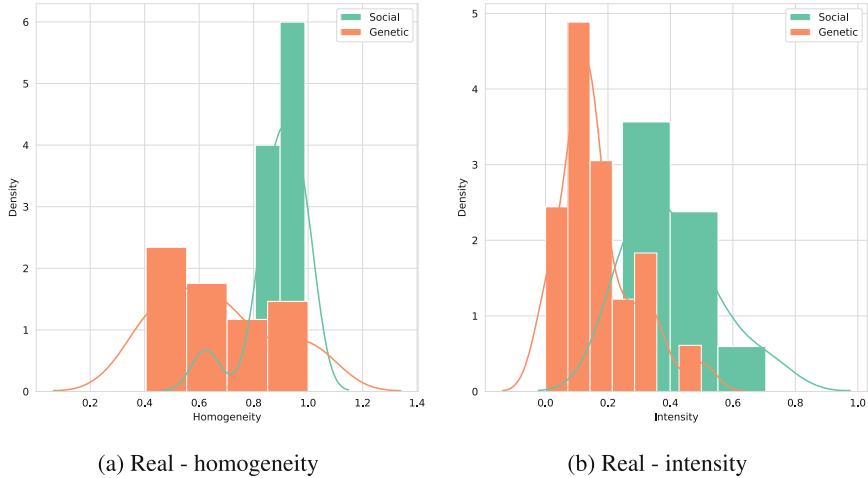


Fig. 5. Distributions of homogeneity and intensity when genetic networks are compared to social ones.

The properties of synthetic networks were plotted with respect to the dropout parameter. The reader can observe apparent linear trend between the dropout (sparseness) and entanglement intensity (Fig. 4(b)). This trend indicates sparser networks are less “intensely” coupled. As intensity directly measures this property, this result outlines one of the *desired* properties of the proposed network generator.

The reader can also observe high density of *networks* in the space of high homogeneity and average or low intensity (Fig. 4(d)). This property directly reflects the sampling procedure, as the majority of the considered networks consist of edges, which co-occur in majority of layers. A similar observation can be observed in Fig. 4(a,c), where denser regions of the homogeneity/intensity space emerge when higher homogeneity is considered. Note that we also visualized the dependence of the synthetic network’s properties on the dropout, as well as the number of layers—both parameters determine a given multiplex’s structure.

7 Discussion and Conclusion

In this paper we demonstrated that two measures for assessing the relation between layers in a given multiplex network offer interesting insights when computed across a wide array of real-world networks. To our understanding, the observed relationship between the intensity and homogeneity of layer entanglement was not yet reported. We showed that real networks cluster based on their type (*e.g.* biological *vs.* social). Apart from experiments on real networks we also generated a large set of synthetic ones, where the analysis outlined the following properties: Intensity is directly correlated with edge dropout parameter—the sparser the network, the lower the intensity. This result indicates the proposed

generator indeed emits networks which adhere to this property. Next, we observe that large parts of the generated networks are subject to high homogeneity with various degrees of entanglement intensity.

The detailed inspection of the synthetic networks with respect to the parameters d and the number of layers (k) reveals that the generative process is more sensitive to dropout (layered patterns of intensity emerge), than to the number of layers (uniformly distributed *w.r.t.* homogeneity). This property indicates the model's properties could also be investigated theoretically, which we leave for further work.

In addition, we may observe (from Fig. 3) that our set of genetic networks tend to match networks with higher dropout, as opposed to social networks which tend to find their way in lower dropout area. This should be further investigated, but this may be related to *homophily* [2, 20]. Homophily is the implied similarity of two entities in a social network, and the property of entities to agglomerate when *being similar*. If the reason of '*being similar*' could be modeled as a layer of interaction, the result of a group of entities in '*being similar*' would lead to the formation of a clique in this layer, hence locating social networks in low dropout areas.

The proposed work offers at least two prospects of multiplex network study which are in our belief worth exploring further. The difference between the genetic and social networks is possibly subject to very distinct topologies which emerge in individual layers. This claim can be empirically evaluated via measurement of *e.g.*, graphlets, communities or other structures. Next, genetic networks are less homogeneous. Further work includes exploration of this fact, as it can be merely a property of the networks considered, empirical methodology used to obtain the networks or some other effect.

We believe that theoretical properties of the proposed network generator can also be further studies, offering potential insights into how multiplex networks behave and whether the human-made aspects are indeed representative of a given system's state.

Acknowledgements. The work of the first author was funded by the Slovenian Research Agency through a young researcher grant. The work of other authors was supported by the Slovenian Research Agency (ARRS) core research programme *Knowledge Technologies* (P2-0103) and ARRS funded research project *Semantic Data Mining for Linked Open Data* (financed under the ERC Complementary Scheme, N2-0078). We also acknowledge Dagstuhl seminar-19061 [16] where many ideas implemented in this paper emerged.

References

1. Battiston, F., Nicosia, V., Latora, V.: Structural measures for multiplex networks. *Phys. Rev. E* **89**(3), 032804 (2014)
2. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. *Science* **323**(5916), 892–895 (2009)
3. Burt, R.S., Schøtt, T.: Relation contents in multiple networks. *Soc. Sci. Res.* **14**(4), 287–308 (1985)

4. Cardillo, A., Gómez-Gardenes, J., Zanin, M., Romance, M., Papo, D., Del Pozo, F., Boccaletti, S.: Emergence of network features from multiplexity. *Sci. Rep.* **3**, 1344 (2013)
5. Chen, B.L., Hall, D.H., Chklovskii, D.B.: Wiring optimization can relate neuronal structure and function. *Proc. Natl. Acad. Sci.* **103**(12), 4723–4728 (2006)
6. Chen, X., Wang, R., Tang, M., Cai, S., Stanley, H.E., Braunstein, L.A.: Suppressing epidemic spreading in multiplex networks with social-support. *New J. Phys.* **20**(1), 013007 (2018)
7. Coleman, J., Katz, E., Menzel, H.: The diffusion of an innovation among physicians. *Sociometry* **20**(4), 253–270 (1957)
8. Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., et al.: The genetic landscape of a cell. *Science* **327**(5964), 425–431 (2010)
9. Cozzo, E., Kivelä, M., De Domenico, M., Solé-Ribalta, A., Arenas, A., Gómez, S., Porter, M.A., Moreno, Y.: Structure of triadic relations in multiplex networks. *New J. Phys.* **17**(7), 073029 (2015)
10. De Domenico, M., Lancichinetti, A., Arenas, A., Rosvall, M.: Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Phys. Rev. X* **5**(1), 011027 (2015)
11. De Domenico, M., Nicosia, V., Arenas, A., Latora, V.: Structural reducibility of multilayer networks. *Nat. Commun.* **6**, 6864 (2015)
12. De Domenico, M., Solé-Ribalta, A., Gómez, S., Arenas, A.: Navigability of interconnected networks under random failures. *Proc. Natl. Acad. Sci.* **111**(23), 8351–8356 (2014)
13. Gomez, S., Diaz-Guilera, A., Gomez-Gardenes, J., Perez-Vicente, C.J., Moreno, Y., Arenas, A.: Diffusion dynamics on multiplex networks. *Phys. Rev. Lett.* **110**(2), 028701 (2013)
14. Kapferer, B.: *Strategy and Transaction in an African Factory: African Workers and Indian Management in a Zambian Town*. Manchester University Press, Manchester (1972)
15. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. *J. Complex Netw.* **2**(3), 203–271 (2014)
16. Kivelä, M., McGee, F., Melançon, G., Henry Riche, N., von Landesberger, T.: Visual analytics of multilayer networks across disciplines (dagstuhl seminar 19061). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (2019)
17. Krackhardt, D.: Cognitive social structures. *Soc. Netw.* **9**(2), 109–134 (1987)
18. Lazega, E., et al.: *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press, Oxford (2001)
19. Magnani, M., Micenková, B., Rossi, L.: Combinatorial analysis of multiple networks. arXiv preprint [arXiv:1303.4986](https://arxiv.org/abs/1303.4986) (2013)
20. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. *Annu. Rev. Sociol.* **27**(1), 415–444 (2001)
21. Mittal, R., Bhatia, M.: Analysis of multiplex social networks using nature-inspired algorithms. In: *Nature-Inspired Algorithms for Big Data Frameworks*, pp. 290–318. IGI Global (2019)
22. Omodei, E., De Domenico, M.D., Arenas, A.: Characterizing interactions in online social networks during exceptional events. *Front. Phys.* **3**, 59 (2015)
23. Padgett, J.F., Ansell, C.K.: Robust action and the rise of the medici, 1400–1434. *Am. J. Sociol.* **98**(6), 1259–1319 (1993)

24. Renoust, B., Melançon, G., Munzner, T.: Detangler: visual analytics for multiplex networks. *Comput. Graphics Forum* **34**(3), 321–330 (2015)
25. Renoust, B., Melançon, G., Viaud, M.L.: Entanglement in multiplex networks: understanding group cohesion in homophily networks. In: Missaoui, R., Sarr, I. (eds.) *Social Network Analysis - Community Detection and Evolution*, pp. 89–117. Springer, Cham (2014)
26. Škrlj, B., Kralj, J., Lavrač, N.: Py3plex: a library for scalable multilayer network analysis and visualization. In: *International Conference on Complex Networks and their Applications*, pp. 757–768. Springer (2018)
27. Škrlj, B., Kralj, J., Lavrač, N.: CBSSD: community-based semantic subgroup discovery. *J. Intell. Inf. Syst.* **53**, 1–40 (2019)
28. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: Biogrid: a general repository for interaction datasets. *Nucleic Acids Res.* **34**(suppl.1), D535–D539 (2006)
29. Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., Baudot, A.: Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* **35**(3), 497–505 (2018)
30. Vickers, M., Chan, S.: Representing classroom social structure. Victoria Institute of Secondary Education, Melbourne (1981)
31. Wang, W., Cai, M., Zheng, M.: Social contagions on correlated multiplex networks. *Physica A Stat. Mech. Appl.* **499**, 121–128 (2018)
32. Wasserman, S., Faust, K.: *Social Network Analysis, Methods and Applications. Structural Analysis in the Social Sciences*. Cambridge University Press, Cambridge (1994)



Introducing Multilayer Stream Graphs and Layer Centralities

P. Parmentier^{1,2}, T. Viard^{1(✉)}, B. Renoust^{3,4}, and J.-F. Baffier^{1,5}

¹ RIKEN AIP, Tokyo, Japan
viard@lipn.univ-parisb.fr

² École Polytechnique, Palaiseau, France

³ Institute for Datability Science, Osaka University, Osaka, Japan

⁴ JFLI, CNRS UMI3527, Tokyo, Japan

⁵ Japan Society for the Promotion of Sciences, Tokyo, Japan

Abstract. Graphs are commonly used in mathematics to represent some relationships between items. However, as simple objects, they sometimes fail to capture all relevant aspects of real-world data. To address this problem, we generalize them and model interactions over time with multilayer structure. We build and test several *centralities* to assess the importance of layers of such structures. In order to showcase the relevance of this new model with centralities, we give examples on two large-scale datasets of interactions, involving individuals and flights, and show that we are able to explain subtle behaviour patterns in both cases.

Keywords: Multilayer graph · Stream graph · Centrality · Density

1 Introduction

Graphs have been widely used since the first definition of the Königsberg Bridges by Euler [9]. Although their formalization and drawing came later [14], graphs have been constantly challenged and their formalism extended in many ways, with, among the most common, orientation, labels, and weights for nodes and links [3], to represent the connections of things in their entirety: friendships, railroads, communications, *etc.*

Recently, new formalisms have emerged to encompass the more complex patterns that arise in real-world data. In particular, the multilayer networks [15] capture multiple families of relationships and entities together. This is useful, for example, to inspect homophily within groups of documents [23]. However, multilayer networks show some limitations in fully capturing interactions that exist over time [21], beyond the dynamic of a graph as a whole. To cope with many individual time-dependant interactions, stream graphs [16] offer a comprehensive formalism to deal with real-world sequences of interactions over time.

In this paper, we are interested in joining both formalisms by proposing the *multilayer stream graph*. After briefly reviewing the state of the art, we

give its definition in Sect. 3, we explore the notion of centrality in Sect. 3.1, while applying this model on two datasets in Sect. 4 before concluding. The key contributions of this paper are (i) the introduction of multilayer stream graphs and (ii) the demonstration of its relevance for the detection of central layers.

2 Related Work

We now introduce the concept of multilayer stream graph. To illustrate our definitions, we will use a toy example of a population of monkeys F_1, F_2, M_1, M_2 (two females and two males) interacting together.

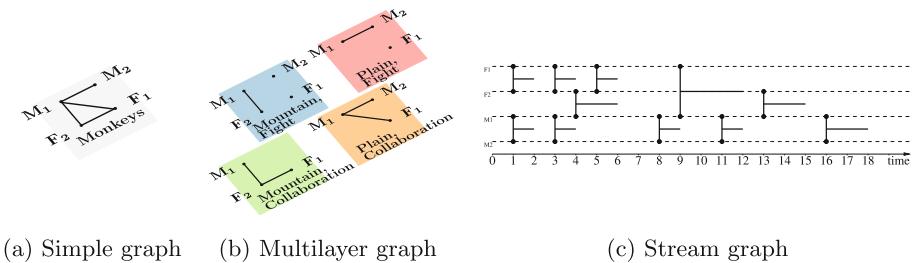


Fig. 1. Example of the interactions of a population of monkeys. (a) The simple social network of monkeys. (b) The multilayer graph between monkeys across places and relationships (collaborating and fighting in the mountain or in the plain). (c) The stream graph describing sequence of interactions between monkeys. We may notice the frequent interactions between monkeys.

2.1 From Graphs to Multilayer and Stream Graphs

A simple graph is a tuple $G = (V, E)$ composed by a set of nodes V and a set of edges $E \subseteq V \otimes V$, where each edge is an unordered pair of two distinct nodes. The *degree* of a node $v \in V$, denoted $d(v)$ is the number of edges in which v appears: $d(v) = |\{(u, v) \in E | u \in V\}|$. The *density* is the probability, given two nodes, that they are connected: $\delta(G) = \frac{2|E|}{|V|(|V|-1)}$. In Fig. 1a, the nodes are monkeys: $V = \{F_1, F_2, M_1, M_2\}$ and there is a link in E between two animals if they have been in contact, hence forming $G = (V, E)$. M_1 is the monkey with the highest degree. In this example, the density $\delta(G) = \frac{2}{3}$.

Multilayer Graphs. Consider now that the monkeys from Fig. 1a to interact in different places, such as the *mountain* or the *plain*, either through *collaboration* or *fight*.

A *multilayer graph* [8, 15] is a set $M = (V_M, E_M, V, \mathcal{L})$, where \mathcal{L} is the *structure*, a finite set of d different sets, named *aspects* such that $\mathcal{L} = \{L_1, \dots, L_d\}$. Each aspect L_i contains elements $l_i^1, \dots, l_i^{n_i}$ which are named *elementary layers*. A *layer* α is then a combination of elementary layers from each aspect:

$\alpha \in L = L_1 \times \dots \times L_d$. V is the set of nodes and each of them can be present on any of the different layers. A node on a layer is called a *node-layer*. The set of nodes-layers is $V_M \subseteq V \times L$. Each node-layer can be linked to another with undirected edges, *i.e.* $E_M \subseteq V_M \otimes V_M$. The *degree of a node-layer* is the number of links in which the node-layer appears. The *degree of a node* is the number of links in which the node appears. The *density* can also be defined for each layer, hence generalized over the multilayer network: $\delta(M) = \frac{2|E_M|}{|V_M|(|V_M|-1)}$.

Figure 1b represents our monkeys example in a multilayer context. The structure \mathcal{L} is made of two aspects: the place and the type of interaction. Consider the layer (*collaboration, mountain*) in which a link between M_1 and F_2 means that the two animals collaborated in the mountain. The layer (*collaboration, mountain*) and (*collaboration, plain*) show higher density than the other layers. A possible interpretation is that this group of monkeys interactions are more collaboration based. Also, F_2 , (*collaboration, mountain*) and M_1 , (*collaboration, plain*) are the two key individuals with the highest collaboration degree.

Stream Graphs. However, this does not take into account the temporal nature of interactions, *i.e.* when and for how long they occur. For example, fights or collaborations could be short or long depending on the circumstances, and occur frequently or not. This information is crucial for finer grained understanding.

A *stream graph* [16] is a tuple $S = (T, W, V, E)$ where T is the time interval of study, V is the set of nodes. The stream graph model does not require a discrete definition of time. A stream graph considers a set of *time instants* in a continuous manner.

The *time-nodes* set $W \subseteq T \times V$ describes the existence of nodes depending on time: $(t, v) \in W$ means that the node v appears at time instant t . The set $E \subseteq T \times V \times V$ contain all the links and their time instants of existence. Given nodes u and v , we call $T_u = \{t, (t, u) \in W\}$ the *set of time instants* at which u appears, and $T_{uv} = \{t, (t, uv) \in E\}$ the *set of time instants* at which the link (u, v) appears.

Several notions have been designed in [16] that extend the model of classical graphs. For example, links may either last a certain amount of time, or be instantaneous (leading to a density equal to zero).

The *number of links* of a set of links is formally defined as the duration of the links divided by the length of T . The *degree* of one node is the *number of links* of the set of links attached to the node. The *density* $\delta(S)$ is the probability, given two nodes and a time instant, that a link exists between the two nodes:

$$\delta(S) = \frac{\sum_{e \in E} |T_e|}{\sum_{u, v \in V \otimes V} |T_u \cap T_v|}$$

Figure 1c provides a new distinctive look at the interactions happening between the individuals over 10 units of time. M_1 is the node with the highest degree of value 0.55. We can also observe that F_1 and F_2 met twice

shortly. F_1 and M_1 are the first to meet. The density of the stream graph is $\delta(S) = \frac{16}{20 \cdot 4 \cdot 3} \approx 0.06$.

2.2 Temporality, Multiple Layers, and Centrality

Now that we have introduced both the multilayer graph and stream graph models, let us briefly review prior works relevant to both temporal and multilayer approaches.

We may first be interested by the dynamics in multilayer networks [1, 2, 7, 11, 15, 18, 20, 21]. In temporal multilayer networks, one goal is often to identify community structures [1, 2, 18], which is not the task we are focusing on this paper. However, to model such networks, Kivelä *et al.* [15] suggest that temporality is only an *aspect* of the multilayer networks that could decompose the multilayer network like any other aspect. The same approach is taken by Pilosof *et al.* [20]. Nonetheless, unlike any aspect, time is submitted to order. Moreover, these analyses only consider juxtaposed time-frames as separated networks. Considering the changes of topologies in an overall graph is however relevant for the study of spreading processes [7, 11], which demonstrated the dependency on spreading from layers coupling. These works also point out how the evolution of centrality is key to studying the network [24], and isolating nodes of interest, such as in citation networks [21].

As of heterogeneity in stream graphs, we may first bring forward the Δ -analysis [16], which provides a way of studying interactions at multiple time scales; it may be regarded as a specific case of multilayer. Some approaches are more hybrid between the two. Vaiana *et al.* [26] are the first to bring a hybrid model between dynamic multilayer and temporal stream of links to capture different functional networks in the brain, but they mostly use the multilayer dynamic approach and introduced a unifying definition as a future work, which is in line with our contribution.

Both dynamic multilayer graph models or stream graphs would work with temporal data, however each imposes a specific point of view on the data. Each time frame in a dynamic multilayer network imposes to choose a time granularity, the choosing of which is not trivial. In addition, it also implies some distortions: either parts of links duration would be excluded (outside of the time frame) or a link would be considered as present over the whole frame considered. Multilayer stream-graphs, our proposed model, are completely agnostic to these issues, since they take each link in its own duration. The focus of this model is not on a whole graph interacting, but closer to the data, on a series of interaction events, no matter the nodes they attach. The structure of the resulting graph is only a consequence of these interactions.

3 Multilayer Stream Graph

Let us now introduce our object. A *multilayer stream graph* S_M is a tuple $(T, V, \mathcal{L}, L_M, V_M, W_M, E_M)$, with T a time interval, V a set of nodes, and

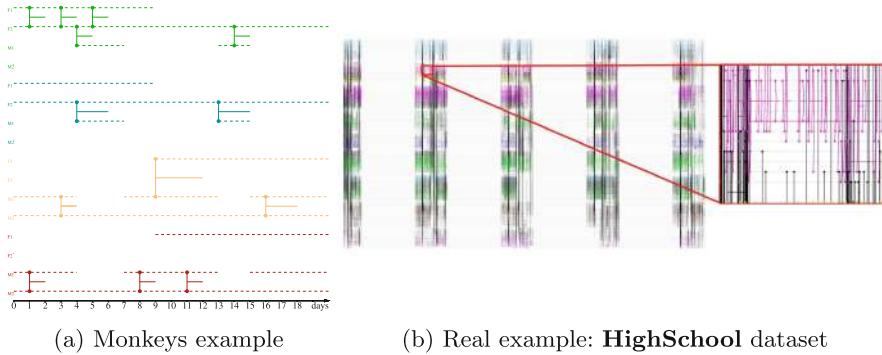


Fig. 2. (a) This example follows the colors of Fig. 1b: in green, (*collaboration, mountain*); (*fight, mountain*), in blue; (*collaboration, plain*), in yellow; and t(*fight, plain*), in red. This example captures 13 interactions of the 4 monkeys in a time frame [0, 20]. (b) The visualization of the much larger multilayer stream graph for dataset **HighSchool** composed of 36732 links involving 329 students over the course of 5 days.

$\mathcal{L} = \{\mathcal{L}_i\}_{i=1}^d$ a set of *aspects*. For a given $i \leq d$, $\epsilon \in \mathcal{L}_i$ is called an *elementary layer*; finally, we call a *layer* an element of $L = \mathcal{L}_1 \times \mathcal{L}_2 \times \dots \times \mathcal{L}_d$. We denote by $V_M \subseteq V \times L$ the set of node layers, and by $W_M \subseteq T \times V \times L$ the set of time-nodes-layers. In other words, $(u, \alpha) \in V_M$ means that node u is present in layer α , and $(t, u, \alpha) \in W_M$ means that node u is present at time t in layer α . $L_M \subseteq I \times L$ is the set of time-layers, where $I = \{[a, b] | a, b \in T, b \geq a\}$ is the set of intervals included in T . (t, α) in L_M means that layer α exists at time t . Finally, we denote by $E_M \subseteq T \times V_M \otimes V_M$ the set of interactions. In other words, $(t, (u, \alpha), (v, \alpha')) \in E_M$ means that node u in layer α and node v in layer α' interacted at time t .

Similarly to stream graphs, we denote $T_{u, \alpha}$ the set of times involving node-layer (u, α) , i.e. $T_{u, \alpha} = t : \exists(t, u, \alpha) \in W_M$, and $T_{(u, \alpha)(v, \beta)}$ the set of times where (u, α) and (v, β) interact, i.e. $T_{(u, \alpha)(v, \beta)} = t : \exists(t, (u, \alpha), (v, \beta)) \in E_M$.

We illustrate this object in ???. Notice that $(t, ((u, \alpha), (v, \beta))) \in E_M$ implies that $(t, u, \alpha) \in W_M$ and $(t, v, \beta) \in W_M$. Similarly, $(t, u, \alpha) \in W_M$ implies that $(t, \alpha) \in L_M$. In other words, a link between two nodes-layers can only exist if the two nodes-layers exist at this time, and nodes-layers can only be present when layers exist.

In the rest of this paper, for the sake of defining the most elementary object possible, we consider multilayer stream graphs to be undirected, unweighted and unlabeled. We show that even this elementary model is relevant for real-world data analysis. We discuss some possible extensions in Sect. 5.

Let us now start by defining various extractions and projections of multilayer stream graphs.

The *induced multilayer graph* by the set $\tau \subseteq T$ is a multilayer graph $M_I(S_M) = (V_{M,I}, E_{M,I}, V, L)$ which gathers all the layers, nodes-layers and links existing over time τ . In this graph, $V_{M,I} = \{v \in V | \exists t \in \tau, (t, v) \in V_M\}$ and $E_{M,I} = \{(v, \alpha), (w, \beta) \in (V \times L) \otimes (V \times L) | \exists t \in \tau, (t, v, \alpha, w, \beta) \in E_M\}$. In other words, it is the graph where nodes are elements of V and one puts a link between two nodes if and only if they have interacted over a duration τ .

Notice that when $\tau = t$, this induced graph is called the *multilayer graph at time t* . Figure 1b shows the multilayer graph induced by T for the multilayer stream graph in Fig. 2a.

Interlayer links are often used to model transit in a multilayer, such as the underground path to change between two lines of a subway station [15]. Given a pair of layers $\alpha, \beta \in L_1 \times \dots \times L_d$, the *interlayer stream graph* is the bipartite stream graph $S^{(\alpha, \beta)} = (T^{\alpha, \beta}, V^{\alpha, \beta}, W^{\alpha, \beta}, E^{\alpha, \beta})$, with $T^{\alpha, \beta} = \{t : \exists(t, u, \alpha) \in W_M, \exists(t, u, \beta) \in W_M, u \in V\}$ the interval of time in which α and β appear simultaneously. $V^{\alpha, \beta} = (V \times \{\alpha, \beta\}) \cap V_M$ is the set of node-layers of the multilayer stream graph restricted to α and β . $W^{\alpha, \beta} = (T^{\alpha, \beta} \times V^{\alpha, \beta}) \cap W_M$ describes all their intervals of existence. Finally, $E^{\alpha, \beta} = \{(t, (u, \alpha)(v, \beta)) \in E_M : t \in T^{\alpha, \beta}, u, v \in V\}$ is the set of interactions between layers α and β .

For a given layer α , we define the *intralayer stream graph* simply as $S(\alpha, \alpha)$, i.e. the interlayer stream graph between layer α and α . We denote it by $S^\alpha = (T^\alpha, V^\alpha, W^\alpha, E^\alpha)$. For example, Fig. 2a captures only intralayer interactions.

The *aggregated stream graph* $S_A(S_M) = (T, V, W_A, E_A)$ is the stream graph where all layer information has been removed. As such, it has the same interval of study T as S_M . Its nodes are the same as in S_M (the set V). Their times of existence are the union of their times of existence on the different layers: $T_u = \bigcup_{\alpha \in L} T_{u,\alpha}$ and $W_A = \bigcup_{u \in V} T_u \times \{u\}$. An edge exists between two nodes of $S_A(S_M)$ if it exists at the same time between two correspondent node-layers of S_M , i.e. $E_A = \{(t, u, v) | \exists(\alpha, \beta) \in L^2, (t, (u, \alpha), (v, \beta)) \in E_M\}$. For example, Fig. 1c is the aggregated stream graph obtained by superimposing all layers in Fig. 2a.

The *degree* of a node in a multilayer stream graph is the number of links (as defined just before) in which the node appears, i.e. $d(u) = \frac{1}{|T|} |\{(t, (u, \alpha)(v, \beta)) \in E_M : t \in T, v \in V, \alpha, \beta \in L\}|$. Similarly, the degree of a node-layer (u, α) is simply $d(u, \alpha) = \frac{1}{|T|} |\{(t, (u, \alpha)(v, \beta)) \in E_M : t \in T, (v, \beta) \in L\}|$.

In Fig. 2a, we can notice that females interact much more in the mountain than in the plain, and the contrary for the males. We can spot that the longest interaction, between F_1 and M_1 , takes place in the plain and lasts 3 days. The node with the highest degree is M_1 ($d = 7$) and the node-layer with the highest degree is $(M_1, (collaboration, plain))$.

The *density* of a multilayer stream graph is the probability, when one takes a random time t and two random node-layers (u, α) and (v, β) that the link $(t, (u, \alpha), (v, \beta))$ is in E_M : $\delta(S_M) = \frac{\sum_{(u, \alpha), (v, \beta) \in E_M} |T_{(u, \alpha)(v, \beta)}|}{\sum_{(u, \alpha), (v, \beta) \in E_M} |T_{(u, \alpha)} \cap T_{(v, \beta)}|}$.

In Fig. 2a, the density of the multilayer stream graph is: $\delta(S_M) = \frac{18}{104} \approx 0.17$. Notice that in comparison, the density of the aggregated stream graph is $\delta(S) = \frac{17}{20*6} \approx 0.14$, and the one of the aggregated graph is $\delta(G) = \frac{4}{6} = \frac{2}{3}$.

Moreover, this definition of density can be readily applied and adapted to specific cases. For example, the interlayer density of interactions between two layers α and β is nothing but $\delta(S^{(\alpha,\beta)})$, the density of the interlayer stream graph. The denominator sum can also be modified to take into account specific aspects of the data, for example by summing on $(u, \alpha), (v, \alpha)$ if interlayer links are not allowed, among others.

3.1 Centralities

One key application on real-world data is the analysis and detection of important nodes, *i.e.* that are *central*. Many notions of centrality coexist for graphs [4, 5], multilayer graphs [12, 22] and stream graphs [6, 16] alike. As of today, no consensus emerges on a global centrality notion, as they all capture different notions of importance [13].

In this section, we develop upon the formalism introduced in Sect. 3 and introduce two centrality definitions on multilayer stream graphs extending from entanglement [22, 23] and inspired by eigenvector centrality [13], taking into account the multifaceted nature of the object while remaining simply explainable and computationally efficient.

As a prerequisite, let us extend the definition of *paths* to multilayer streams graphs. A path from (t, u, α) to (t', v, β) is a sequence $(t_i, (u_i, \alpha_i), (v_i, \beta_i))_{i=0}^k$ of elements of E_M such that $(u_0, \alpha_0) = (u, \alpha)$, $(v_k, \beta_k) = (v, \beta)$, $t_0 \geq t$, $t_k \leq t'$ and for all $i = 0..k$, $(u_{i+1}, \alpha_{i+1}) = (v_i, \beta_i)$ and $t_{i+1} \geq t_i$. A common variant, defined in [16], are γ -paths, *i.e.* paths for which the condition $t_{i+1} \geq t_i$ becomes $t_{i+1} \geq t_i + \gamma$. In other words, in γ -paths traversing an edge costs γ . This is especially useful for modelling transportation networks, as we will see in Sect. 4.

Let us now introduce the new notions of centrality, that we call *superimposed layer centrality* and *juxtaposed layer centrality*. Both aim at giving an intuition of the importance of *layers* in the multiplex stream graph.

A group of layers is *superimposed* if each node can be present on each layer at any time (also referred as multiplex networks [15]). In other words, saying that two layers are superimposed means that it is possible to have interlayer links between those layers. This typically corresponds to layers describing diverse types of relationships across the same set of nodes. For instance, in Fig. 2a, the layers (*mountain, collaboration*) and (*mountain, fight*) are superimposed.

Given a superimposed multilayer stream graph and a time $t \in T$, for all nodes $u \in V$ and all layers $\alpha_{i=0..k}$, let us compute $X_{\alpha_i}(t, u)$ as the probability that a random walker starting from node u at time t will cross a link involving layer α_i . One then obtains a $|V| \times k$ matrix corresponding to the relative importance of each layer α_i for each node u .

We define the *superimposed layer centrality* as the maximal eigenvalue of Σ_X , the matrix of covariances of all random walkers. In Σ_X , each term of the matrix is computed as $X_{\alpha_i, \alpha_j} = E[(X_{\alpha_i} - E[X_{\alpha_i}])(X_{\alpha_j} - E[X_{\alpha_j}])]$. Intuitively,

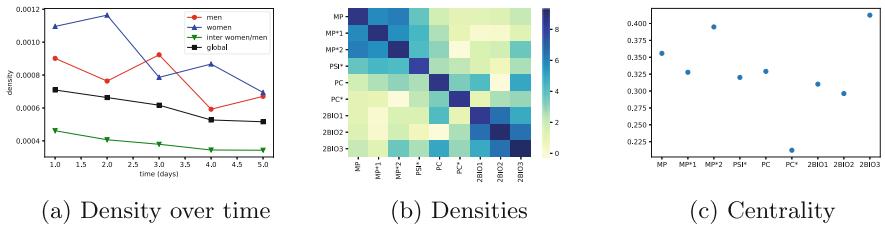


Fig. 3. (a) Density in multilayer stream graph computed for each day, inside the layers of men, the layers of women, between the two groups of layers and inside the whole group. (b) The matrix of the log of densities between the classes. (c) Score of centrality for the different layers.

the eigenvalues of Σ_X give a ranking of the layers by decreasing importance. The maximal eigenvalue corresponds to the maximal variance for a linear combination of X_{α_i} corresponding to the eigenvector of Σ_X .

However, not all datasets contain meaningful superimposed layers. To this end, we define another notion of centrality, the *juxtaposed layer centrality*.

A group of layers is *juxtaposed* if each node can only be present in one layer. This is typically the case for non-superposed states: age, gender, class number, etc. Notice however that the layer associated to each node can in principle change over time. For example if one were to consider the aspect $age = \{baby, child, adolescent, adult, elderly\}$ in Fig. 2a, the layers (*child, mountain, collaboration*) and (*adult, mountain, collaboration*) are juxtaposed. In this case, studying the relations between layers is particularly relevant.

We then consider the interlayer density matrix Δ , i.e for each pair i, j of layers, one computes the interlayer density $\delta(S_M(\alpha_i, \alpha_j))$ ¹.

From the multilayer stream graph displayed in Fig. 2a, we can alternatively consider two layers of: *male* interactions, and *female* interactions. The matrix of densities for these layers (*male* and *female*) is $\begin{pmatrix} 1/5 & 1/7 \\ 1/7 & 1/6 \end{pmatrix}$. One can notice that in this (toy) example, interactions are stronger in the males than in the females and that the females and the males interact less with their opposite gender.

The *juxtaposed layer centrality* then correspond for each layer to its entry in the eigenvector associated with the maximum eigenvalue of Δ . Notice that in both cases, the Perron-Frobenius theorem [10, 19] states that a irreducible non-negative matrix has a maximum positive eigenvalue with an eigenspace of range 1. In our case, we know that Δ is non-negative by definition of the densities and irreducible unless we can share the layers into different groups that do not interact together.

¹ With the case where $i = j$ simply returns the intralayer density $\delta(S_M(\alpha_i))$.

4 Results

We now demonstrate multilayer stream graphs for the analysis of two real-world datasets. All implementations are available to the public².

4.1 Data

The first one records interactions among high-school students [17]. Each student is associated to a class, and interactions can be of three kinds: (i) face-to-face, (ii) self-declared friendship, and (iii) Facebook friendship. Notice that only (i) is time-dependent, (ii) is directed, and (iii) is undirected.

This dataset comprises of 36,732 links involving 329 students over the course of 5 days. *Superimposed aspects* are the interaction type (face-to-face, friendship, Facebook friendship), whereas *juxtaposed aspects* are the gender of each student (female, male, or undefined) and the class. The (relatively) small size of this dataset makes it visualizable, see Fig. 2b. We will hereafter refer to this dataset as **HighSchool**.

The second one documents all domestic flights in the United States since 1987 [25]. Since the whole dataset is too large to be efficiently processed, we focus instead on a longitudinal study across the years, using all the flights in January 1988, 1995, 2010 and 2019.

Each of these datasets involves 346 airports and contains roughly 500000 flights. There are no *juxtaposed aspects*, however the company operating the flight is a natural *superimposed aspect*. There is a maximum of 17 companies. Finally, while the stream graph is too large to be visualized, we display the induced graph over a map of the United States in Fig. 4a. We will hereafter refer to this dataset as **US-Flights**.

4.2 Experiments

We now devise two experiments in order to shed light into some patterns present in the two datasets described in Sect. 4.1.

The intrinsic structure of the two datasets allow us to demonstrate the relevance of both notions of centrality described in Sect. 3.1. We show that our formalism is able to shed some light on patterns in both datasets, which in turn serves as a proof of concept for our work.

The **HighSchool** dataset features gender information about the participants. In this context, we investigate the interaction patterns between the participants, taking into account this information.

Let us consider the multilayer stream graph describing the interactions among students, the layers here being the gender ($\{M, F\}$)³ and the class label

² <https://github.com/TiphaineV/multiplex-streams/src/visualisation>.

³ The dataset also contains a few ‘U’, for Undefined, that corresponds to interactions with teachers. We do not consider them in the rest of this work.

($\{MP, MP^*1, MP^*2, PSI^*, PC, PC^*, 2BIO1, 2BIO2, 2BIO3\}$), corresponding to usual French names for such schools.

Regarding gender, let us consider the multilayer stream graph induced by every 24 hours. We obtain 5 such stream graphs, corresponding to the 5 days of the dataset recording. On each daily multilayer stream graph, we compute the inter and intralayer densities, as well as the graph density (*i.e.* discarding all gender and temporal information). We show the result in Fig. 3. Several conclusions can be observed: first of all, the graph density (legend “global”) does not adequately capture the subtleties in the data. It averages the intra and inter-densities, that are in reality following two different modes; in other words, individuals interact more with individuals of the same gender as them.

We also study the densities of interactions between the different classes. Figure 3a shows the density matrix for each class. For readability, it displays the absolute value of the logarithm of the densities, as it makes blocks more apparent. While as for gender, the intra-density is higher than the inter-density, we can discern larger blocks grouping layers together: $\{MP, MP^*1, MP^*2\}$ $\{2BIO1, 2BIO2, 2BIO3\}$. These blocks correspond to specialty topics, as MP corresponds to mathematics and physics, while BIO corresponds to biology. This result is intuitive, and so this serves as an argument that our model captures the interaction subtleties in the data.

Figure 3 shows the superimposed layer centrality values for each class, as defined in Sect. 3.1.

$2BIO3$ and MP^*2 are the most central: we can see that they are the most central among two clusters of layers, regrouping the MP classes and the BIO classes. In terms of data analysis, this makes sense since these two groups correspond to common specialty subdivisions in the French system: favouring Biology (BIO) or Mathematics and Physics (MP). Notice finally that the PC (Physics and Chemistry) layers in this school are the least central; one can then assume that the students in these classes interact less in time with other classes, which comes as a surprise considering that one specialty, Physics, is shared with the MP students.

Notice however that the **HighSchool** dataset does not have juxtaposed layers that we can study. In order to demonstrate the interest of the juxtaposed layer centrality, we focus instead on the **US-Flights** dataset.

On this dataset, we show in Fig. 4 the correlation between the probability of coverage by a random walker and the layer centrality value we compute. For each company α corresponding to each layer, and for a given $t > t_0$, we compute the random variable $X_\alpha(v, t) = \sum_{(t, (u, \alpha), (v, \alpha)) \in E_M} (t_{max} - t)$. In other words, the probability to take a plane decrease with time, until it reaches 0 after a certain amount of time.

Given these probabilities, we compute then the covariance matrix of those variables, as explained in Sect. 3.1. The eigenvalues of the covariance matrix corresponds to the centrality score. Notice that, however, just like in graphs, the centrality scores themselves mean little, as it is their relative order that carries importance.

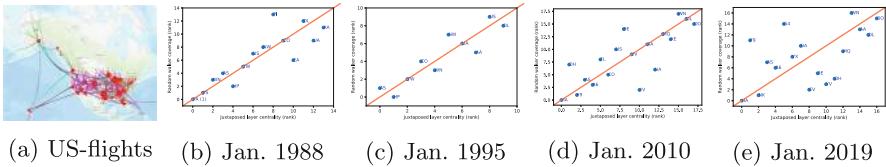


Fig. 4. (a) Induced multilayer graph of the US companies flights. The size of the nodes reflect of their PageRank. (b–e) Comparison between rank by coverage and juxtaposed layer rank of companies for the 4 subdatasets extracted from **US-Flights**. In red, the function $y = x$, corresponding to a perfect correlation.

In order to assess the usefulness of our metric, we show in Fig. 4 the rank of each company compared to the relative coverage of each company by a random (temporal) walker. In the four subdatasets that we consider, we can see that our centrality is well correlated to the random walker coverage, though however this is especially true for the older subdatasets (1988 and 1995).

We notice that the layer centrality fits less and less with the one of random walker over time. This is due to the fact that the score of each layer in the eigenvector tends to be the same. This means that a lot of companies tends to look like the other ones. We can guess that with the improvement of the study of the market, the carriers have found what are the more interesting routes and concentrates on the same ones. The number of flight has increased a lot (436,951 per month in 1988 to 583,986 in 2019) but probably on the most popular routes rather than to create new connections.

5 Conclusion

In this paper, we devise a new formalism that bridges the gap between two recent advances in the state-of-the-art: multilayer graphs and stream graphs. We propose a new framework that generalizes both objects, and define some elementary notions on it, in order to show its relevance. Furthermore, we introduce two notions of *layer centrality* that capture the relative importance of layers over time. We experiment on two interaction datasets, of individual contacts and flight information, and show the relevance of the formalism and centralities at capturing subtle patterns in the data.

This work is intended only as a validation for the multilayer stream graph model, and as such it opens numerous perspectives. The first of them relates to the formalism itself: while the model we define is straightforwardly usable, it can be extended in many ways. While some of these are straightforward, such as orientation, others require more thorough work, such as weighting, or proper label utilization.

Another interesting axis depends on the data itself. While many examples of multilayer stream graphs exist in real life, all the relevant information is not typically captured in datasets, typically because the current models cannot use the extra information. We hope this paper serves as a wider call to researchers of many disciplines, to use our model and tailor it to their needs.

References

1. Amelio, A., Pizzuti, C.: Evolutionary clustering for mining and tracking dynamic multilayer networks. *Comput. Intell.* **33**(2), 181–209 (2017)
2. Bassett, D.S., Porter, M.A., Wymbs, N.F., Grafton, S.T., Carlson, J.M., Mucha, P.J.: Robust detection of dynamic community structure in networks. *Chaos Interdiscip. J. Nonlinear Sci.* **23**(1), 013142 (2013)
3. Biggs, N., Lloyd, E.K., Wilson, R.J.: *Graph Theory*, pp. 1736–1936. Oxford University Press, Oxford (1986)
4. Brandes, U.: A faster algorithm for betweenness centrality. *J. Math. Sociol.* **25**(2), 163–177 (2001)
5. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **30**(1–7), 107–117 (1998)
6. Costa, E.C., Vieira, A.B., Wehmuth, K., Ziviani, A., da Silva, A.P.C.: Time centrality in dynamic complex networks. *CoRR* abs/1504.00241 (2015). <http://arxiv.org/abs/1504.00241>
7. De Domenico, M., Granell, C., Porter, M.A., Arenas, A.: The physics of spreading processes in multilayer networks. *Nat. Phys.* **12**(10), 901–906 (2016)
8. De Domenico, M., Solé-Ribalta, A., Cozzo, E., Kivelä, M., Moreno, Y., Porter, M.A., Gómez, S., Arenas, A.: Mathematical formulation of multilayer networks. *Phys. Rev. X* **3**(4), 041022 (2013)
9. Euler, L.: Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, pp. 128–140 (1741)
10. Frobenius, G.F., Frobenius, F.G., Frobenius, F.G., Frobenius, F.G., Mathematician, G.: Über matrizen aus positiven elementen. *Königliche Akademie der Wissenschaften* (1908)
11. Gallotti, R., Barthelemy, M.: The multilayer temporal network of public transport in Great Britain. *Sci. Data* **2**, 140056 (2015)
12. Ghalmame, Z., Cherifi, C., Cherifi, H., El Hassouni, M.: Centrality in complex networks with overlapping community structure. *Sci. Rep.* **9** (2019)
13. Gupta, N., Singh, A., Cherifi, H.: Centrality measures for networks with community structure. *Physica A Stat. Mech. Appl.* **452**, 46–59 (2016)
14. Hopkins, B., Wilson, R.J.: The truth about königsberg. *Coll. Math. J.* **35**(3), 198–207 (2004)
15. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. *J. Complex Netw.* **2**(3), 203–271 (2014)
16. Latapy, M., Viard, T., Magnien, C.: Stream graphs and link streams for the modeling of interactions over time. *Soc. Netw. Anal. Min.* **8**(1), 61 (2018)
17. Mastrandrea, R., Fournet, J., Barrat, A.: Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLoS ONE* **10**(9), e0136497 (2015)
18. Mourchid, Y., Renoust, B., Cherifi, H., El Hassouni, M.: Multilayer network model of movie script. In: International Conference on Complex Networks and their Applications, pp. 782–796. Springer (2018)
19. Perron, O.: Zur theorie der matrices. *Mathematische Annalen* **64**(2), 248–263 (1907)
20. Pilosof, S., Porter, M.A., Pascual, M., Kéfi, S.: The multilayer nature of ecological networks. *Nat. Ecol. Evol.* **1**(4), 0101 (2017)
21. Renoust, B., Claver, V., Baffier, J.F.: Multiplex flows in citation networks. *Appl. Netw. Sci.* **2**(1), 23 (2017)

22. Renoust, B., Melançon, G., Munzner, T.: Detangler: visual analytics for multiplex networks. *Comput. Graphics Forum* **34**(3), 321–330 (2015)
23. Renoust, B., Melançon, G., Viaud, M.L.: Entanglement in multiplex networks: understanding group cohesion in homophily networks. In: Social Network Analysis–Community Detection and Evolution, pp. 89–117. Springer (2014)
24. Taylor, D., Myers, S.A., Clauset, A., Porter, M.A., Mucha, P.J.: Eigenvector-based centrality measures for temporal networks. *Multiscale Model. Simul.* **15**(1), 537–574 (2017)
25. United States Department of Transportation: Bureau of transportation statistics (2019). https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236. Accessed 16 Mar 2019
26. Vaiana, M., Muldoon, S.F.: Multilayer brain networks. *J. Nonlinear Sci.* 1–23 (2018)



Better Late than Never: A Multilayer Network Model Using Metaplasticity for Emotion Regulation Strategies

Nimat Ullah^(✉) and Jan Treur

Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
nimatullah09@gmail.com, j.treur@vu.nl

Abstract. Adaptivity in emotion regulation strategies has always been considered as one of the key factors for health. As the choices of emotion regulation strategies change as per context, the priorities of strategies also change with time. This phenomenon is called plasticity. This paper focuses on network-oriented modeling of the concept of metaplasticity from recent neurological literature which controls the plasticity. Simulation results are presented for the elaboration of the concept in choice of emotion regulation strategies with age.

Keywords: Plasticity · Metaplasticity · Emotion regulation · Gender · Reappraisal · Expressive suppression

1 Introduction

In the choice of emotion regulation strategies, flexibility per context is a well-established fact now in cognitive and social sciences [1]. Research considering the changes in emotion regulation strategies with time can be found, for example, in [2]. A factor “changes with time” can be described by changes in emotion regulation strategies with age [3]. A vast body of literature is available showing that changes in the choice of emotion regulation strategies, caused by various factors, take place as a person grows [4]; this is also highlighted by socioemotional selectivity theory (SST) [2]. These changes are referred to as *plasticity* in the literature from the neurocognitive sciences.

Recently, an increasing amount of work has been reported about *metaplasticity* [5, 6] or plasticity of plasticity. For emotion regulation, these concepts apply to the adaptive changes that take place in the choice of emotion regulation strategies over time with age.

This paper extends the work presented in [7] which focuses on age and gender differences in choice of emotion regulation strategies. When we say “gender” we refer to the unary biological as well as social role, i.e., male vs female throughout the paper. In this paper the concepts of plasticity and metaplasticity [5, 6] have been applied for age differences in choice of emotion regulation strategies by using the modeling approach for multi-layered adaptive networks and its supporting software environment described in [8–10]. Complexity of the emerging behaviour addressed in this paper lies in the multiple orders of adaptivity, and the dynamic and adaptive interaction between the layers of the obtained network.

Plasticity of plasticity in emotion regulation is a novice concept in the field of AI and network-oriented modeling. In the multi-layered network-oriented modeling approach used from [9, 10], a base network model is extended by on top of it adding two layers, respectively, for first-order adaptation of (some of) the base network connections, and for second-order adaptation by control of the first-order adaptation speed and intensity of the changes that take place over time. Simulation results are reported to illustrate the network behaviour emerging from the interaction (or co-evolution) between the three layers (base network dynamics and first- and second-order adaptation dynamics). In rest of the paper, Sect. 2 presents a theoretical background for the model. Section 3 presents the multilayer network model, Sect. 4 presents simulation experiments of the model. Finally, Sect. 5 concludes the paper.

2 Background

Shift in choice of emotion regulation strategies occurs as a person grows and it involves many factors that influence this shift from one strategy to another strategy. According to SST [2], this shift is because of the time constraint being experienced by older adults which alters their motivational goals. Similarly, [11] states that younger and older adults use different kinds of emotion regulation strategies and age-specific developmental increase and decrease takes place in the use of emotion regulation strategies. Moreover, According to SST [2] older adults turn to more use of antecedent-focused strategies like reappraisal from response-focused strategies like suppression. In line with these findings, [12] found an increased use of reappraisal and decrease use of suppression with age (from 20 to 60). A reason for this shift from response-focused strategies to antecedent-focused strategies may be found in the “strength and vulnerability integration theory” [13] stating that as physiological flexibility decreases with age, it becomes difficult to implement response-focused strategies. Therefore, older adults may use more antecedent-focused strategies.

Similarly, [14] state that older people are “more likely” to reappraise than younger adults. In terms of control in emotional situations, older people are better in controlling their emotions [15, 16] and quicker in returning to a positive mood after a negative mood [17, 18] in comparison to younger adults. Confirming these findings, [19, 20], not only cognitive reappraisal is considered to be efficient in downregulating negative emotional experiences, it also helps in decreasing the psychological distress, which is reversed to suppression. In case of suppression, distress still remains high even if the expression of emotion is suppressed successfully [21]. In line with this discussion, [22] suggest that being an effective emotion regulation strategy, it is cognitive reappraisal that helps older people in retaining a positive emotional state. In contrast, younger adults prefer confrontational coping, as reported by [23]. The reason of an increased use of reappraisal can also be because older adults need fewer cognitive resources [24] as compared to their younger counterparts to down-regulate negative emotional response. Increased use of reappraisal and decreased use of suppression is also supported by series of findings like [25] reporting that emotional wellbeing improves with age from adulthood to early old age, and [26] reporting that use of reappraisal helps in emotional wellbeing when compared to expressive suppression. This supports the idea

that decrease in negative affect and increase in positive affect occurs due to the increase in the use of reappraisal as an emotion regulation strategy with increasing age [3].

So, the more general concept of plasticity also applies to the experience of emotions. Also, in other literature this is indicated. For instance, [27, 28] put forward that the intensity of negative emotions decreases with age while the intensity of positive emotions either remains stable or increases with age. Moreover, older people tend to consider a stressful situation less threatening [29] in comparison to younger adults and give weaker negative reaction [30, 31]. Studies like [32] also found decreased rate of depression/anxiety in older people in comparison to younger adults. Similarly, [33] also support the notion that frequency of negative affect decreases with age while positive affect remains stable. Various individual differences can be found in self-regulation in adults [34]. However, overall, findings suggest that some development takes place in emotional experience and regulatory capabilities into the second half of life. At the same time, studies like [35] also found decline in cognitive resources with age that too is subjected to individual differences.

The concept of plasticity in emotion regulation strategies has long ago been defined by [36] stating that maturity in cognition is bound to improvements in cognitive reappraisal and older people exhibit more of this cognitive maturity than their younger counterparts [37]. Similarly, [38, 39] state that flexibility in goal adjustment increases with age. These findings provide a strong base for the network model introduced here.

3 The Multilayered Network Model

This section describes the computational model presented in this paper as well as the Network-Oriented Modeling approach for adaptive networks based on network reification [8–10] that has been employed for designing the network model, and performing simulations with it.

The multi-layered network model presented in Fig. 1. demonstrates the phenomena of plasticity and metaplasticity. Table 1 provides overview of the various states of the model. The first (bottom) layer describes the base level, which shows the basic processes of the two strategies, i.e., expressive suppression and cognitive reappraisal. Expressive suppression suppresses the expression of the emotion while sensory representation and the negative belief about the stimulus still remain high. In contrast, reappraisal changes the beliefs about the stimulus which, as a result, decreases the intensity of the negative emotion while increasing positive emotions about the stimulus.

The second layer describes first-order network adaptation at the first reification level, which demonstrates the Hebbian learning process taking place throughout one's life in various forms. Here, the person learns about which emotion regulation strategy to use over time. This happens by changing the \mathbf{W} states that provide reified representations for the connection weights used at the base level. Initially, the person is using expressive suppression in younger ages. The use of reappraisal increases with the increase in age, which discourages the use of expressive suppression at the base level.

Table 1. Overview of the states of the multi-layered network model in Fig. 1.

State	Explanation	Level
X_1	ws_s	Base level
X_2	ss_s	
X_3	srs_s	
X_4	ps_a	
X_5	es_a	
X_6	ss_b	
X_7	srs_b	
X_8	fs_b	
X_9	ps_b	
X_{10}	es_b	
X_{11}	bs_-	
X_{12}	bs_+	
X_{13}	cs_{reapp}	
X_{14}	cs_{sup}	
X_{15}	$ms_{dstress}$	
X_{16}	$\mathbf{W}_{fs_b, cs_{reapp}}$	First reification level
X_{17}	$\mathbf{W}_{fs_b, cs_{sup}}$	
X_{18}	$\mathbf{Mw}_{fs_b, cs_{reapp}}$	Second reification level
X_{19}	$\mathbf{Hw}_{fs_b, cs_{reapp}}$	
X_{20}	$\mathbf{Hw}_{fs_b, cs_{sup}}$	
X_{21}	$\mathbf{Mw}_{fs_b, cs_{sup}}$	

The third layer describes second-order adaptation at the second reification level, which controls the speed as well as persistence factor of the first-order learning phenomena at the first reification level. It uses **H** states and **M** states as reified representations. The **H** states control the adaptation speed factors for the first-order adaptation, modeled by their respective **W** states, and the **M** states control the persistence level of the first-order adaptation, modeled by their respective **W** states.

Hebbian learning can take place for connections between any two states at the base level. For instance, the connection from fs_b to cs_{reapp} in the base model is adaptive but the connection weight is no state of the base level. Instead, this weight is represented by state $\mathbf{W}_{fs_b, cs_{reapp}}$ at the first reification level. As learning itself is subjected to change too, its change in adaptation speed and persistence is controlled by states at the 2nd reification level. For instance, the speed and persistence of $\mathbf{W}_{fs_b, cs_{reapp}}$ (first reification level state) are controlled by the second reification level states $\mathbf{Hw}_{fs_b, cs_{reapp}}$ and $\mathbf{Mw}_{fs_b, cs_{reapp}}$.

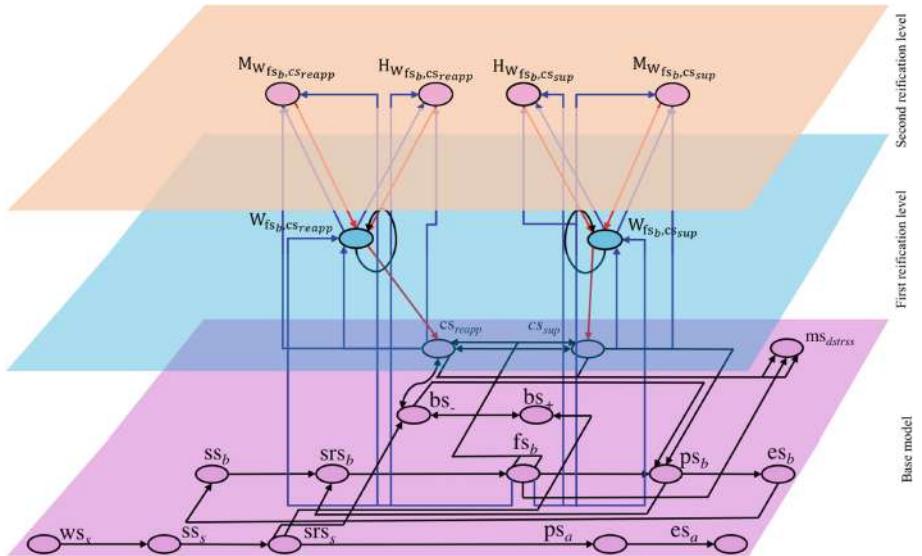


Fig. 1. Multi-layered adaptive network model for emotion regulation strategies with age.

representing this adaptation speed and persistence, respectively. Similarly, there is possibility of many other layers on top of this.

The full specification of the multi-layered network model by role matrices can be found in Box 1 and Box 2. Each matrix addresses some network characteristic and has rows according to all states X_j with in that row the data for that characteristic. Here **mb** specified the states with incoming connections to state X_j . In general, the green cells with values indicate the static (non-adaptive) network characteristics, and the red cells with names X_i indicate adaptive characteristics where the value of X_i plays the role of that characteristic. As an example, in the connectivity role matrix **mew** for connection weights, for the control states X_{13} and X_{14} (cs_{reapp} and cs_{sup}) at the base level it is indicated in the first column that the connection from the first base state X_8 (the feeling state fs_b) indicated in the first column in **mb** is adaptive and is represented by X_{16} resp. X_{17} ($W_{fs_b,cs_{reapp}}$ resp. $W_{fs_b,cs_{sup}}$).

The role matrices can be used as input for the modeling environment described in [9] and can be executed then. It represents the conceptual representation of the model and shows how each node is influenced by other nodes in the network. Starting from base network model to the second reification level, Box 1 and 2 specify these influences. For more background on the role matrices specification format used here, and the modeling environment, see [9, 10].

mb	connectivity:	1	2	3		mcw	connectivity:	1	2	3
base connectivity										
X_1	WS _s	X_1				X_1	WS _s	1		
X_2	SS _s	X_1				X_2	SS _s	1		
X_3	SRS _s	X_2				X_3	SRS _s	1		
X_4	ps _a	X_3				X_4	ps _a	0.1		
X_5	es _a	X_4				X_5	es _a	0.2		
X_6	ss _b	X_{10}				X_6	ss _b	1		
X_7	srs _b	X_9	X_6			X_7	srs _b	0.5	0.15	
X_8	fs _b	X_7				X_8	fs _b	1		
X_9	ps _b	X_8	X_{11}	X_{14}		X_9	ps _b	0.4	0.5	-0.9
X_{10}	es _b	X_9				X_{10}	es _b	1		
X_{11}	bs ₋	X_3	X_{13}	X_{12}		X_{11}	bs ₋	0.6	-0.7	-0.4
X_{12}	bs ₊	X_3	X_{11}			X_{12}	bs ₊	0.4	-0.4	
X_{13}	CS _{reapp}	X_8	X_{11}			X_{13}	CS _{reapp}	X_{16}	0.2	
X_{14}	CS _{sup}	X_8	X_{13}			X_{14}	CS _{sup}	X_{17}	-1	
X_{15}	MS _{dstrss}	X_8	X_{13}	X_{14}		X_{15}	MS _{dstrss}	0.4	-0.4	0.4
X_{16}	W _{fs_b,cs_{reapp}}	X_8	X_{13}	X_{16}		X_{16}	W _{fs_b,cs_{reapp}}	1	1	1
X_{17}	W _{fs_b,cs_{sup}}	X_8	X_{14}	X_{17}		X_{17}	W _{fs_b,cs_{sup}}	1	1	1
X_{18}	MW _{fs_b,cs_{reapp}}	X_8	X_{13}	X_{16}		X_{18}	MW _{fs_b,cs_{reapp}}	1	1	1
X_{19}	HW _{fs_b,cs_{reapp}}	X_8	X_{13}	X_{16}		X_{19}	HW _{fs_b,cs_{reapp}}	1	1	1
X_{20}	HW _{fs_b,cs_{sup}}	X_8	X_{14}	X_{17}		X_{20}	HW _{fs_b,cs_{sup}}	0.6	0.8	0.8
X_{21}	MW _{fs_b,cs_{sup}}	X_8	X_{14}	X_{17}		X_{21}	MW _{fs_b,cs_{sup}}	0.6	0.8	0.5

Box 1 Role matrices for connectivity

In Box 1, matrix **mb** represents for any node of the network its incoming connections. These connections in **mb** are either between states at the same level or from lower level to higher level, i.e. no downward connection from a higher to a lower level, as the downward connections are the connections which effectuate adaptivity and are specified in the other role matrices. To the right, in matrix **mcw** the values between 0–1 represent connection weights of the incoming connections while the X_i represent the respective states in the higher level that represent and control the (incoming) adaptive connection to that specific state.

In Box 2 below, it can be seen that the Hebbian learning states in first reification level has downward incoming connections from the second reification level one each for speed factor and for persistence.

mcfw aggregation:		1	2	3		mcfp aggregation:		1	2	3		ms timing:		
combination	function weights	allogistic	hebb	id		combination parameters	func.	allogistic	hebb	id		speed factors		1
X_1	ws_s					X_1	ws_s					X_1	ws_s	0
X_2	ss_s					X_2	ss_s					X_2	ss_s	1
X_3	srs_s					X_3	srs_s					X_3	srs_s	1
X_4	ps_a					X_4	ps_a					X_4	ps_a	1
X_5	es_a					X_5	es_a					X_5	es_a	1
X_6	ss_b					X_6	ss_b					X_6	ss_b	1
X_7	srs_b	1				X_7	srs_b	10 0.3				X_7	srs_b	1
X_8	fs_b					X_8	fs_b					X_8	fs_b	1
X_9	ps_b	1				X_9	ps_b	10 0.3				X_9	ps_b	1
X_{10}	es_b					X_{10}	es_b					X_{10}	es_b	1
X_{11}	bs_-	1				X_{11}	bs_-	8 0.2				X_{11}	bs_-	1
X_{12}	bs_+	1				X_{12}	bs_+	8 0.2				X_{12}	bs_+	1
X_{13}	cs_{reapp}	1				X_{13}	cs_{reapp}	5 0.8				X_{13}	cs_{reapp}	0.1
X_{14}	cs_{sup}	1				X_{14}	cs_{sup}	12 0.2				X_{14}	cs_{sup}	0.4
X_{15}	ms_{dstrss}	1				X_{15}	ms_{dstrss}	8 0.5				X_{15}	ms_{dstrss}	0.5
X_{16}	$W_{fs_b,cs_{reapp}}$		1			X_{16}	$W_{fs_b,cs_{reapp}}$		X_{18}			X_{16}	$W_{fs_b,cs_{reapp}}$	X_{19}
X_{17}	$W_{fs_b,cs_{sup}}$		1			X_{17}	$W_{fs_b,cs_{sup}}$		X_{19}			X_{17}	$W_{fs_b,cs_{sup}}$	X_{20}
X_{18}	$MW_{fs_b,cs_{reapp}}$	1				X_{18}	$MW_{fs_b,cs_{reapp}}$	12 0.2				X_{18}	$MW_{fs_b,cs_{reapp}}$	0.3
X_{19}	$HW_{fs_b,cs_{reapp}}$	1				X_{19}	$HW_{fs_b,cs_{reapp}}$	4 0.2				X_{19}	$HW_{fs_b,cs_{reapp}}$	0.3
X_{20}	$HW_{fs_b,cs_{sup}}$	1				X_{20}	$HW_{fs_b,cs_{sup}}$	10 0.3				X_{20}	$HW_{fs_b,cs_{sup}}$	0.3
X_{21}	$MW_{fs_b,cs_{sup}}$	1				X_{21}	$MW_{fs_b,cs_{sup}}$	10 0.3				X_{21}	$MW_{fs_b,cs_{sup}}$	1

Box 2 Role matrices for aggregation and timing

4 Simulation Results

The scenarios addressed in this work are inspired by [7] where both age and gender have been considered for the difference in choice of emotion regulation strategies. The extended multi-layered network model introduced here only considers age for the choice in emotion regulation strategies. Novelty of the model is that it is a multi-layered adaptive network model with dynamic states which change in an adaptive manner with Hebbian learning. The Hebbian learning itself is controlled in an adaptive manner as well, as in real life. Table 2 provides the initial values of the states of the model.

Table 2. Initial values of the states

State	ws_s	All other base states	$W_{fs_b,cs_{reapp}}$	$W_{fs_b,cs_{sup}}$	$HW_{fs_b,cs_{reapp}}$	$HW_{fs_b,cs_{sup}}$	$HW_{fs_b,cs_{reapp}}$	$MW_{fs_b,cs_{sup}}$
Initial value	1	0	0.3	0.3	0.5	0.5	0.9	0.9

Figure 2 shows the speed factor and persistence factor reification states of the two reified \mathbf{W} states given in Fig. 3. It can be seen in the graph that both the factors are initially increasing for both the reified states but after some time $HW_{fs_b,cs_{sup}}$ and

$M_{W_{fsb},cs_{sup}}$ get decreasing. This phenomenon demonstrates metaplasticity wherein the learning itself is dynamic, i.e. increases/decreases with time. This makes the person either stick to previous emotion regulation strategies or switch from one strategy to another strategy after learning takes place over the years.

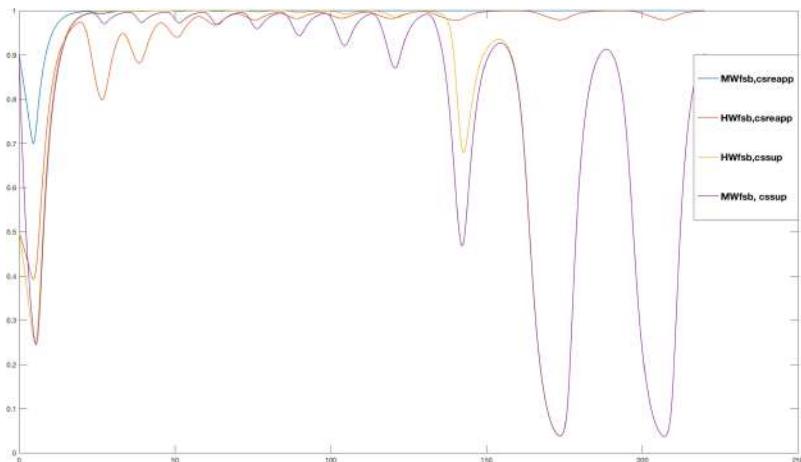


Fig. 2. Second-order reified representation states for speed and persistence factors.

In Fig. 3 the reified states are given, where $W_{fsb,cs_{reapp}}$ increases slowly and gradually while $W_{fsb,cs_{sup}}$ decreases slowly and gradually until it gets equal to zero. Initially, younger adults use suppression; therefore, $W_{fsb,cs_{sup}}$ is high. On the other hand $W_{fsb,cs_{reapp}}$ is, though, low but increasing slowly with increase in age. Finally, $W_{fsb,cs_{reapp}}$ is high enough to activate cs_{reapp} instead of cs_{sup} . This demonstrates the learning that takes place over the years and changes priorities over time.

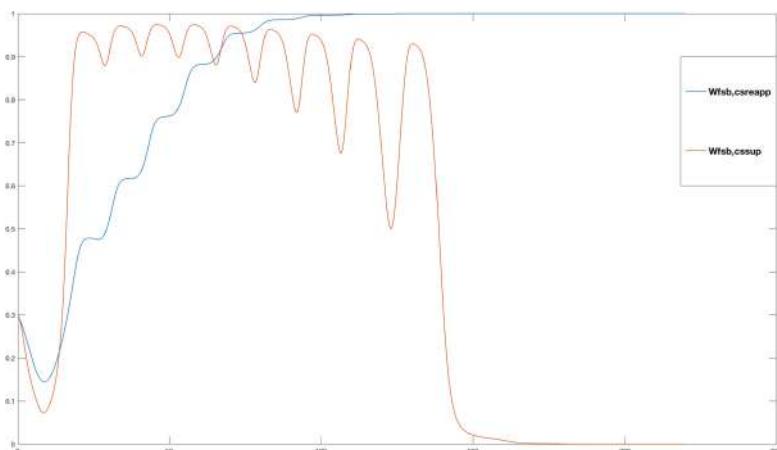


Fig. 3. First-order reified representation state over time (Age).

Figure 4 demonstrates the entire scenario in which choice of emotion regulation strategies changes as a person grows. Initially, suppression gets activated (in young age) and suppresses the body states by not letting the person to express his or her emotions while the negative belief and intensity of the stimulus stay high during this whole process. In the later stages of life, the person switches from suppression to reappraisal. It can be seen that this shift between strategies doesn't not take place at once. Tendency towards reappraisal is increasing over time and finally it becomes the major emotion regulation strategy. Effective states of this graph can be further studied in detail in Fig. 5.

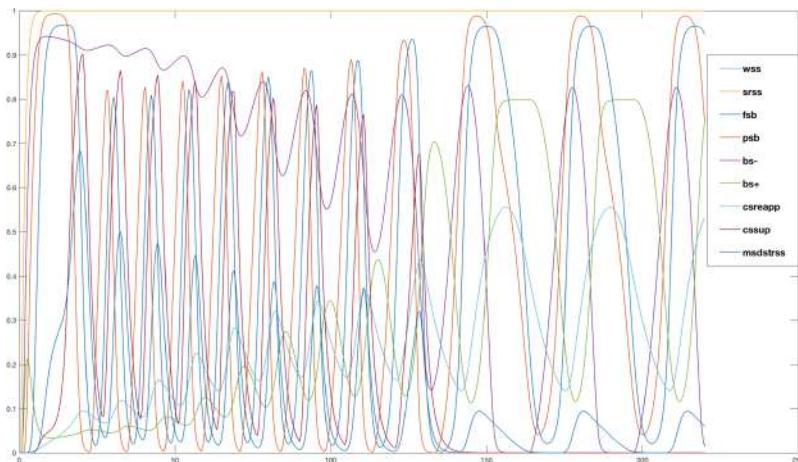


Fig. 4. Switching from suppression to reappraisal over time.

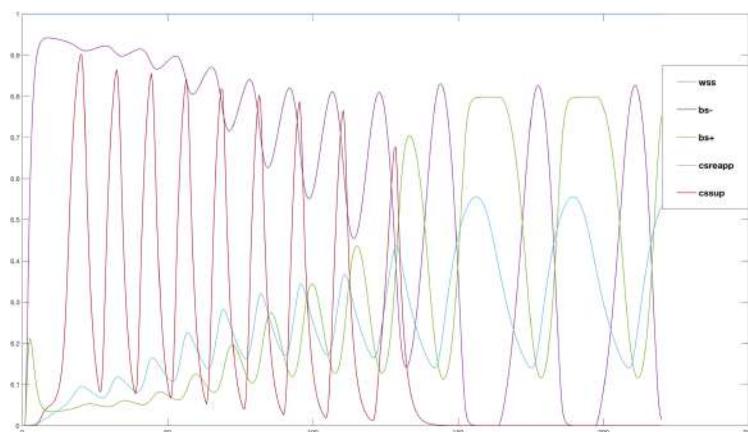


Fig. 5. Demonstration of the effective states over time.

Figure 5 shows only the effective states of Fig. 4. Initially, as suppression gets activated, it can be seen that bs^- still remains high and bs^+ remains low. This shows suppression of expression while the intensity of emotions still remains the same. At the same time, due to Hebbian learning, activation of reappraisal is continuously decreasing while suppression is increasing when, finally, activation of reappraisal is low enough to make reappraisal effectively get activated and change belief of the person. Therein, it can be seen the bs^- decreases while bs^+ increases. This demonstrates the phenomenon of metaplasticity exactly as described by literature from cognitive neuroscience.

5 Conclusion

Change in the choice of emotion regulation strategies, over time, is a proven fact in social and psychological literature so far. This paper brings this phenomenon into the field of network modeling within computer science and artificial intelligence. The layered approach used in this paper makes the phenomenon very dynamic and adaptive which is exactly as it takes place in real life. First, it establishes the fact that all such phenomena are prone to changes, second, the speed and intensity of this change in choice itself is changing over time.

Moreover, the layered network modeling approach used for this network model also takes the application of network-oriented modeling a step forward. This layered and abstract approach makes it possible to model every real-life phenomenon in a real but relatively easy way. The complexity of the network model lies in the dynamics of the different layers (with various adaptive parameters like speed factors for strategy choice adaptation over time, and persistency of learning), and the interaction or co-evolution of these layers. No other network models that address these second-order adaptive emotion regulation processes are known to the authors.

References

1. Aldao, A., Sheppes, G., Gross, J.J.: Emotion regulation flexibility. *Cogn. Ther. Res.* **39**(3), 263–278 (2015)
2. Carstensen, L.L., Isaacowitz, D.M., Charles, S.T.: Taking time seriously: a theory of socioemotional selectivity. *Am. Psychol.* **54**(3), 165–181 (1999)
3. Nakagawa, T., Gondo, Y., Ishioka, Y., Masui, Y.: Age, emotion regulation, and affect in adulthood: the mediating role of cognitive reappraisal. *Jpn. Psychol. Res.* **59**(4), 301–308 (2017)
4. Allen, V.C., Windsor, T.D.: Age differences in the use of emotion regulation strategies derived from the process model of emotion regulation: a systematic review. *Aging Ment. Health* **23**(1), 1–14 (2019)
5. Abraham, W.C., Bear, M.F.: Metaplasticity: the plasticity of synaptic plasticity. *Trends Neurosci.* **19**(4), 126–130 (1996)
6. Abraham, W.C.: Metaplasticity: tuning synapses and networks for plasticity. *Nat. Rev. Neurosci.* **9**(5), 387–399 (2008)

7. Zhenyu, G., Liu, R., Ullah, N.: A temporal-causal network model for age and gender difference in choice of emotion regulation strategies. In: Nguyen, N.T., et al. (eds.). LNAI 11683, pp. 106–117. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28377-3_9
8. Treur, J.: Multilevel network reification: representing higher order adaptivity in a network. In: Proceedings of ComplexNetworks 2018, vol. 1. Studies in Computational Intelligence, vol. 812, pp. 635–651. Springer (2018)
9. Treur, J.: Modeling higher-order adaptivity of a network by multilevel network reification. *Netw. Sci.* (2019, in press)
10. Treur, J.: Network-oriented modeling for adaptive networks: designing higher-order adaptive biological, mental and social network models. Springer (2020, to appear). <https://www.researchgate.net/publication/334576216>
11. Zimmermann, P., Alexandra, I.: Emotion regulation from early adolescence to emerging adulthood and middle adulthood: age differences, gender differences, and emotion-specific developmental variations. *Int. J. Behav. Dev.* **38**(2), 182–194 (2014)
12. John, O., Gross, J.J.: Healthy and unhealthy emotion regulation: personality processes. *Individ. J. Pers.* **72**(6), 1301–1334 (2004)
13. Charles, S.T.: Strength and vulnerability integration: a model of emotional well-being across adulthood. *Psychol. Bull.* **136**(6), 1068–1091 (2010)
14. Charles, S.T., Laura, L.C.: Emotion regulation and aging. In: Gross, J.J. (ed.) *Handbook of Emotion Regulation*, pp. 307–327. The Guilford Press, New York (2007)
15. Lawton, M.P., Kleban, M.H., Rajagopal, D., Jennifer, D.: Dimensions of affective experience in three age groups. *Psychol. Aging* **7**(2), 171–184 (1992)
16. Phillips, L.H., Henry, J.D., Hosie, J.A., Milne, A.B.: Age, anger regulation and well-being. *Aging Ment. Health* **10**(3), 250–256 (2006)
17. Carstensen, L.L., Pasupathi, M., Mayr, U., Nesselroade, J.R.: Emotional experience in everyday life across the adult life span. *J. Pers. Soc. Psychol.* **79**(4), 644–655 (2000)
18. Larcom, M.J., Derek, M.I.: Rapid emotion regulation after mood induction: age and individual differences. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* **64**(6), 733–741 (2009)
19. Silje, H., Kraft, P., Corby, E.: Emotion regulation: antecedents and well-being outcomes of cognitive reappraisal and expressive suppression in cross-cultural samples. *J. Happiness Stud.: Interdiscip. Forum Subj. Well-Being* **10**(3), 271–291 (2009)
20. Oliver, P.J., Gross, J.J.: Healthy and unhealthy emotion regulation: personality processes, individual differences, and life span development. *J. Pers.* **72**(6), 1301–1333 (2004)
21. Gross, J.J., et al.: Emotion and aging: experience, expression, and control. *Psychol. Aging* **12**(4), 590–599 (1997)
22. Dannii, Y.Y., Wong, C.K.M., Lok, D.P.P.: Emotion regulation mediates age differences in emotions. *Aging Ment. Health* **15**(3), 414–418 (2011)
23. Folkman, S., Lazarus, R.S., Pimley, S., Novacek, J.: Age differences in stress and coping processes. *Psychol. Aging* **2**(2), 171–184 (1987)
24. Scheibe, S., Blanchard-Fields, F.: Effects of regulating emotions on cognitive performance: what is costly for young adults is not so costly for older adults. *Psychol. Aging* **24**(1), 217–223 (2009)
25. Scheibe, S., Carstensen, L.L.: Emotional aging: recent findings and future trends. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* **65B**(2), 135–144 (2010)
26. Debora, C.: Cognitive reappraisal and expressive suppression strategies role in the emotion regulation: an overview on their modulatory effects and neural correlates. *Front. Syst. Neurosci.* **8**, 1–6 (2014)
27. Christina, R., Li, S., Smith, S.: Intraindividual variability in positive and negative affect over 45 days: do older adults fluctuate less than young adults? *Psychol. Aging* **24**(4), 863–878 (2009)

28. Mroczek, D.K., Kolarz, C.M.: The effect of age on positive and negative affect: a developmental perspective on happiness. *J. Pers. Soc. Psychol.* **75**(5), 1333–1349 (1998)
29. Charles, S.T., Almeida, D.: Daily reports of symptoms and negative affect: not all symptoms are the same. *Psychol. Health* **21**(1), 1–17 (2006)
30. Kira, S.B., Fingerman, K.L., Almeida, D.M.: Age differences in exposure and reactions to interpersonal tensions: a daily diary study. *Psychol. Health* **20**(2), 330–340 (2005)
31. Kira, S.B., Fingerman, K.L.: Age and gender differences in adults' descriptions of emotional reactions to interpersonal problems. *J. Gerontol. Ser. B Psychol. Sci. Soc. Sci.* **58**(4), 237–245 (2003)
32. Regier, D.A., et al.: One-Month prevalence of mental disorders in the United States: based on five epidemiologic catchment area sites. *Arch. Gen. Psychiatry* **45**(11), 977–986 (1988)
33. Charles, S.T., Reynolds, C.A., Gatz, C.A.: Age-related differences and change in positive and negative affect over 23 years. *J. Pers. Soc. Psychol.* **80**(1), 136–151 (2001)
34. Rothbart, M.K., Ahadi, S.A., Evans, D.E.: Temperament and personality: origins and outcomes. *J. Pers. Soc. Psychol.* **78**(1), 122–135 (2000)
35. Verhaeghen, P.: Aging and executive control: reports of a demise greatly exaggerated. *Curr. Dir. Psychol. Sci.* **20**(3), 174–180 (2011)
36. Labouvie-Vief, G., DeVoe, M., Bulka, D.: Speaking about feelings: conceptions of emotion across the life span. *Psychol. Aging* **4**(4), 425–437 (1989)
37. Labouvie-Vief, G., Blanchard-Fields, F.: Cognitive ageing and psychological growth. *Ageing Soc.* **2**(2), 183–209 (1982)
38. Heckhausen, J., Schulz, R.: A life-span theory of control. *Psychol. Rev.* **102**(2), 284–304 (1995)
39. Brandtstädtter, J., Renner, G.: Tenacious goal pursuit and flexible goal adjustment: explication and age-related analysis of assimilative and accommodative strategies of coping. *Psychol. Rev.* **5**(1), 58–67 (1990)



Comparison of Opinion Polarization on Single-Layer and Multiplex Networks

Sonoko Kimura¹(✉), Kimitaka Asatani², and Toshiharu Sugawara¹

¹ Waseda University, Tokyo 1698555, Japan

kmrnsnk@asagi.waseda.jp, sugawara@waseda.jp

² The University of Tokyo, Tokyo 1138656, Japan

asatani@gmail.com

Abstract. This paper investigates how opinions are polarized by simulating opinion formation with Q-learning in multiplex networks. People sometimes change their opinions to accommodate themselves to the surrounding people in communities, but opinions may still be polarized. To investigate the mechanism of opinion polarization, many studies including studies using agent-based simulations were conducted, but most of these simulations were performed by assuming that people belong to a single community. A number of studies assumed multiple communities, but they usually considered only simple opinion formation methods and more studies are needed. In this paper, we propose an opinion formation model on multiplex networks using Q-learning for agents to identify better individual opinions and analyze how opinions are polarized or agreed on various network structures. Our experiments indicate that opinions are more likely to lead to a consensus on multiplex networks than on single-layer networks. They also suggested that opinions are easily polarized when their cluster coefficient were high and the characteristic path length were longer.

Keywords: Opinion formation · Opinion polarization · Multiplex networks

1 Introduction

People have their own opinions, but they also hear other's opinions at the same time, so they express their opinions or may change them to accommodate themselves to the majority opinion in their communities. In addition, because people belong to multiple communities/groups such as groups of friends, groups of colleagues, and communities of local residents and because these communities may have different opinions, people try to express their opinions or modify them to eliminate conflicts with the majority opinions for the consensus in all communities they belong to. However, they are sometimes unable to find a compromise. Meanwhile, due to the recent spread of social networking services (SNSs) such

This work was partly supported by KAKENHI (17KT0044, 19H02384).

as Twitter, Facebook, and Instagram, people can easily use multiple SNSs and so belong to the associated communities. Therefore, people are likely to face different opinions and need to change their opinions, or at least, they must have consistent personal opinions within themselves.

In addition, the development of SNSs has enhanced the phenomenon of the echo chamber effect, i.e., users of SNSs are likely to only be exposed to articles that are aligned with their opinions/beliefs, reinforcing their opinions [11]. As a result, the opinions in a community on an SNS are polarized, and people think that only their opinions are correct and they are majority. This kind of polarization may occur in each community, but since people belong to multiple communities, they have many chances to encounter different and stronger opinions; this makes generating consistent personal opinions difficult. These discussions motivate us to study opinion formation and the mechanism of polarization in multiplex networks in which each layer of network corresponds to a community in an SNS.

Opinion polarization has been studied by many researchers [4,8]. For example, Garimella et al. [8] used the real Twitter data set to analyze the echo chamber effect to investigate the characteristics of users who express polarized opinions. Banisch and Olbrich [4] aimed to identify the mechanism and conditions of generating opinion polarization using an agent-based simulation with their game-theoretic model. This is a repeated game on a network where agents learn the best response to their surrounding agents using Q-learning. Then, they insisted that even if there were no opinion polarization in the initial state, opinions were polarized eventually, and each agent supported its current opinion more confidently. They also showed that the consensus, which was the state where all agents support the same opinion, was more likely to occur on dense networks.

Opinion formation on multiplex networks has also been studied recently because people usually belong to multiple communities [3,9,10,14]. For example, Amato et al. [3] proposed a model describing the competition between two options (opinions) in 2-layer multiplex networks. They extended the model proposed by Abrams and Strogatz (the AS model) [1] to multiplex networks. In their model, agents belong to two different networks and change their options by imitating/copying the option of one of its neighbors or the option of its counterpart in the other network. Their simulation results showed that the final state of the simulation was not only the consensus state but also the conflict state where the two options coexisted. However, they used the voter's model, and we think that it is too simple for studying opinion formation in a community because it did not consider the degree of confidence for the current opinions within agents.

Therefore, we attempt to analyze how the opinion polarization differs between single-layer and multiplex networks because the results of the above-mentioned studies [3,4] indicate that they had quite different phenomena. We generated multiplex networks whose individual networks are *random geometric graphs* (RGGs) or Erdős-Rényi networks (ER networks) [7], which were used in the previous studies [4], and introduced a game-theoretic opinion formation model with Q-learning. Note that our opinion formation model in multiplex networks

is based on that of Banisch and Olbrich [4], but we further extended it to model the interactions between networks in different layers, referring to the model in Amato et al. [3]. Moreover, we conducted an experimental simulation of opinion formation in the *connecting nearest neighbor* (CNN) [12], which has the scale-free property, and compared how opinions are polarized when the networks are RGGs, ER networks, and CNN networks. We also analyzed the features of opinion formation on multiplex networks by comparing with those on single-layer networks; these results are helpful to clarify the mechanism of the polarization.

2 Opinion Formation Model on Multiplex Networks

2.1 Networks and Agents

We propose a model for describing intra-network and inter-network opinion formation on multiplex networks, each of which represents a community or connections among people that may be in a cyberspace like Twitter and Facebook or in the real world, such as groups of friends and colleagues. An example of multiplex networks is illustrated in Fig. 1.

Multiplex networks $\mathcal{M} = \{N_W(1), \dots, N_W(L)\}$ consist of L networks, and the l -th layer network is denoted by $N_W(l)$, where L is a positive integer and $1 \leq l \leq L$. Network $N_W(l)$ is represented as the graph $N_W(l) = (V^l, E^l)$, where $V^l = \{v_1^l, v_2^l, \dots, v_n^l\}$ and $E^l \ni (v_i^l, v_j^l)$ are the sets of nodes and edges, respectively. Network $N_W(l)$ is generated independently and has a different topological structure. However, we assume that $V = V^1 = \dots = V^L$, meaning that any node $v \in V$ appears in all networks. We use the same index for each user; i.e., $V \ni v_i = v_i^1 = \dots = v_i^L$, which corresponds to a user, and we call it an *agent*. We often use the element $v_i \in V$ as a representative of v_i^1, \dots, v_i^L . From the definition, $E^l \neq E^{l'}$ if $l \neq l'$, meaning that agents belong to all communities but have different neighbors in each community.

Suppose that agent v_i^l supports a certain opinion in $N_W(l)$. The opinion supported by v_i^l is represented as O_i^l ($= -1$ or 1), where 1 and -1 correspond to incompatible counter-opinions. Agent v_i^l in $N_W(l)$ has the Q-values, $Q_i^l(1)$ and $Q_i^l(-1)$, that represent the degree of support for each opinion in the l -th network. Then, v_i^l 's opinion in $N_W(l)$ can be defined by comparing the Q-values, but we introduced a small fluctuation probability $0 < \varepsilon \ll 1$, which corresponds to an ε -greedy strategy in reinforcement learning to explore better opinions; i.e.,

$$O_i^l = \arg \max_{o \in \{-1, 1\}} Q_i^l(o) \quad (1)$$

with probability $1 - \varepsilon$, and O_i^l is randomly selected from $\{-1, 1\}$ with probability ε . Note that we often denote the counter-opinion of O_i^l as $-O_i^l$. It is possible

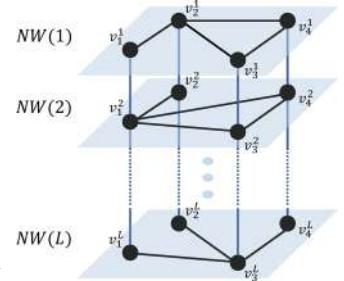


Fig. 1. Structure of multiplex networks.

for agents to support different opinions in each network. In this case, $O_i^l \neq O_i^{l'}$, and their associated Q-values have different values in each network. However, initially, v_i^1, \dots, v_i^L have the same Q-values such that $-0.5 \leq Q_i^l(1), Q_i^l(-1) \leq 0.5$. This can be expressed by

$$Q_i^1(o) = Q_i^2(o) = \dots = Q_i^L(o),$$

for $\forall v_i \in V$, where $o = -1$ or 1 . However, $O_i^1 = O_i^2 = \dots = O_i^l = \dots = O_i^L$ may not be satisfied due to the ε -greedy strategy. Agents form their opinion through two kinds of interactions: *intra-network opinion formation*, which corresponds to opinion formation through interactions with their neighboring agents in the same community, and *inter-network opinion formation*, which corresponds to opinion formation or opinion decisions reached through the thought process within each agent.

2.2 Intra-network Opinion Formation

Agents form their opinions through interactions with their surrounding agents to accommodate themselves to the community's opinion. For this purpose, each agent exchanges its opinions with its neighboring agents in the same network and updates the Q-values $Q_i^l(O_i^l)$ as follows [4].

We introduce a discrete time period whose unit is called a *time step*. First, agent v_i^l is randomly selected from each network for every time step (so a total of L agents are selected), and v_i^l chooses its neighbor agent v_j^l in $N_W(l)$ randomly (therefore, $(v_i^l, v_j^l) \in E^l$). Then, v_i^l expresses its opinion O_i^l to v_j^l , and then, v_j^l responds by expressing r_{ij}^l , which indicates whether it agrees or disagrees with v_i^l 's opinion. The value of r_{ij}^l can be defined as

$$r_{ij}^l = O_i^l \cdot O_j^l.$$

Then, v_i^l updates its Q-values by

$$Q_i^l(O_i^l) \leftarrow (1 - \alpha)Q_i^l(O_i^l) + \alpha r_{ij}^l, \text{ and} \quad (2)$$

$$Q_i^l(-O_i^l) \leftarrow Q_i^l(-O_i^l), \quad (3)$$

where parameter α is the learning rate for Q-values.

2.3 Inter-network Opinion Formation

It is possible that communities in the real-world have different opinions. However, if the user supports different opinions among the networks, her/his inconsistency and conflict will be observed by other acquaintances, and so the user may be criticized due to the inconsistency in her/his opinion by neighboring agents in a certain network, resulting in unstable feelings. To decrease the number of such conflicts, the users in a community sometimes change their opinion by copying a more confident opinion or assertion formed in other networks. We try to model

such an arrangement behavior. For example, agent v_i^l in network $N_W(l)$ changes its opinion O_i^l to opinion O_i^m supported by the same agent v_i^m as a member of $N_W(m)$ ($m \neq l$) if O_i^m is more strongly supported by v_i^m in $N_W(m)$. Thus, inter-network opinion formation corresponds to the process of decreasing the number of conflicts in opinions expressed in all communities.

This formation process is modeled as follows. Agent $v_i^l \in V^l$ is randomly selected every time step for $1 \leq \forall l \leq L$ and chooses another network $N_W(m)$ ($l \neq m$) randomly. Then, v_i^l changes its opinion O_i^l in $N_W(l)$ to its counterpart's opinion O_i^m in $N_W(m)$ with the following probability $P_{O_i^l \leftarrow O_i^m}$:

$$P_{O_i^l \leftarrow O_i^m} = \beta \cdot \left(\gamma \frac{|\Delta Q_i^m| - |\Delta Q_i^l| + 2}{4} + (1 - \gamma) \frac{\max Q_i^m - \max Q_i^l + 2}{4} \right), \quad (4)$$

where

$$\Delta Q_i^l = Q_i^l(1) - Q_i^l(-1), \quad \text{and} \quad \max Q_i^l = \max_O Q_i^l(O).$$

Note that when v_i^l changes its opinion, $Q_i^l(o)$ is set to $Q_i^m(o)$ for $o = -1, 1$. Parameter γ ($0 \leq \gamma \leq 1$) represents the degree of balance between the first and second terms in Formula (4), and parameter β ($0 \leq \beta \leq 1$) is the weight to determine the frequency for inter-network opinion formation.

Because the Q-value represents how much the agent is convinced by the current opinion, the difference between the Q-values expresses the degree of support for the current opinion that is opposed to another opinion. The first term of Formula (4) represents the difference between the differences of Q-values $|\Delta Q_i^l|$ and $|\Delta Q_i^m|$ in $N_W(l)$ and $N_W(m)$, and the second term represents the difference between the maximum Q-values. Therefore, both terms represent how strongly the agent is convinced by the opinion O_i^m compared with opinion O_i^l .

2.4 Feature Values of Opinion Formation

We measured six feature values, the *majority rate*, two kinds of *support strengths*, *dispersion*, *agreement rate*, and *consistency rate*, to understand the features of the converged opinion formation for various networks. The majority rate is defined as the rate of the agents that support the majority opinion in each network. The majority and minority support strengths are defined as follows. First, let us denote the set of agents that support $o \in \{-1, 1\}$ by S_o and the majority and minority opinions in $N_W(l)$ by o^a and o^b , respectively. Then, the majority and minority support strengths are calculated by using the average of the absolute values of the differences in Q-values $|\Delta Q_i^l|$ for all agents in $N_W(l)$, i.e.,

$$\frac{\sum_{v_i \in S_{o^a}} |\Delta Q_i^l|}{|S_{o^a}|} \quad (\text{Majority support strength})$$

$$\frac{\sum_{v_i \in S_{o^b}} |\Delta Q_i^l|}{|S_{o^b}|} \quad (\text{Minority support strength})$$

These values represent how strongly the majority and minority opinions are supported by the members in S_1 and S_{-1} . The dispersion is defined in DiMaggio et al. [6]’s study to measure the opinion variance. This is defined as the variance σ_l^2 of the distribution of ΔQ_i^l in the network $N_W(l)$, i.e.,

$$\sigma_l^2 = \frac{1}{N} \sum_{i=1}^N (\Delta Q_i^l - \overline{\Delta Q^l})^2, \quad (5)$$

where $\overline{\Delta Q^l}$ is the mean value of ΔQ_i^l for $\forall v_i^l \in V^l$.

The agreement rate is the rate of the agreement with the neighboring agents. Suppose that $N_W(l) = (V^l, E^l)$. This is calculated as follows.

$$\frac{1}{|E^l|} \sum_{(v_i^l, v_j^l) \in E^l} \frac{(2 - |O_i^l - O_j^l|)}{2}.$$

Note that $|O_i^l - O_j^l| = 0$ if v_i^l and v_j^l have the same opinion; otherwise, $|O_i^l - O_j^l| = 2$. The feature values explained above are calculated for each network. Therefore, we used their mean values for multiplex networks. Finally, the consistency rate is introduced to measure the agents’ internal consistency. This is defined as the rate of the agents supporting the same opinion in all networks to which the agent belongs; $|C|/|V|$, where $C = \{v_i \in V \mid O_i^1 = \dots = O_i^L\}$. Therefore, it can be defined only for multiplex networks.

3 Experiments and Discussion

3.1 Experimental Setting

We experimentally investigate the difference of the characteristics between opinion formation on a single-layer network (single community) and on multiplex networks (multiple communities) of the RGGs and ER networks and how different network structures of communities affect the distribution of opinion agreement/disagreement in the first experiment (Exp. 1). Note that the RGG is generated by placing nodes at random positions on a unit plane, $[0, 1] \times [0, 1]$, and generating a link between two nodes if their distance is less than radius r . We also investigate how opinions are polarized and whether the opinions that agents have in different networks are consistent. We then introduced CNN networks that have high-clustering coefficients, small-world, and scale-free properties [2], which are often observed in human society, and investigated the differences in opinion formation on the CNN networks with those on RGGs and ER networks in the second experiment (Exp. 2).

Table 1 lists the parameters used in these experiments. The parameters used to generate networks in our experiments are listed in Table 2, and their network characteristics are listed in Table 3. We set the values in Table 2 so that their degrees were almost identical. We also conducted these experiments using the Watts-Strogatz (WS) model [13] and Barabási-Albert (BA) model [5], but we

Table 1. Parameter values in experiments

Parameter	Description	Value
N	Number of agents (nodes)	1000
L	Number of networks in multiplex networks	2
α	Learning rate	0.05
ε	Exploration rate	0.1
β	Weight of inter-network opinion formation	1.0
γ	Balancing factor for opinion formation	0.5

Table 2. Parameters to generate networks.

Parameter	Description	Value
r	Neighborhood radius when generating links (RGG)	0.06
p	Probability of generating a link (ER model)	0.01
u	Probability of changing a potential edge to an edge (CNN model)	0.8

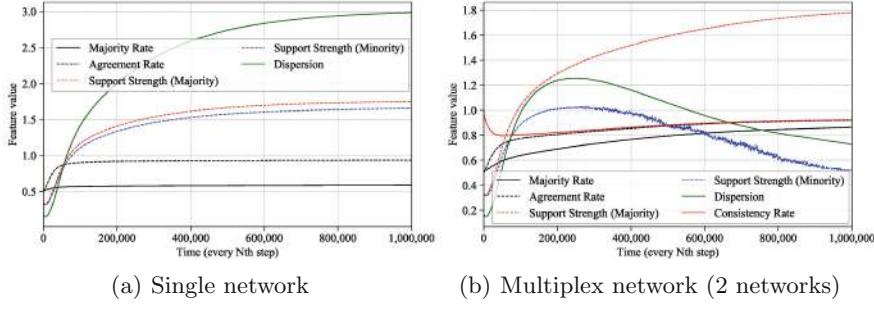
omit their results because the main characteristics of WS networks and BA networks are similar to those when using the RGGs and ER networks, respectively. In each experimental run using the multiple networks, they were generated using the same network generation model but were generated independently, so their structures were also different. The figures shown below are the average values of the 100 independent experimental trials using different random seeds.

3.2 Opinion Formation in the RGG and the ER Model

In Exp. 1, we investigated the difference of the features of opinion formations between the models on single-layer and multiplex networks of the RGG and the ER model. Note that the experiment on a single-layer network was examined by Banisch and Olbrich [4], but we reproduced their results using larger networks whose nodes are 1000 for comparison. All feature values of the opinion formation on the RGGs and the ER networks are shown in Figs. 2 and 3, respectively.

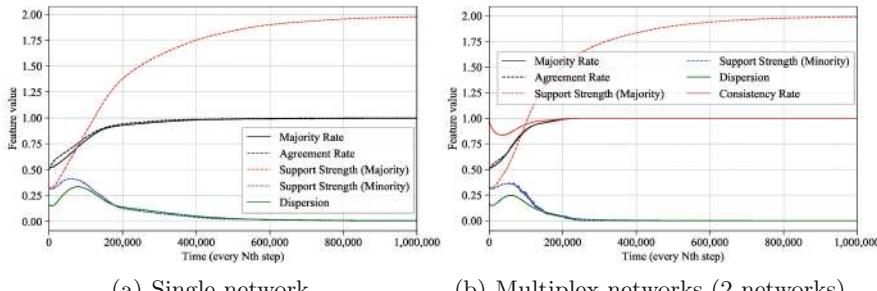
Table 3. Characteristics of networks.

Description	RGG	ER model	CNN model
Average degree (number of neighbors)	10.747	10.002	9.935
Average clustering coefficient	0.605	0.010	0.419
Average characteristic path length	11.787	3.262	4.424
Average power-law exponent			-1.330



(a) Single network

(b) Multiplex network (2 networks)

Fig. 2. Characteristics of the opinion formation in the RGG.

(a) Single network

(b) Multiplex networks (2 networks)

Fig. 3. Characteristics of the opinion formation in the ER model.

Figure 2(a) shows the feature values of opinion formation on the single-layer RGGs. The majority rate was around 0.6, the ratio of the number of agents supporting the majority opinion to the number of agents supporting the minority opinion was around 6:4, and this proportion seemed stable from the early steps of the experiment. However, because the agreement rate increased from 0.5 to around 0.93, agents with different opinions coexisted in half in the initial state, but nearby agents in the network gradually came to support the same opinion. Then, agents gradually supported their opinions more strongly through interaction with agents who agree with them. This is represented as the increase of both support strengths to over 1.65; with this increase, the dispersion also increased. During this process, the opinions were polarized on the networks.

The features of opinion formation on multiplex networks of RGGs, which are plotted in Fig. 2(b), are quite different. The decrease of the minority support strength indicates that the strength of supporting the minority opinion became weak, and the agents gradually changed their opinions to the majority opinion. This result means there was almost consensus; actually, the majority rate increased to around 0.86 and was still slightly increasing at the end of this experiment. In the same way, the dispersion increased at first but then decreased and became near 0.73. The first increase can be explained as follows. If $|\Delta Q_i^l - \bar{\Delta Q}^l|$

is large, v_i^l supports the current opinion O_i^l confidently; otherwise, it wonders which opinion it should support. Therefore, at first (until 300,000 time steps), the agent exchanged her/his opinions with both agents supporting the majority and minority opinions; in this situation, both support strengths were more than 1.0, which was relatively large, and the dispersion was around 1.25, so agents' opinions seemed diverse. After that, it decreased because the $|\Delta Q_i^l|$ of agents became similar values to each other because the number of agents supporting the majority opinion increases, and then agents exchanging their opinions were more likely to exchange opinions with agents supporting the majority opinions. In the same way, the consistency rate slightly decreased at first due to the intra-network opinion formation, but it gradually recovered due to the consensus in inter-network opinion formation.

However, the agreement rate increased slowly and seemed almost identical to that on the single-layer networks. Therefore, a small number of agents supported the minority opinion confidently because the minority support strength was still high (0.52) although it was gradually decreasing. Thus, the opinions were less polarized than in the single-layer networks, but small-scale polarization still existed. The polarization seemed to be resolved gradually but it might take longer time.

Figure 3(a) plots the feature values of opinion formation on the single-layer ER networks over time. We can see that all agents in the network reached the consensus because the majority rate and the agreement rate became 1.0. Since all agents support the majority opinion, the majority and minority support strengths approached 1.98 and 0.0. The dispersion also approached to almost 0.0 because all agents supported the same opinion with high strength, which was around 2.0.

The feature values of opinion formation on the multiplex ER networks is plotted in Fig. 3(b). This figure indicates that the transitions of the feature values were quite similar to those in the single-layer ER network, although the convergence speed was higher on multiplex networks than on single-layer networks. In addition, all networks reached consensus on the same opinion because the consistency rate converged to 1.

3.3 Polarization in CNN Networks

In Exp. 2, we investigated the opinion formation on the CNN networks and compared the difference of opinion formation between single-layer and multiplex networks. The feature values of the opinion formation on the single-layer CNN networks were plotted in Fig. 4.

Figure 4(a) indicates the slightly different tendency of opinion formation on single-layer networks compared to opinion formation from those on the RGGs and ER single-layer networks. Because the agreement rate became very high (around 0.98, which is between the agreement rate in ER networks and that in RGGs), there are many agents whose neighboring agents also agreed with them. Therefore, agents are more likely to exchange opinions with agents with the same opinion. Thus, the majority support strength increased to around 1.95.

The minority support strength also increased to over 1.64. The majority rate rose to around 0.86 and became stable. This means that both opinions are probably strongly supported in the local sub-communities in the CNN networks. Therefore, the opinions will remain polarized.

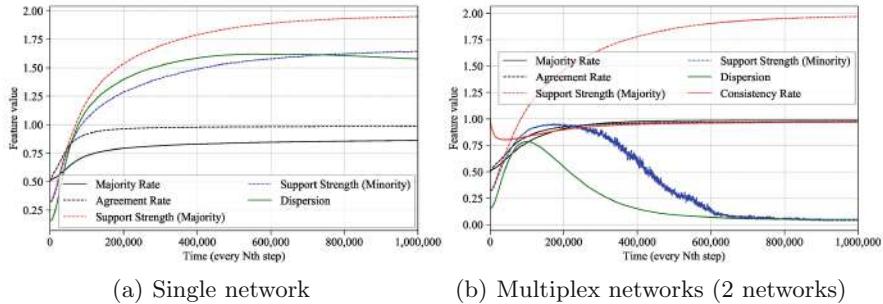


Fig. 4. Characteristics of the opinion formation in the CNN model.

Figure 4(b), which plots the feature values of the opinion formation in the multiplex CNN networks, also indicates that the features of opinion formation are different from those in the single-layer CNN networks. The majority rate and the agreement rate approached 0.98; hence, the networks almost reached consensus. Because the consistency rate approached to 0.97 but is not equal to 1.0, all agents in the networks almost reached consensus on the same opinion, but a small number of agents supported the minority opinion.

3.4 Discussion

Comparing the RGG and the ER networks, opinions in the RGG were likely to be polarized. The RGGs have the high cluster coefficients, so agents in a same subgroup who are densely connected to come to support the same opinions independently; thus, the opinions tended to be polarized. Our experimental results also indicate that opinion formations in multiplex networks were likely to reach consensus, compared with those in single-layer networks for all network types. Although agents in the RGGs were likely to reach consensuses individually in local groups, the inter-network opinion formation on multiplex networks could provide chances to exchange their opinions with agents in different groups. In this way, opinions were less polarized on multiplex networks. For the ER networks, we can see that the consensus was reached in both networks, but the convergence was faster in multiplex networks than in single-layer networks.

The majority rate was higher in the CNN single-layer networks than in the RGG single-layer networks. Opinions in the CNN multiplex networks almost reached consensus but opinions in the RGG multiplex networks still remained polarized. Actually, the dispersion and the minority support strength of RGG

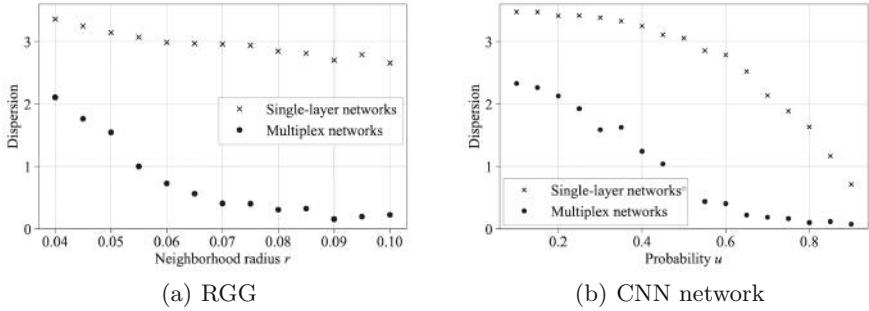


Fig. 5. Dispersion in RGGs and CNN networks.

multiplex networks were higher than those of the CNN multiplex networks. We think that these were also due to the difference in the values of the average clustering coefficient and the average characteristic path length (see Table 3).

We investigated the values of dispersion at the 1,000,000 time step by changing the parameters of r (in the RGG). The results are plotted in Fig. 5(a). Note that according to the increase of r from 0.6 to 1.0, the characteristic path length decreased from 11.79 to 6.36, but the cluster coefficients changed within a narrow range [0.60, 0.62]. This figure indicates that dispersion gets smaller if r is larger (so the characteristic path length is also shorter). We also investigated similar experiments using the CNN networks with various values of u . We varied the values of u from 0.1 to 0.9; their clustering coefficients ranged from 0.07 to 0.47, and their characteristic path lengths ranged from 9.84 to 3.27 (the degrees of network also increased). This result also indicates a tendency similar to that shown in Fig. 5(b). Therefore, these results, including the results of the BA networks and WS networks, suggested that opinions were easily polarized when the clustering coefficient of the network is high. Then, the inter-network opinion formation could reduce the polarization, but it took a longer time to reach the consensus if the characteristic path length was longer.

The proposed model describes the situations in which a person expresses different opinions in each network in the constraint of consistency of the opinions. We simulate the situations that people discuss something in the social networks and their workplaces/schools. The consistency constraint may come from a fear that discovery of inconsistency by other people. Our simulation result suggests that an opinion is likely to converge when agents consider the consistency of the expressed opinions in each network. The result indicates that the effect of some hidden network, which is hard to be observed due to data availability, plays an important role for opinion convergence. The consistency of opinion mediates the opinion dynamics between each layer of networks. Our study reveals the importance of agent's consistency of opinion for opinion formation.

4 Conclusion

We investigated the opinion polarization on multiplex networks with a model that forms opinions within each network using Q-learning and formed opinions between networks to try to make their opinions consistent across networks. The experiments comparing single-layer and multiplex networks showed that the agents tended to reach consensus on multiplex networks. This is because the inter-network opinion formation can be regarded as the interaction of each agent with the agents outside its local group and can make agents express more confident opinions formed on one of other networks. From the comparison of each network structure, opinions were polarized in the RGG and were also polarized in the CNN model although not as much as in the RGG, while the model reached consensus in the ER model. The larger the average cluster coefficient and the longer the average characteristic path length, the more likely the opinions will be polarized. In future work, we would like to investigate the effect on polarization/consensus on the networks.

References

1. Abrams, D.M., Strogatz, S.H.: Modelling the dynamics of language death. *Nature* **424**(6951), 900 (2003). <https://doi.org/10.1038/424900a>
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002). <https://doi.org/10.1103/RevModPhys.74.47>
3. Amato, R., Kouvaris, N.E., Miguel, M.S., Díaz-Guilera, A.: Opinion competition dynamics on multiplex networks. *New J. Phys.* **19**(12), 123019 (2017). <https://doi.org/10.1088/1367-2630/aa936a>
4. Banisch, S., Olbrich, E.: Opinion polarization by learning from social feedback. *J. Math. Sociol.* **43**(2), 76–103 (2019). <https://doi.org/10.1080/0022250X.2018.1517761>
5. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. *Science* **286**(5439), 509–512 (1999)
6. DiMaggio, P., Evans, J., Bryson, B.: Have American's social attitudes become more polarized? *Am. J. Sociol.* **102**(3), 690–755 (1996). <https://doi.org/10.1086/230995>
7. Erdős, P., Rényi, A.: On random graphs I. *Publicationes Mathematicae (Debrecen)* **6**, 290–297 (1959)
8. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Political discourse on social media: echo chambers, gatekeepers, and the price of bipartisanship. In: Proceedings of the 2018 World Wide Web Conference, pp. 913–922, Switzerland (2018). <https://doi.org/10.1145/3178876.3186139>
9. Halu, A., Zhao, K., Baronchelli, A., Bianconi, G.: Connect and win: the role of social networks in political elections. *EPL (Europhys. Lett.)* **102**(1), 16002 (2013). <https://doi.org/10.1209/0295-5075/102/16002>
10. Nguyen, V.X., Xiao, G., Xu, X.J., Li, G., Wang, Z.: Opinion formation on multiplex scale-free networks. *EPL (Europhys. Lett.)* **121**(2), 26002 (2018). <https://doi.org/10.1209/0295-5075/121/26002>
11. Sunstein, C.R.: The law of group polarization. *J. Polit. Philos.* **10**(2), 175–195 (2002). <https://doi.org/10.1111/1467-9760.00148>

12. Vázquez, A.: Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E* **67**, 056104 (2003). <https://doi.org/10.1103/PhysRevE.67.056104>
13. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* **393**(6684), 440–442 (1998). <https://doi.org/10.1038/30918>
14. Xu, W.J., Zhong, L.X., Huang, P., Qiu, T., Shi, Y.D., Zhong, C.Y.: Evolutionary dynamics in opinion formation model with coupling of social communities. *Adv. Complex Syst.* **18**(01n02), 1550003 (2015). <https://doi.org/10.1142/S0219525915500034>



Learning of Weighted Multi-layer Networks via Dynamic Social Spaces, with Application to Financial Interbank Transactions

Chris U. Carmona^{1,3}(✉) and Serafin Martinez-Jaramillo^{2,3}

¹ Department of Statistics, University of Oxford, Oxford, UK
carmona@stats.ox.ac.uk

² Center for Latin American Monetary Studies, Mexico City, Mexico
³ Banco de Mexico, Mexico City, Mexico

Abstract. We propose a general network model suited for longitudinal data of multi-layer networks with directed and weighted edges. Our formulation built upon the latent social space representation of networks. It consists of a hierarchical formulation: deep levels of the model represent latent coordinates of agents in the social space, evolving in continuous time via Gaussian Processes; meanwhile, top levels jointly manage incidence and strength of interactions by considering a Zero-Inflated Gaussian response. Learning of the model is performed through Bayesian Inference. We develop an efficient MCMC algorithm targeting the posterior distribution of model parameters and missing data (available in GitHub). The motivation for our model lies in the context of Financial Networks, specifically the analysis of transactions between commercial banks. We evaluate the model in synthetic data, as well as our main case study: the network of inter-bank transactions in the Mexican financial system. Accurate predictions are obtained in both cases estimating out-of-sample link incidence and link strength.

1 Introduction

In a number of natural, social and economic systems the dynamic interaction between agents across time can be recorded as longitudinal network-valued data, with directed and weighted edges. Examples include: migration of people and/or animals between regions, value of trading between countries, financial transactions between banks, etc. These interactions often occur in multiple layers of connectivity, thus generating *multi-layer* networks, which should be jointly modeled for an adequate understanding of the system under study.

The analysis and understanding of Networks have advanced rapidly during the last few years. Nevertheless, there is still a recognized underdevelopment of statistical analysis for Dynamic Networks (Crane 2018). The lack of available models is even more pronounced for dynamic complex systems with weighted,

directed and multilayered interactions, as most of the seminal advancements in dynamic network models are centered in binary undirected interactions.

In particular, there exist an important gap in current models for banking interactions that *simultaneously* incorporate features such as time-dependency across layers, external predictors, and other desirable characteristics that potentially improve the predictive performance of models. There is little to precedent of statistical inference of the underlying dynamic structure of weighted multi-layer networks in the context of financial system. Arguably, this may be consequence of two factors: (1) the novelty of methods that can handle the combined complexity of such systems, and (2) the unavailability of open-access datasets due to the sensitivity of such data, which in turn deters potential academics interested on statistical models for financial networks.

A successful implementation of this techniques will uncover critical hidden structures and dependencies in the financial system, significantly improving the performance of models for prediction, stress testing, and ultimately, systemic risk assessment. An adequate understanding of the underlying characteristics of the complex financial networks provides important insights that translate into concrete policy insights and policy applications to financial stability and macro-prudential regulation (Battiston and Martinez-Jaramillo 2018).

1.1 Related Literature: Latent Space Models for Networks

One of the current frontiers in statistical models for networks is being developed under the *Latent Social Space* representation. This framework, originally studied in the seminal work of Hoff et al. (2002), assumes that each agent is positioned within a latent social space, and the probability of interaction with other agent depends on the relative distance between the two. The model has been expanded in recent years to accommodate for more complex features (see Kim et al. (2018a) for a review).

In Ward et al. (2013), a comprehensive model for longitudinal data of the world-trade network is introduced. Their model presents valuable features, such as incorporating network dependencies jointly for both link incidence and strength, supporting directed relations by considering *sender and receiver* latent spaces; as well as including external information as predictors. Nevertheless, this model is not suited to jointly model multi-layer networks, and the dynamics are considered in discrete time, which complicates inference when there is uneven sampling in time.

Durante and Dunson (2014b, 2014a) introduced continuous-time dynamics by considering Gaussian Processes for the evolution of latent coordinates of agents across time, later expanded in to incorporate locally adaptive dynamics Durante and Dunson (2016). Durante et al. (2017) introduced a model for multi-layer networks, the formulation considers one shared latent space to capture the global structure between agents, and K layer-specific latent spaces, which characterises the idiosyncratic structure of each layer. These models, however, are suitable only for undirected (symmetric) unweighted (binary) edges.

Sewell and Chen (2015, 2016, 2017) proposed an alternative formulation for dynamic networks with directed and weighted edges. Instead of considering two latent spaces (sender/receiver) as in Ward et al. (2013), they consider one latent space, together with a set of node-specific parameters $r_{1:n}$ to reflect the *social reach* of agents, and two global parameters, β_{IN} and β_{OUT} , to express the importance of popularity and sociability, respectively. multi-layer networks and continuous time are not considered in their formulation.

Recently, Linardi et al. (2017) developed a model for inter-bank data based on Sewell and Chen (2015) model. Their aim is to characterize only the presence or absence of links between banks in discrete time, in a single layer.

1.2 Our Contribution

We fill a gap in the current literature of statistical network models by combining features from preceding research into a comprehensive model. Compared with previous network models, we expanded the inferential tools to allow the inference and prediction of a more general class of dynamic networks, accommodating for links with direction, weight, or both. Moreover, the simultaneous incorporation of additive effects and external covariates had not been discussed before, such factors reveal important insight about the underlying structure of the network and its evolution.

The correct implementation of this features together is not trivial. We developed the *DynMultiNet* package¹ based on R and C++. The application of the model and software developed are not restricted to financial networks, there is plethora of fields in which our model could be beneficial for the understanding and prediction of such complex networks.

In the context of financial stability, our work is pioneer. No previous work has performed statistical inference on the joint network of transactions between banks. Our probabilistic framework opens the door for future work in anomaly detection, stress testing and risk management which considers agents interactions adequately.

This work is motivated by the study of transactions between commercial banks within the Mexican Banking System. In Sect. 5 we apply the proposed methodology to real transactions observed in the Mexican Banking System between 2010 and 2018. Our model achieves accurate in- out-of-sample estimation of probabilities of connection and transaction value.

The model is defined and discussed in Sect. 2, followed by a description of the MCMC method developed for estimation in Sect. 3.

2 The Model

Consider a dynamic system of V interactive agents across time, with activity recorded as a multi-layer network with K levels. Let $Y^{(k)}(t) = \{y_{ij}^{(k)}(t)\} \in \mathbb{R}^{V \times V}$

¹ Available at <https://github.com/christianu7/DynMultiNet>.

be the weighted adjacency matrix for the network in layer k at time t , with $y_{ij}^{(k)}(t)$ measuring the interaction from agent i to agent j , for $i, j = 1, \dots, V$, $t = t_1, \dots, t_T$ and $k = 1, \dots, K$.

The network model consists of a hierarchical model. The observational model is given by a mixture between a Gaussian distribution and a probability mass at zero. The first component accounts for the strength of positive interactions between agents, while the second comprise the pairs with no activity. This is

$$y_{ij}^{(k)}(t) \sim \lambda_{ij}^{(k)}(t) * \mathcal{N}(\mu_{ij}^{(k)}(t), \sigma_{(k)}^2) + [1 - \lambda_{ij}^{(k)}(t)] * \delta_{y_{ij}^{(k)}(t)}(0). \quad (1)$$

independently for $i, j = 1, \dots, V$, $t = t_1, \dots, t_T$ and $k = 1, \dots, K$. Here, $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2 and $\delta_y(0)$ the Dirac measure concentrated at zero.

$\lambda_{ij}^{(k)}(t)$ is interpreted as the *probability of incidence* for an interaction from agent i to agent j in layer k at time t , and $\mu_{ij}^{(k)}(t)$ is the *expected strength* of such the interaction -if it existed-. $\sigma_{(k)}^2$ is the variance of all interactions within the corresponding layer k . We incorporate the latent space representation of networks through $\mu_{ij}^{(k)}(t)$ and $\lambda_{ij}^{(k)}(t)$.

We consider *multiple* social spaces to deal with the directed and multi-layer nature of our networks. To accommodate for directed interactions, we follow Ward et al. (2013) and duplicate the social spaces into the *sender* and *receiver* spaces. Additionally, following Durante et al. (2017), we will use $K + 1$ spaces for a K -layered network: one global space capturing the systemic interaction between agents, and K layer-specific spaces that speak for idiosyncratic behavior for activity of type k .

$$\begin{aligned} \mu_{ij}^{(k)}(t) &= \theta^{(k)}(t) + & \gamma_{ij}^{(k)}(t) &= \eta^{(k)}(t) + \\ & u_i(t)' v_j(t) + u_i^{(k)}(t)' v_j^{(k)}(t) + & a_i(t)' b_j(t) + a_i^{(k)}(t)' b_j^{(k)}(t) + \\ & s_{\mu,i}(t) + p_{\mu,j}(t) + & s_{\lambda,i}(t) + p_{\lambda,j}(t) + \\ & \beta_\mu(t)' x_{ij}^{(k)}(t), & \beta_\lambda(t)' x_{ij}^{(k)}(t). \end{aligned} \quad (2)$$

Let us start by describing the expected strength. Here $u_i(t) \in \mathbb{R}^H$ is a *dynamic* vector of latent coordinates which represents the location of node i at time t within the *global sender space*; similarly, $v_j(t)$ is the position of agent j in the *global receiver space*. $u_i^{(k)}(t)$ and $v_j^{(k)}(t)$ denote the corresponding coordinates within the *layer-specific spaces*. The baseline processes $\theta^{(k)}(t) \in \mathbb{R}$ captures the average intensity of interactions between all agents in layer k .

Now, $s_{\mu,i}(t)$ and $p_{\mu,j}(t)$ are additive effects, induced by the sender agent i and receiver agent j , respectively. Additive effects represent agent i “sociability” (out-degree) and agent j “popularity” (in-degree). They capture significant heterogeneity in activity levels across nodes (Ward et al. 2013; Hoff 2018; Kim et al. 2018b).

External covariates are also a desirable feature that incorporate exogenous variation to our system (Ward et al. 2013; Durante and Dunson 2014a; Kim et al. 2018b). We introduce them using a vector of P edge-specific covariates $x_{ij}^{(k)}(t) \in \mathbb{R}^p$. We define $\beta_\mu(t) = (\beta_{\mu,1}(t), \dots, \beta_{\mu,P}(t)) \in \mathbb{R}^p$ and $\beta_\lambda(t) \in \mathbb{R}^p$ as the corresponding -dynamic- coefficients for the strength and incidence of interactions.

The dynamic of the network is captured by assuming changes in the parameters described before. We adopt smooth trajectories in our formulation introducing *Gaussian Processes* (GPs) with squared exponential correlation. $C(t, t') = \exp\left(\frac{t-t'}{\delta}\right)^2$, as the priors for the coordinates,

$$\begin{aligned}\theta^{(k)}(\cdot) &\sim GP(\bar{\theta}^{(k)}, C_\mu), \\ u_{i,h}(\cdot) &\sim GP(\bar{u}_{i,h}, C_\mu), \quad v_{i,h}(\cdot) \sim GP(\bar{v}_{i,h}, C_\mu), \\ u_{i,h}^{(k)}(\cdot) &\sim GP(\bar{u}_{i,h}^{(k)}, C_\mu), \quad v_{i,h}^{(k)}(\cdot) \sim GP(\bar{v}_{i,h}^{(k)}, C_\mu),\end{aligned}\tag{3}$$

independently for $k = 1, \dots, K$, $i = 1, \dots, V$ and $h = 1, \dots, H$, with $C_\mu(t, t') = \exp\left(\frac{t-t'}{\delta_\mu}\right)^2$. The parameter and δ_μ control the smoothness of changes across time for these latent coordinates. The current implementation of our model uses a single smoothness parameter for all latent coordinates to ease the MCMC implementation². We use a constant (non-zero) mean function for the GPs centering parameter.

More elaborate dynamics can be considered for the latent locations. One option is incorporating non-stationarity through the GPs covariance Rasmussen and Williams (2005). Another possibility are *Nested Gaussian Processes* (Durante and Dunson 2016) which induce Locally Adaptive trajectories for the latent factors.

The dimension of the social spaces H is fixed. We suggest choosing H by optimizing a metric for predictive performance, such as the WAIC Watanabe (2009) or the Pareto-smoothed approximation Vehtari et al. (2017) to leave-one-out cross-validation.

The probability of incidence $\lambda_{ij}^{(k)}(t)$ is treated in a similar way to the strength of interaction. However, we need to map the latent similarity measure among units into the probability space. We use a logistic link for this mapping, obtaining

$$\begin{aligned}\lambda_{ij}^{(k)}(t) &= \frac{1}{1 + \exp(-\gamma_{ij}^{(k)}(t))}, \\ \gamma_{ij}^{(k)}(t) &= \eta^{(k)}(t) + a_i(t)' b_j(t) + a_i^{(k)}(t)' b_j^{(k)}(t).\end{aligned}\tag{4}$$

We introduced a separate baseline processes $\eta^{(k)} \in \mathbb{R}^H$ and new set of latent coordinates $a_{i,h}, b_{i,h}, a_{i,h}^{(k)}, b_{i,h}^{(k)} \in \mathbb{R}^H$. These terms follow similar dynamics as we

² It is possible to optimize for the parameter δ for each agent using Variational Inference (Tran et al. 2015), which will be explored in future works.

defined previously for θ, u and v ,

$$\begin{aligned}\eta^{(k)}(\cdot) &\sim GP(\bar{\eta}^{(k)}, C_\mu), \\ a_{i,h}(\cdot) &\sim GP(\bar{a}_{i,h}, C_\mu), \quad b_{i,h}(\cdot) \sim GP(\bar{b}_{i,h}, C_\mu), \\ a_{i,h}^{(k)}(\cdot) &\sim GP(\bar{a}_{i,h}^{(k)}, C_\mu), \quad b_{i,h}^{(k)}(\cdot) \sim GP(\bar{b}_{i,h}^{(k)}, C_\mu),\end{aligned}\tag{5}$$

independently for $k = 1, \dots, K$, $i = 1, \dots, V$ and $h = 1, \dots, H$, with $C_\lambda(t, t') = \exp\left(\frac{t-t'}{\delta_\lambda}\right)^2$.

We use separate sets of latent coordinates for the incidence and for the strength of interactions. In case studies datasets, we observed that the empirical distribution of the log-strength $\log(y_{ij}^{(k)}(t) + 1)$ for existing links (i.e. $y \geq 0$) is concentrated far from zero, even for pairs with low activity. Using a zero-truncated gaussian distribution to account for both incidence and strength (as in Sewell and Chen (2016)) would notably reduce the performance of the model.

We preserve GPs for the dynamics of additive effects and coefficients associated with external covariates. We use different parameters for the GPs for each external covariates. In financial networks, agents may react fast to changes in stock markets levels and slow to changes in macroeconomic variables such as employment or GDP growth. We have

$$\begin{aligned}s_{\mu,i}(\cdot) &\sim GP(\bar{s}_{\mu,i}, C_\mu), \quad s_{\lambda,i}(\cdot) \sim GP(\bar{s}_{\lambda,i}, C_\lambda), \\ p_{\mu,i}(\cdot) &\sim GP(\bar{p}_{\mu,i}, C_\mu), \quad p_{\lambda,i}(\cdot) \sim GP(\bar{p}_{\lambda,i}, C_\lambda), \\ \beta_{\mu,l}(\cdot) &\sim GP(\bar{\beta}_{\mu,l}, C_{\beta_l}), \quad \beta_{\lambda,l}(\cdot) \sim GP(\bar{\beta}_{\lambda,l}, C_{\beta_l}),\end{aligned}\tag{6}$$

independently for each agent $i = 1, \dots, V$ and each covariate $l = 1, \dots, P$.

The formulation of the model allows to straightforwardly handle missing values, infer missing links and perform forward and backward predictions.

Sampling from missing link is an optional step. Inference of all the latent parameters does not rely on the imputed values, as they can be integrated out for the computation of conditional posteriors.

3 Estimation

We adopt a Bayesian approach to learn the parameters in the model. The posterior distribution is computed via a Gibbs sampler. We follow Durante and Dunson (2014b) and employ a Polya-Gamma data augmentation strategy Polson et al. (2013) to sample the dynamic terms related to the link incidence (i.e. $\eta, a, b, a^{(k)}, b^{(k)}, s_\lambda, p_\lambda, \beta_\lambda$). Under this approach, every dynamic latent in the model has a multidimensional Gaussian distribution as conditional posterior. The main complication arise in correctly setting the updating sequence to maintain conditional independence, and in arranging the associated design and response matrices. We use Metropolis-Hastings (MH) steps to sample the variance of the weights $\sigma^{(k)}$.

We provide an implementation of the sampler using the R and C++ programming languages³, as well as detailed derivation of the posterior in the author's website.

4 Simulation Study

We conduct a simulation study to evaluate the performance of our model in a synthetic dataset, constructed to mimic a possible generating process in a financial application⁴.

Consider a system with $V = 15$ agents, whose interactions are recorded in a Weighted multilayer network with $K = 2$ layers. We assume a total of $T = 30$ observational times, which are randomly (uniform) distributed in the interval $(0, 30)$.

Probabilities of positive interactions are constructed considering two components. First, *core-periphery* static probabilities that differentiate between highly-connected and dissociated nodes, as shown in heat maps on the left side of Fig. 1 (top layer 1, bottom layer 2). Second, dynamic probabilities that emulate a scenario of financial crisis happening at time $t = 10$, displayed on Fig. 1. Two of these trajectories affect the probability $\lambda_{ij}^{(k)}(t)$, one for each layer: a sharp increase in connectivity before the crisis, followed by a prolonged decrease (top-left); a systemic decrease of trust that start few periods before the crisis (top-right).

The expected weights of the links are created through an inverse logistic transformation of the static probabilities, then adding a fixed value for each layer, and finally including two additional trajectories. The trajectories here assume that strength of connection increases/decreases between central/periphery agents after the crisis (bottom-left/bottom-right).

Finally, both expected weights and probabilities were perturbed considering agent-specific trajectories during the period.

For inference, we choose a value $H = 3$ for the dimension of the latent spaces and a value $\delta = 20$ for the smoothing parameter. We ran 2500 Gibbs iterations, and discarded the first 500 as warm-up period. We verified good mixing of the parameters in the observational equations.

In Fig. 2 we compare the generative vs. estimated probabilities of connection for three relevant times: before, at, and after the crisis ($t = 0.3, 9.1, 24.6$ respectively). The structure of the network is learned adequately. Figure 3 shows an overall comparison of the estimated probabilities and weights across all agents, layers and times. The panels compare horizontally three types of estimation: in-sample, imputation, and projective forecasting. Top panels correspond to probabilities, bottom panels to weights. The model effectively captures the underlying structure of the network in all periods, but it performs much better when more information is available, as expected.

³ Available at <https://github.com/christianu7/DynMultiNet>.

⁴ Code to replicate this experiment is available at the author's website.

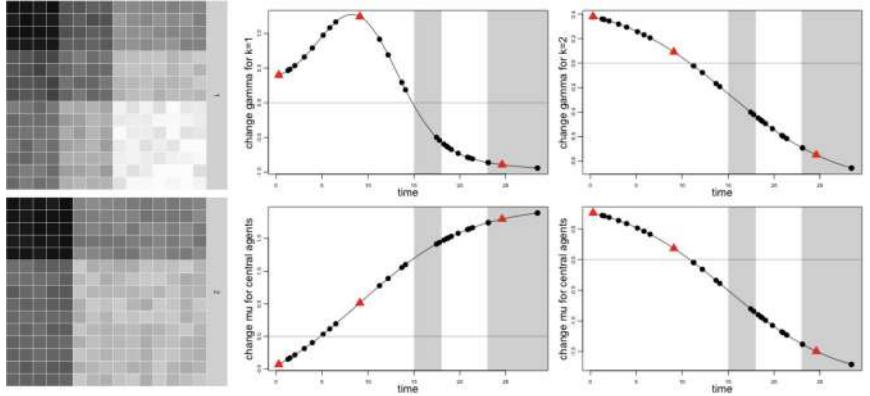


Fig. 1. Base probabilities and systemic trends in synthetic data. Left panel shows base probabilities for the $V = 15$ agents and two $K = 2$ layers. Right panel show the systemic trends that modify $\mu_{ij}^{(k)}(t)$ and $\mu_{ij}^{(k)}(t)$ through time for all agents. The points corresponding values at observational times. Shaded areas display periods with missing data.

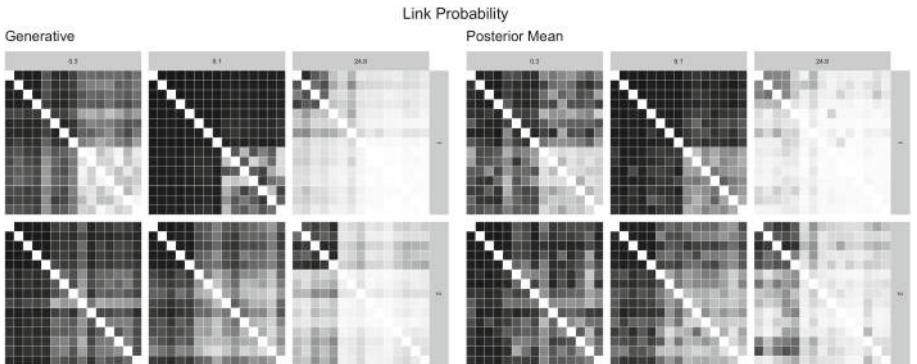


Fig. 2. Comparison of real vs. estimated probabilities. We show the probability of connection $\lambda_{ij}^{(k)}(t)$ for three times ($t = 0.3, 9.1, 24.6$) and the $K = 2$ layers. The left panel shows the original generative probabilities, while the right panel show the estimation given by the posterior mean.

Lastly, Fig. 4 displays how the model captures the dynamics of the network. For the pair of agents ($i = 1, j = 7$), we show the evolution of probability of connection and link strength. The smooth dynamic of the latent elements in the model effectively captures changes in the adjacency matrices.

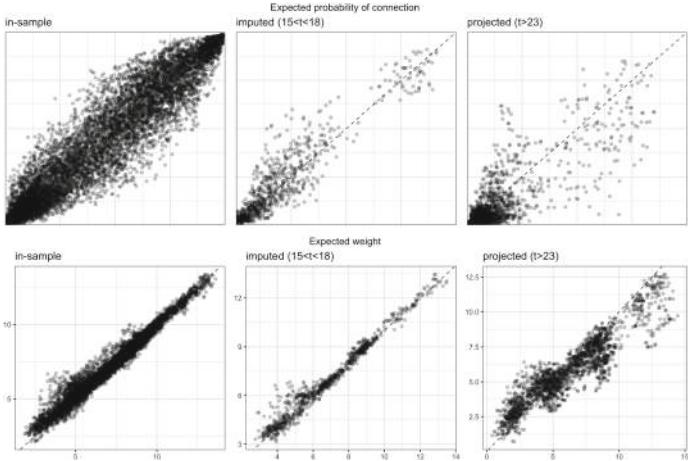


Fig. 3. Overall comparison of real vs. estimated probabilities and weights. We display the prediction accuracy for probability $\lambda_{ij}^{(k)}(t)$ (top) and mean strength $\mu_{ij}^{(k)}(t)$ (bottom). We compare the prediction under three conditions: Period with observations (left), period with missing data within sample $15 < t < 18$ (center), and forecasted values for the future $t > 23$ (right).

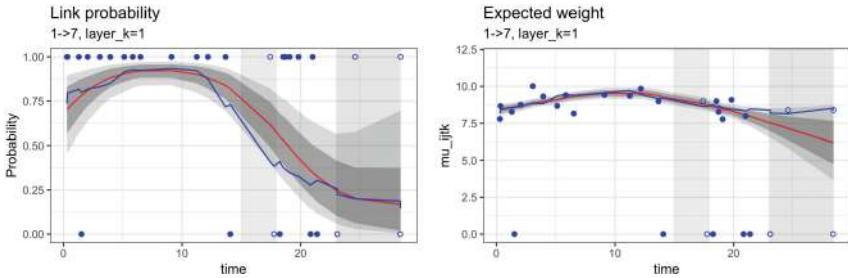


Fig. 4. Dynamics of probability of connection $\lambda_{ij}^{(k)}(t)$ (left) and expected weight $\mu_{ij}^{(k)}(t)$ (right) for the pair of agents ($i = 1, j = 7$) in layer $k = 1$. Red lines show the posterior mean for both quantities, together with their 95% and 75% credibility intervals. Blue lines are the true generative quantities. Blue circles show the network data: solid for observed, empty for missing data.

5 Case Study

5.1 Interbank Activity in the Mexican Financial System

The data used for this case study is derived from a database on exposures built and operated by the Central Bank of Mexico (Banxico) with the specific purpose of studying contagion and systemic risk. Banxico gathers information using daily, weekly, and monthly regulatory reports. These reports contain every single funding transaction between most financial institutions in the Mexican Financial

System, on a daily basis, in local and foreign currency. For empirical network studies using this data see Poledna et al. (2015); Molina-Borboa et al. (2015); de la Concha et al. (2018).

We aim to characterise the underlying probabilistic structure that dictates interactions between banks across markets. The results of the inferential process will provide enlightening information about the underlying connectivity of the banking system. We focus our analysis on interbank transaction within the Mexican financial system in three layers:

OCIMN + OCIME: Unsecured lending. This network is built using the information from the regulatory report which contain the unsecured loans in domestic currency (OCIMN) and in foreign currency (OCIME). A link ij is the sum of all the transacted loans granted by i to j in any currency during the time period of aggregation (daily or monthly).

REPO: Repurchase agreement transactions. This network is built using the information from the regulatory report on Repo transactions. A link ij is the sum of all the Repos transacted during the time period of aggregation, regardless of the collateral, currency and maturity.

CVT: Purchase and sale of securities, bilateral transactions. This network is built with the information from the regulatory report on the bond market activity by banks. A link ij represents the total amount of all the securities sold by i to j regardless of the type of securities and the terms of such purchases.

We discuss here the results from two main experiments. We analyze financial data from $V = 46$ banks in the layers described above. We used aggregated monthly data during two periods of analysis to train the model:

- 2008-01-01 to 2009-12-31 (two years observed)
- 2008-01-01 to 2010-12-31 (three years observed)

The predictive horizon was set as 1 year forward in both cases. For inference, we set $\delta = 24$, $H = 5$.

There is a clear structure captured by the model, which allows to easily identify the centrality and relevance of each actor in the system. The estimated propensity of connection between agents in the system for a given month is illustrated in Fig. 5. Each bank is represented by a circle in this graph, whereas the distance between them and the width of line linking them represents the probability of being connected.

The dynamic of the interaction between banks is also learned by the model. In Fig. 6, the evolution of the relation between two banks is depicted. The left panel shows the probability of connection between them as a function of time. We see that there is a clear gradual increase in unsecured transactions from bank “11” to bank “1” during 2019 (time 13 to 24). The right panel shows the expected volume for the transaction (conditional on its occurrence). The blue points in both graphs correspond to the observed outcome (0/1 for link existence in the left and a positive weight for the transaction volume in the right.).



Fig. 5. Force-directed visualization of unsecured market for a (randomly) selected day, estimated by the model. The width of the edges are proportional to the probability of connection between two agents on that day. The size of the nodes is given by their aggregated in-strength.

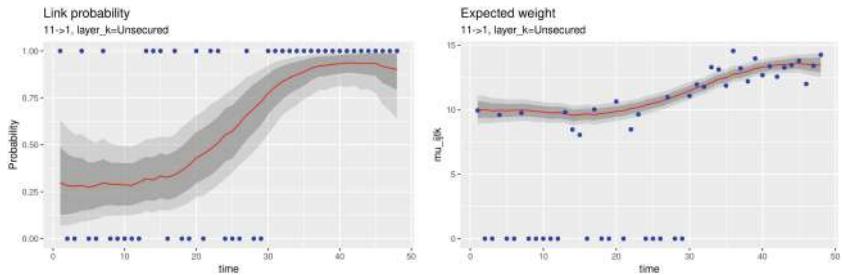


Fig. 6. For a (random) pair of banks ($i = 11, j = 1$), we show the dynamic of the strength of unsecured transactions from bank i to bank j across time (monthly activity, time = 0 corresponds to Jan-2008). On the left, the posterior distribution for the probability of connection, Mean \hat{p} in red, and the observed activity represented by the blue dots (0/1). On the right, The expected amount for the transaction (conditional on being connected), and the actual amount also represented by the blue dots. All quantities are estimated simultaneously within a comprehensive Bayesian model.

We evaluate predictive accuracy of edge formation for each layer as measured by the ROC curve and the associated area under the ROC Curve, AUC. Out-of-sample performance was obtained by predicting 12 months following after the training period and comparing to the observed (left-out) values. In Fig. 7 we show the Area-Under-the-Curve for predictions made for both in-sample and out-of-sample periods. The performance is outstanding, reaching almost 1 for in-sample edges; slowly decaying for out-of-sample data, but still with values above 0.85 for all layers.

Regarding the prediction of weights, in Fig. 8 we compare observed vs predicted weights for the three layers. The left plot corresponds to in-sample observations, reaching a strong correlation of 0.89; whereas the right plot corresponds to predicting the first six months after the training data, showing a correlation of 0.77.

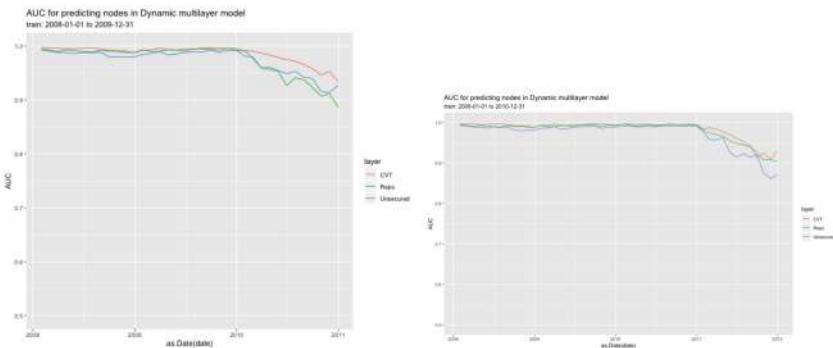


Fig. 7. Assessing model accuracy. Area Under the ROC curve.

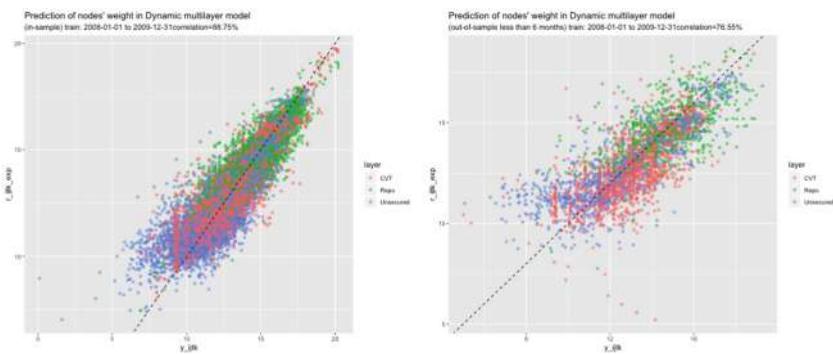


Fig. 8. Assessing model accuracy. Comparison of observed vs predicted link weights.

6 Conclusions

This work provides a comprehensive statistical model that can be useful to capture in a very broad sense complex aspects of interconnectedness in weighted multilayer networks. There very limited number of statistical models which comprise weighted and directed multiplex structures, something which is a crucial element in financial networks applications, in particular those related to systemic risk analysis. The proposed model proved to be effective in capturing important aspects of the complexity of financial networks.

The possible implications in the field of stress testing and systemic risk measurement are many and important. Until now, most of the studies on financial networks rely on limited information to build financial networks (one layer) or single shots (limiting the analysis of the dynamics of such structures).

The model and the software developed usefulness is not restricted to financial networks, there is plethora of fields in which it could be useful and benefit from all these novel techniques which were applied in this work.

References

- Battiston, S., Martínez-Jaramillo, S.: Financial networks and stress testing: challenges and new research avenues for systemic risk analysis and financial stability implications. *J. Financ. Stab.* **35**, 6–16 (2018)
- Crane, H.: Probabilistic Foundations of Statistical Network Analysis, 1st edn. Chapman and Hall/CRC (2018)
- de la Concha, A., Martínez-Jaramillo, S., Carmona, C.: Multiplex financial networks: revealing the level of interconnectedness in the banking system. In: Complex Networks & Their Applications VI, pp. 1135–1148. Springer (2018)
- Durante, D., Dunson, D.B.: Bayesian dynamic financial networks with time-varying predictors. *Stat. Probab. Lett.* **93**, 19–26 (2014a)
- Durante, D., Dunson, D.B.: Nonparametric Bayes dynamic modelling of relational data. *Biometrika* **101**(4), 883–898 (2014b)
- Durante, D., Dunson, D.B.: Locally adaptive dynamic networks. *Ann. Appl. Stat.* **10**(4), 2203–2232 (2016)
- Durante, D., Mukherjee, N., Steorts, R.C.: Bayesian learning of dynamic multilayer networks. *J. Mach. Learn. Res.* **18**, 1–29 (2017)
- Hoff, P.D.: Additive and multiplicative effects network models (2018)
- Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. *J. Am. Stat. Assoc.* **97**(460), 1090–1098 (2002)
- Kim, B., Lee, K.H., Xue, L., Niu, X.: A review of dynamic network models with latent variables. *Stat. Surv.* **12**, 105–135 (2018a)
- Kim, B., Niu, X., Hunter, D.R., Cao, X.: A dynamic additive and multiplicative effects model with application to the united nations voting behaviors (2018b)
- Linardi, F., Diks, C.G.H., van der Leij, M., Lazier, I.: Dynamic interbank network analysis using latent space models. *SSRN Electron. J.* (2017)
- Molina-Borboa, J., Martínez-Jaramillo, S., Lopez-Gallo, F.: A multiplex network analysis of the Mexican banking system: link persistence, overlap. *J. Netw. Theory Finance* **1**(1), 99–138 (2015)
- Poledna, S., Molina-Borboa, J.L., Martínez-Jaramillo, S., van der Leij, M., Thurner, S.: The multi-layer network nature of systemic risk and its implications for the costs of financial crises. *J. Financ. Stab.* **20**, 70–81 (2015)
- Polson, N.G., Scott, J.G., Windle, J.: Bayesian inference for logistic models using Pólya-Gamma latent variables. *J. Am. Stat. Assoc.* **108**(504), 1339–1349 (2013)
- Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (2005)
- Sewell, D.K., Chen, Y.: Latent space models for dynamic networks. *J. Am. Stat. Assoc.* **110**(512), 1646–1657 (2015)
- Sewell, D.K., Chen, Y.: Latent space models for dynamic networks with weighted edges. *Soc. Netw.* **44**, 105–116 (2016)
- Sewell, D.K., Chen, Y.: Latent space approaches to community detection in dynamic networks. *Bayesian Anal.* **12**(2), 351–377 (2017)
- Tran, D., Blei, D., Airoldi, E.M.: Copula variational inference. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28, pp. 3564–3572. Curran Associates, Inc. (2015)

- Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**(5), 1413–1432 (2017)
- Ward, M.D., Ahlquist, J.S., Rozenas, A.: Gravity’s rainbow: a dynamic latent space model for the world trade network. *Netw. Sci.* **1**(01), 95–118 (2013)
- Watanabe, S.: Algebraic Geometry and Statistical Learning Theory. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press (2009)



Influence of Countries in the Global Arms Transfers Network: 1950–2018

Sergey Shvydun^{1,2}

¹ National Research University Higher School of Economics,
Myasnitskaya Str. 20, 101000 Moscow, Russia
shvydun@hse.ru

² V.A. Trapeznikov Institute of Control Sciences of Russian Academy
of Science, Profsoyuznaya Str. 65, 117342 Moscow, Russia

Abstract. Using the SIPRI Arms Transfers Database covering all trade in military equipment over the period 1950–2018, we examine the relationship between countries from a novel empirical perspective. We consider the arms transfers network as a multiplex network where each layer corresponds to a particular armament category. First, we analyze how different layers overlap and elucidate main ties between countries. Second, we consider different patterns of trade in order to identify countries specializing on particular armament categories and analyze how they change their export structure in dynamic. We also examine how countries influence each other at different layers of multiplex network. Finally, we analyze the influence of countries in the whole network.

Keywords: Arms transfers · Influence · Multiplex network · Dynamic

1 Introduction

The transfer of weapons and other military equipment have a central focus in the field of security studies and a pressing matter for policymakers and political activists. It was justified that arms are employed as a potential instrument of influence [1]. Arms transfers are also significant and positive predictors of increased probability of war [2]. Arms trade ought to be understood as both economic and political transactions. Despite this, network-analytic approaches to arms transfers are surprisingly infrequent. Some deploy social network analysis for primarily descriptive purposes, especially for visualization [3] and calculating basic statistics, even if they go on to model dyadic relations using standard econometric methods [4]. Unfortunately, very few works consider the arms trade network as dynamic [5] or take in into account how countries interact to each other with respect to different armament categories. However, there is no doubt that application of social network analysis to the arms trade has a huge potential for researchers and practitioners engaged in power transition studies [6].

In this paper, we address the arms trade from a novel empirical perspective. We consider the arms transfers network as a multiplex network. The main goal of the paper is to analyze relationships between countries with respect to the network structure. We analyze how network layers overlap in order to understand whether the trade of product α is accompanied with the trade of product β and how it is changed in dynamic.

Another important issue is to understand whether countries tend to specialize in particular armament categories and how their export structure is changing over time. Finally, we need to understand how countries influence each other globally and at different armament categories and which countries are the most influential players.

The structure of the paper is organized as follows. First, we describe the arms trade dataset and present its preliminary analysis. Second, we analyze the export structure of countries and its dynamic. Third, we examine how countries influence each other at different layers of multiplex network as well as in the aggregated multiplex network. The final Section concludes.

2 Data Description

2.1 SIPRI Arms Transfers Database

To analyze the global arms trade network we used the SIPRI Arms Transfers Database from the Stockholm International Peace Research Institute (SIPRI) [7]. The database contains information on all transfers of major conventional weapons from 1950 to 2018 (the current version was published on 11 March 2019).

Each arm deal is characterized by the type and number of weapon systems ordered and delivered, the years of deliveries and its financial value. SIPRI has developed a unique pricing system to measure arms transfers using a common unit—the SIPRI trend-indicator value (TIV). According to [8], the TIV measures transfers of military capability rather than the financial value of arms transfers.

To understand and demonstrate main features of the global arms transfers' network, we analyze the data and present its main statistic.

2.2 Data Analysis

The trade dynamics in terms of TIV values is illustrated in Fig. 1.

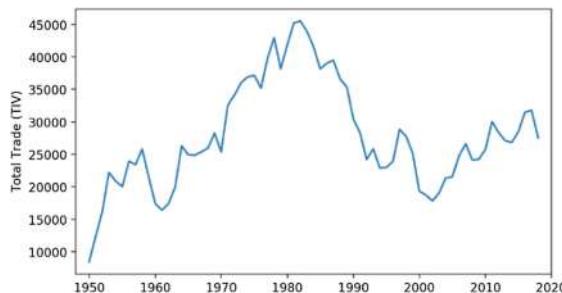


Fig. 1. Dynamics of global arms trade in TIV.

According to Fig. 1, the volume of international arms trade increased over time, peaked in 1982, and following the end of the Cold War, there was a steady decline in global arms transfers. The following years the trade reached the lowest point in 2002

and amounted to approximately 39% of the highest level. As for the last years, there is a rise of arms trade, which is 68% higher comparing to 2002.

As for the main exporters, the United States and the Soviet Union were largest players of international arms transfers until the collapse of the Soviet Union and combined up to 80% of the total trade. France, the United Kingdom and Germany are the next major exporters during the whole period. From 1990s, the United States and Russia became the main exporters, however, their share slightly decreased from 62–68% in 1990s to 51–58% in 2010s. There is also a rise of arms transfers of China in 2010s that accounts for 5–7% of the total value.

All weapons and other military equipment are combined in the following armament categories: air defense systems, aircrafts, armored vehicles, artillery, engines, missiles, naval weapons, satellites, sensors, ships and other weapons. In Fig. 2 we present how the structure of global trade evolves over time.

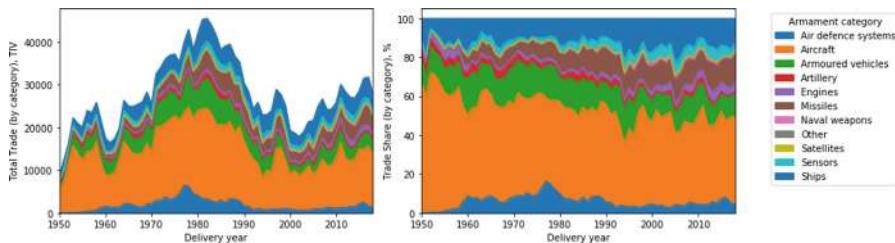


Fig. 2. Structure of the arms trade by armament category (TIV and share).

According to Fig. 2, aircrafts is the major armament category that varied from 65–70% of the total value in 1950s to 40–45% in 2010s. Ships and armored vehicles are stable categories that account for 10–20% during all years. We can also observe the rise of missiles trade in late 1950s, which reached 15% in 2018. Additionally, due to the increase in aircrafts and missiles, there has been a high demand for air defense systems from 1960, which peaked in late 1970s (16%), after which the trade declined to 4–5%.

All weapons and other military equipment are also divided by their status (see Fig. 3). Most arms transfers include new weapons (85–95%) but there are also two peaks of the second hand weapons: 34% after the Second World War and 22% in 1994, which can be explained by the collapse of the Soviet Union.

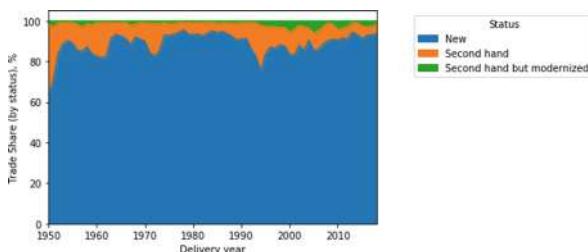


Fig. 3. Structure of the arms trade by status (share).

We constructed the global arms transfer network for each year and illustrated the basic descriptive statistic in Fig. 4. One can observe that the total number of countries engaged in arms transfers has increased from 54 in 1951 to 134 in 2016. In 2017–2018 the total number of countries decreased, however, it can be explained by the fact that some countries have not reported their statistics yet. We also calculated the total number of components in each network and detected that 51 of 69 networks contain only 1 component while all other networks contain one giant component that includes 95% of nodes (except the network in 1954). We can also observe that countries are engaged in trade relations with 5 countries in average while the global network is sparse as its density varies from 0.04 to 0.06.

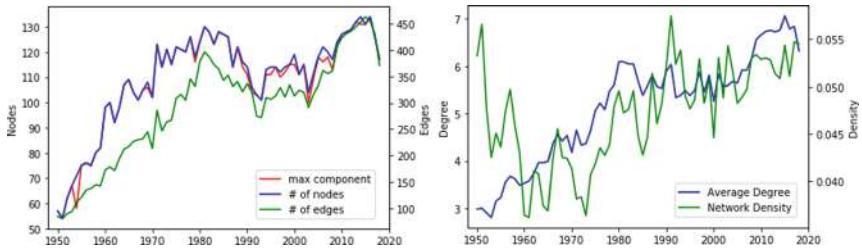


Fig. 4. Arms trade network statistics

One can also attempt to analyze main ties in the arms transfer network. We can interpret the main ties as pairs of countries that have a long history of trade relations and calculate it as

$$t_{ij} = \frac{|\{y|(i,j) \in E_y | (j,i) \in E_y\}|}{|\{y|i \in N_y \& j \in N_y\}|}, \quad (1)$$

where E_y is a set of edges in arms trade network in year y , N_y is a set of nodes in arms trade network in year y . In other words, t_{ij} indicates how often countries i and j trade with each other while they are presented in a network.

Overall, main ties were observed between the largest exporters of weapon – the United States and the Soviet Union (after 1991 - Russia) – and some of their importers. However, some other ties between other states are also observed (see Table 1, numbers in brackets correspond to the total number of years).

According to Table 1, main partners of the United States are Canada, some European countries and, interestingly, South Korea and Turkey. The arms trade between the USA and these countries was during the whole period (69 networks). The main partners of the Soviet Union are European countries of the Eastern Bloc, North Korea and South Yemen that were highly supported by the USSR. Contrary to the Soviet Union, the main partners of Russia are China, India and Iran.

Table 1. Main ties in arms trade network

Country A	Country B	t_{ij}
United States	Canada (69), France (69), Greece (69), Italy (69), Japan (67), Netherlands (69), South Korea (69), Spain (67), Turkey (69)	1
Soviet Union	Bulgaria (42), Czechoslovakia (42), East Germany (GDR, 39), Hungary (42), North Korea (42), Poland (42), Romania (42), South Yemen (22)	1
Russia	China (27), India (27), Iran (27)	1
United States	Argentina (67), Australia (66), Brazil (64), Colombia (66), Denmark (65), Germany (64), Israel (68), Norway (67), Philippines (63), South Vietnam (21), Taiwan (67), Thailand (67), United Kingdom (68)	0.95–0.98
Russia	Kazakhstan (22)	0.95
United Kingdom	India (65), Yugoslavia (38)	0.9–0.94
France	India (62), Spain (59)	0.88–0.89
China	Ukraine (24)	0.88
Russia	Viet Nam (24), Algeria (23), Egypt (23)	0.85–0.88

In addition to t_{ij} value, we also considered ties with the longest consecutive trade relations. Interestingly, most consecutive ties were between France and India (62 years), China and Pakistan (55), China and Germany (53), Sweden and Pakistan (45) as well as the United Kingdom with France (52), India (50), Japan (45) and Australia (43).

2.3 Export Structure Patterns

Since all weapons and other military equipment are combined in armament categories, we analyzed the export structure of all countries. The main goal of the analysis is to consider different patterns of trade, identify countries specializing on particular armament categories and analyze how they change their export structure in dynamic.

To perform such analysis, we calculated the export structure of all countries with respect to TIV measure and applied agglomerative clustering algorithm. As a result, we obtained 22 different patterns that describe their export structure. For instance, cluster #1 contains countries specializing on sensors: Denmark (late 1990s–2000s), the Netherlands (1960s, 1980s–2000s) and Switzerland (1960s–1990s). Cluster #19 are the countries that export both missiles and sensors: Denmark (2010s), France (early 1990s and 2010s), Israel (1995–2015), Sweden (early 1980s, 2010s) and some other countries. Cluster #20 contains all countries specializing mostly on aircrafts and armored vehicles. Some of the clusters are demonstrated in Fig. 5.

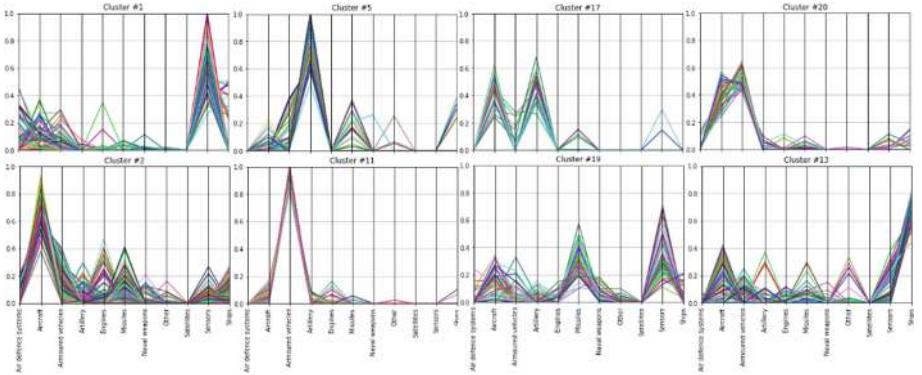


Fig. 5. Example of export structure patterns.

One can analyze how countries change their structure patterns over time. For instance, in early 1950s the Soviet Union was in cluster #2 (aircrafts are the major category (40–90%); other categories are also exported) and in 1956 it moved to cluster #4 (aircrafts is still the main category (20–50%); other categories are also exported). The United States is also in cluster #2 but in some years it moves to cluster #4. China was in cluster #12 (aircrafts are more than 90% of the total export) in 1960s, however, in late 1990s it moved to cluster #4. On the other hand, Germany was in cluster #16 in 1950s (ships are more than 90% of the total export), in 1970s–2000s it was in cluster #18 (ships, missiles and armored vehicles are major categories) and in 2010s it moved to cluster #8 (aircrafts and ships). Overall, one can say that the performed analysis is a valuable tool to analyze how countries evolve their export structure and how it relates to their influence.

2.4 Layers Overlap

Since we consider the arms trade network as a multiplex network, it is also necessary to observe how different layers of the network overlap. In other words, if country A transfers, for instance, missiles to country B , does it also transfer some other armament category to its partner?

To perform such analysis, we calculated the following overlap measure between layers α and β

$$\text{overlap}_y^{\alpha\beta} = \frac{|\{(i,j)|(i,j) \in E_y^\alpha \& (i,j) \in E_y^\beta\}|}{|\{(i,j)|(i,j) \in E_y^\alpha\}|}, \quad (2)$$

where E_y^α and E_y^β are sets of edges in arms trade network for category α and β in year y . In other words, $\text{overlap}_y^{\alpha\beta}$ indicates how often countries that export category α also export category β . One should mention that $\text{overlap}_y^{\alpha\beta}$ measure is not symmetric.

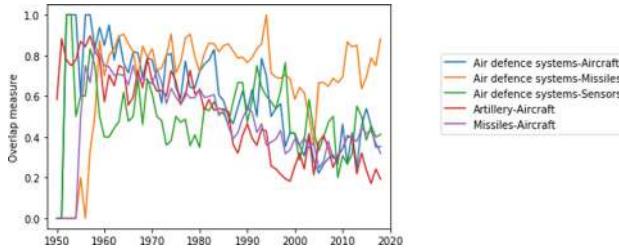


Fig. 6. Overlap measure for armament categories (TOP-5).

The overlap measure for armament categories is provided in Fig. 6. The highest overlap is observed between air defense systems and missiles (orange line, average value is 0.69). In other words, countries that trade air defense systems in 70% cases also transfer missiles to the same country. Air defense systems also overlap with aircrafts (average value is 0.61) and sensors (average value is 0.51) layers. The overlap measure between artillery/missiles and aircrafts layers is around 0.5 in average.

Similarly, we can define the overlap between layers that represents the status of the weapons (see Fig. 7).

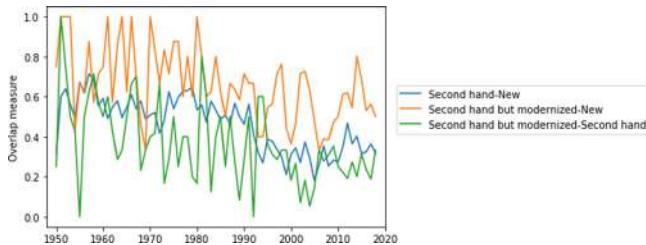


Fig. 7. Overlap measure for status layers (TOP-3).

Contrary to armament layers, we can observe that overlap values for status layers are lower than for armament categories. The highest measure is between second hand but modernized weapons and new weapon layers, which is seems very natural.

3 Influence Analysis

3.1 Methodology

Next, let us consider the arms trade network as a multiplex network and analyze its key participants in each layer as well as in the whole network. Obviously, the United States, the Soviet Union and Russia are likely to be the most important participants in the global arms trade network as it is observed in Sect. 2. However, which countries have high influence values on a particular layer of the network? Are there countries that have relatively low export values and high impact on other participants of the network?

To define the most important elements in a network, various centrality concepts are usually used. Degree, betweenness, closeness, PageRank and Hits centralities are among the most popular centrality measures [9–14]. We calculated these measures, however, since in this work we are focused on the influence in a network, the results are provided for a LRIC measure [15, 16].

First, we need describe how LRIC measure defines the influence in the arms trade network. The LRIC model considers the threshold model and assumes that each country i has some threshold of influence q_i^α at layer α , which indicates the level when this country becomes affected. The threshold value depends on some external parameters of a node (e.g. arms production) or may be calculated with respect to a network structure.

Definition 1. A critical group for country i at layer α is a subset of countries whose group influence exceeds the threshold value q_i^α . More formally, $\Omega(i, \alpha) \subseteq V$ is a critical group for node i if

$$\sum_{j \in \Omega(i, \alpha)} w_{ji}^\alpha \geq q_i^\alpha \quad (3)$$

Definition 2. Country k is pivotal for some group $\Omega(i, \alpha)$ if its exclusion from this group makes the group non-critical. Formally, $\Omega^p(i, \alpha) \subseteq \Omega(i, \alpha)$ is a subset of pivotal countries of group $\Omega(i, \alpha)$ if $\forall k \in \Omega^p(j, l)$

$$\sum_{j \in \Omega(i, \alpha) \setminus \{k\}} w_{ji}^\alpha < q_i^\alpha \quad (4)$$

Definition 3. If country j is pivotal for country i , the direct influence can be evaluated as

$$c_{ji}^\alpha = \max_{\Omega_k(i, \alpha) : j \in \Omega_k^p(i, \alpha)} \frac{w_{ji}^\alpha}{\sum_{h \in \Omega_k(i, \alpha)} w_{hi}^\alpha} \quad (5)$$

Thus, the main idea of the LRIC measure is to elucidate all pivotal members for each node, remove insignificant edges and transform the initial trade network to the network of influence according to formula (5). Additionally, the LRIC measures evaluates how nodes influence each other indirectly by considered various paths between nodes in the network of influence. We should note that LRIC measure has a good correspondence with a notion of influence defined as “manipulation of the arms transfer relationship in order to coerce or induce a recipient- to conform its policy or actions to the desires of the supplier-state” [17].

One could also mention that the largest nodes in terms of our-degree and PageRank values are not always the most influential members according to LRIC measure. This fact can be explained by the following reasons:

- Country A with high export values can trade weapons to less number of countries compared to country B , which export can be more diverse;
- Country A with high export values can trade weapons to many countries but it is not necessarily pivotal for all them. On the other hand, country B can export less weapons but be pivotal for all its partners;
- Very long connections do not play an important role in annual data, as the contagion effect requires time in order to affect distant members.

Thus, we calculated classical measures and LRIC index¹ at each particular layer.

3.2 Influence in Arms Categories

In this paper we evaluated the influence of countries for each armament category. However, to evaluate the LRIC index we first need to define the threshold of influence for each country.

In general, trade relations among countries cannot be considered separately from information of the level of production in these countries. For instance, country A may have high import values and, consequently, be highly dependent on other countries according to the network structure. However, if the total import accounts for a small proportion of the overall production level, none of exporting countries actually influences country A . On the other hand, if the level of production in country A is very low, the country becomes highly dependent on its exporters. Thus, the production level could be a good proxy for the threshold of influence. More characteristics of the recipient which can be used to shape the perceptions of the recipient toward the influence are discussed in [1].

Unfortunately, we did not find any detailed information describing countries production level of a particular armament category. Therefore, we tried to take into account this feature and calculated the threshold of influence of country i as

$$q_i^\alpha = \lambda \cdot \max\left(\sum w_{ki}^\alpha, \sum w_{ik}^\alpha\right), \quad (6)$$

where w_{ki}^α (w_{ik}^α) are the TIV measures of all weapons from armament category α traded from country k to country i (country i to country k) and λ is an additional parameter, $0 < \lambda \leq 1$. The main idea of formula (6) is that if the total export of country i is higher than its total import, country A will be less dependent on import as it sells more than it buys. As for λ value, we considered $\lambda = 0.25$ (low influence threshold), $\lambda = 0.5$ (medium influence threshold), $\lambda = 0.75$ (high influence threshold).

¹ The LRIC library is available in Python and can be downloaded from <https://github.com/SergSHV/srlrc>.

The results for the medium influence threshold are provided in Fig. 8.

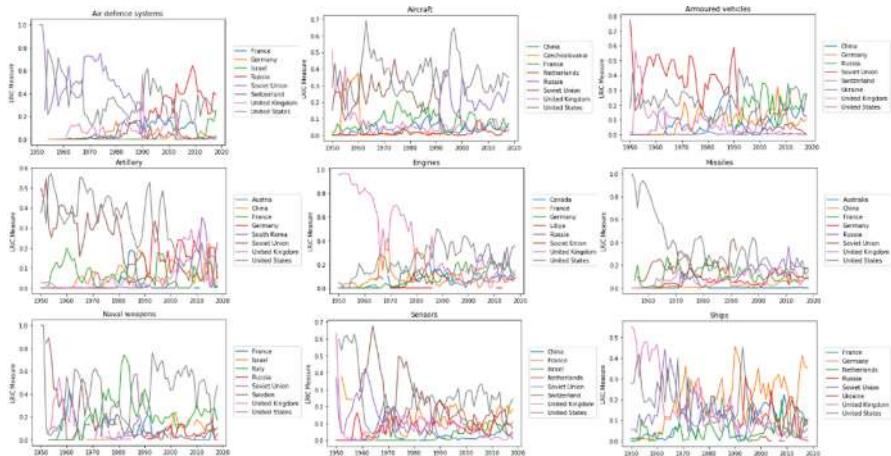


Fig. 8. Influence of countries in armament categories.

According to Fig. 8, the United States, the Soviet Union and the United States are among the most influential members in most armament categories. However, these countries are not necessarily the most influential participants across all layers. For instance,

- Switzerland was the most influential country during 1960–1980s in sensors trade network;
- Italy played an important role in early 1980s in naval weapons trade network;
- South Korea have the highest values in early 2010s by artillery;
- Germany was the most influential country in late 2000s by armored vehicles, it is also among key players in the ships trade network from 1970.

Overall, we can observe two main features. First, some countries specialize on particular armament categories and play an important role in arms transfers. Second, although during 1950–1970s United States, the Soviet Union and the United Kingdom monopolized some categories in terms of LRIC values, their relative importance has decreased and now it is comparable to some other exporters.

3.3 Influence in Multiplex Arms Transfers Network

After evaluating the total influence of countries in different armament categories, we can aggregate this information and estimate the total influence of countries in a aggregated multiplex network.

There are several models how to assess the influence in such networks. One of the main approaches is based on introduced concepts of supra-adjacency and supra-Laplacian (i.e., super-Laplacian) matrices, which are focused to solve eigenvalue

problems in multiplex networks [18, 19]. In this approach, the same nodes, which are presented in different layers, should be connected to each other in order to preserve the main idea: the centrality of a node depends not only on its neighbors of a specific layer but also on all neighbors of other layers. Unfortunately, in our case it is not clear how different layers of armament categories are related to each other and whether the influence on layer α should be taken into account on layer β .

In this paper we consider another approach proposed in [20] which can be applied in case there is no clear dependency across different layers. The approach includes the following steps. First, one should obtain information on pairwise influence of nodes at each individual layer. Second, it is necessary to calculate the total number of layers when node i influences node j more than vice versa. These calculations can be performed with respect to the importance of layers (evaluated, for instance, by sum of degrees). Third, an aggregated network is constructed. For instance, it is possible to construct a majority graph where node i influences node j if it influences this node in more layers. Finally, various concepts from social choice theory are applied to the aggregated network.

Following the paper [20] we aggregated all layers with respect to LRIC matrix that defines how nodes influence each other on a particular layer. In the aggregated network we calculated the Borda score which shows on how many layers node i influences other nodes with respect to the importance of layers. Additionally, we calculated the Copeland score that evaluates the total number of dominated countries by a majority relation. The results are provided in Fig. 9.

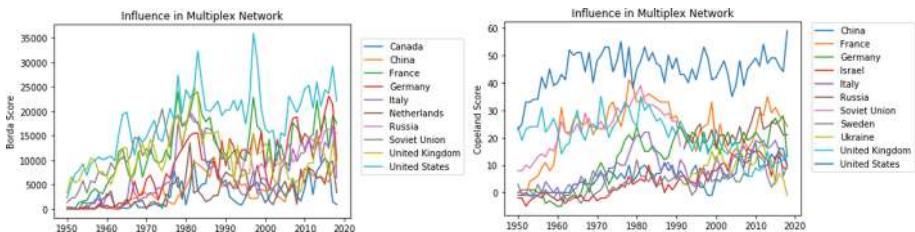


Fig. 9. Influence in multiplex network (Borda and Copeland Scores)

According to Fig. 9, the United States, the United Kingdom, the Soviet Union, France and Russia are the most influential countries in the aggregated multiplex network. For instance, the United States dominates 30–50 other countries by majority relation which was constructed based on information how countries influence each other according to different layers. We can also observe that the influence of the United Kingdom decreased during last years. On the other hand, we can also observe some other countries emerged like Canada, China, Germany, Italy and the Netherlands (by Borda Score) or China, Israel, Italy, Sweden and Ukraine. These countries were among most influential members during short periods. Overall, we can see that these three rankings correlate with each other, provide similar results and reveal countries that influence other countries with respect to several products simultaneously.

4 Conclusion

The study of the arms transfer is ripe for the use of network theories and methods. In our paper we analyzed the network with respect to its multiplex structure. There were observed that some layers have high overlap values like air defense systems and missiles layers. We also divided all countries according to their export structure, identified countries specializing on some categories and analyzed the dynamic of main exporters.

Our main focus is on influence analysis. We considered the threshold model and applied it to each individual layer. Although, there are some evident leaders like the United States, the United Kingdom, the Soviet Union and Russia, our models also detected some other countries that have high influence values on particular layers. The more detailed analysis of the obtained results is required. We also aggregated information on pairwise influence of countries at different layers and presented the overall influence structure.

Acknowledgments. The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and supported within the framework of a subsidy by the Russian Academic Excellence Project ‘5–100’. The influence analysis of countries (Sect. 3) was funded by the Russian Science Foundation under grant № 17-18-01651.

References

1. Sislin, J.: Arms as influence: the determinants of successful influence. *J. Confl. Resolut.* **38**(4), 665–689 (1994)
2. Craft, C., Smaldone, J.: The arms trade and the incidence of political violence in sub-saharan Africa, 1967–97. *J. Peace Res.* **39**(6), 693–710 (2002)
3. Kinsella, D.: Changing structure of the arms trade: a social network analysis. Presented at the Annual Meeting of the American Political Science Association. Philadelphia, PA (2003)
4. Akerman, A., Seim, A.L.: The global arms trade network 1950–2007. *J. Comp. Econ.* **42**(3), 535–551 (2014)
5. Thurner, P.W., Schmid, C.S., Cranmer, S.J., Kauermann, G.: Network interdependencies and the evolution of the international arms trade. *J. Confl. Resolut.* **63**(7), 1736–1764 (2019)
6. Kinsella, D.: Power transition theory and the global arms trade: exploring constructs from social network analysis. Political Science Faculty Publications and Presentations (2013)
7. SIPRI: The SIPRI Arms transfers database (2019). Accessed 1 July 2019
8. Simmel, V., Holtom, P., Bromley, M.: Measuring international arms transfers. Stockholm International Peace Research Institute (SIPRI) (2012)
9. Freeman, L.C.: Centrality in social networks: conceptual clarification. *Soc. Netw.* **1**, 215–239 (1979)
10. Bonacich, P., Lloyd, P.: Eigenvector-like measures of centrality for asymmetric relations. *Soc. Netw.* **23**, 191–201 (2001)
11. Katz, L.: A new status index derived from sociometric index. *Psychometrika* **18**, 39–43 (1953)
12. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *J. ACM* **46**(5), 604–632 (1999)

13. Freeman, L.C.: A set of measures of centrality based upon betweenness. *Sociometry* **40**, 35–41 (1977)
14. Rochat, Y.: Closeness centrality extended to unconnected graphs: the harmonic centrality index. *ASNA* (2009)
15. Aleskerov, F., Meshcheryakova, N., Shvydun, S.: Centrality measures in networks based on nodes attributes, long-range interactions and group influence. arXiv preprint [arXiv:1610.05892](https://arxiv.org/abs/1610.05892)
16. Aleskerov, F.T., Meshcheryakova, N.G., Shvydun, S.V.: Power in network structures. In: Kalyagin, V.A., Nikolaev, A.I., Pardalos, P.M., Prokopyev, O. (eds.) *Models, Algorithms, and Technologies for Network Analysis. Springer Proceedings in Mathematics & Statistics*, vol. 197, pp. 79–85. Springer (2017)
17. Wheelock, T.: Arms for Israel: the limit of leverage. *Int. Secur.* **3**(Fall), 123–137 (1978)
18. Cozzo, E., de Arruda, G.F., Rodrigues, F.A., Moreno, Y.: Multilayer networks: metrics and spectral properties. In: Garas, A. (eds.) *Interconnected Networks. Understanding Complex Systems*. Springer, Cham (2016)
19. Solé-Ribalta, A., De Domenico, M., Kouvaris, N.E., Díaz-Guilera, A., Gómez, S., Arenas, A.: Spectral properties of the Laplacian of multiplex networks. *Phys. Rev. E* **88**, 032807 (2013)
20. Shvydun, S.: Influence assessment in multiplex networks using social choice rules. *Procedia Comput. Sci.* **139**, 182–189 (2018)

Biological Networks



Network Entropy Reveals that Cancer Resistance to MEK Inhibitors Is Driven by the Resilience of Proliferative Signaling

Joel Maust¹, Judith Leopold¹, and Andrej Bugrim^{2(✉)}

¹ University of Michigan Medical School, Ann Arbor, MI 48109, USA

² Silver Beach Analytics, Inc., St. Joseph, MI 49085, USA

andrej@silverbeachtech.com

Abstract. Recently MEK kinase inhibitors emerged as a promising treatment for KRAS-mutant tumors. However, clinical success remains elusive due to drug resistance. To better understand the mechanism of such resistance, we consider drug response as a transition from proliferative to apoptotic state of the molecular network and search for network metrics linked to likelihood of such transition. We focus on the dynamic network entropy – a statistical property related to stability and robustness of network states. We calculate network entropy metrics for approximately 400 cell lines from the Cancer Cell Line Encyclopedia, representing a broad variety of cancers. We investigate correlation between these metrics and cellular response to a MEK-kinase inhibitor drug PD-0325901. We find that network entropy rates of proteins and pathways related to the cell cycle exhibit the most significant differences between groups of sensitive and resistant cell lines. Our results suggest that resistance to MEK kinase inhibition is driven by the overall resilience of the network of proliferative signaling. We confirm this experimentally by observing synergy between MEK and CDK4/6 inhibitors in select cancer cell lines with high network entropy rates of the G2/M transition pathway. Our findings show that network entropy metrics can become a promising predictor of drug sensitivity. They can be used where gene-level markers are not available, provide insights into functional mechanisms of resistance and guide combination therapy selection.

Keywords: Protein interaction networks · Network entropy · Drug resistance · Network robustness

1 Introduction

Oncogenic mutations of KRAS define the largest genomic subset of many tumor types, including pancreatic (~90%), colorectal (~50%), and non-small cell lung cancers (~25%) [1–3]. Overall KRAS is mutated in 20% of all cancer patients, which translates to approximately 3 million new cases a year globally [4]. Recently MEK kinase inhibitors emerged as a promising treatment for patients with tumors harboring mutations in KRAS gene. However, the complexity of KRAS mutant signaling is challenging and intervention directed against downstream signaling kinases, including MEK, have led to disappointing clinical activity [5, 6]. MEK is a central player in the

KRAS signaling pathway and the number of MEK inhibitor-based therapeutic strategies and trials to address diversity of resistance mechanisms has steadily risen during the last decade [7–10]. Most of these strategies are based on combining a MEK inhibitor with another drug, targeting mechanisms associated with resistance.

Nevertheless, adequate understanding of resistance mechanisms remains elusive. This is not surprising, given that cellular behavior, including responses to drugs is defined by the collective dynamics of large-scale gene-regulatory and protein-interaction networks. The full statistical description of these networks would require knowledge of joint probability distribution of the high-dimensional vectors of genomic states. In practice this distribution is hard to evaluate, because the number of degrees of freedom far exceeds the number of available data points (biological samples). Consequently, many bioinformatics workflows rely on average values and variance of individual genomic variables, and do not account for collective dynamics and higher-order statistical measures.

A promising way to circumvent this problem is to apply ideas from statistical physics to the analysis of protein interaction networks in living cells [11–13]. Physical systems can be described in terms of state variables: pressure, volume, entropy, free energy, etc. [14]. Small number of these quantities fully characterize macroscopic states and state transitions, whereas the number of underlying microscopic states is large. Likewise, in living cells there could be large number of different gene expression profiles corresponding to the same cellular phenotype [15–18]. Using the analogy with physical systems, gene expression profiles could be considered “microscopic genomic states”. It is reasonable to assume the existence of some collective properties of protein networks that play the role of macroscopic state variables and can provide a low-dimensional description of a biological system. In this respect stability of and transitions between cellular phenotypes may bear more than a superficial resemblance to phase transitions in statistical physics [19].

Using this reasoning, we look for biological analogs of macroscopic state variables that determine behavior of the protein network in response to a drug. We start with the hypothesis that drug resistance is correlated with cellular robustness and network’s ability to dissipate perturbations. We reason that networks that do not dissipate perturbations very effectively are likelier to respond to treatment by undergoing significant irreversible state transitions, resulting in a “drug-sensitive” cellular phenotype. In contrast, resilient networks that effectively dissipate drug-induced perturbations are likelier to remain in their original state, leading to drug resistance. Such “dynamical robustness” can be linked to the concepts of “network entropy” and “network entropy rates” [20–22] which we investigate in this paper.

2 Network Entropy and Cellular Robustness

Generally defined, network entropy refers to the measure of uncertainty in the transmission of signaling information along the edges of the protein-protein interaction (PPI) network. The edges can be thought of as possible options for transmitting a signal

with edge weights representing relative probabilities of transmission along different routes. Entropy of an individual node i is defined as [20]:

$$S_i = - \sum_{j \in N_i} p_{ij} \log p_{ij} \quad (1)$$

where N_i are all the nearest neighbors of the node i and p_{ij} are probabilities of signal transmission to each of them. Network entropy is closely related to the presence of the multitude of “alternative network paths.” Node’s entropy can be viewed as a measure of signaling promiscuity. It reaches its maximum value when all probabilities are equal. For a node with one predominant transmission route the entropy is low. The relation between signaling entropy and cellular robustness can be understood by considering propagation of perturbations on the network. A high entropy hub node randomly channels perturbations to different paths and they are more likely to dissipate. However, this local dissipation also depends on how likely a perturbation is to reach the hub in the first place. Unlike local entropy, this probability depends on the global topology and weights of all edges of the network. Propagation of perturbations can be modeled as a random walk on the PPI network with the stochastic matrix defined by probabilities p_{ij} . Stationary distribution of this random walk is a vector of visitation probabilities π_i for every node. The local “entropy rate” for each network node and the global entropy rate of the entire network are defined as [20]:

$$H_i = -\pi_i \sum_{j \in N_i} p_{ij} \log p_{ij} = \pi_i S_i \quad (2a)$$

$$H = \sum_{i \in N} H_i \quad (2b)$$

Here N_i is the set of the nearest neighbors of node i , N is the set of all network nodes. As seen from the Eqs. (2a-2b), entropy rate is additive: network entropy rate is the sum of the local rates of its nodes. Local rate is a product of two factors: the probability π_i that a perturbation reaches a node and the local signaling entropy of that node, S_i characterizing its ability to dissipate perturbations. It is intuitively clear that a network where high visitation probabilities correspond to high entropy nodes is the most efficient in dissipating perturbations and is therefore robust. Formally, this relation between network entropy rate and its robustness is based on the fluctuation theorem for networks [20] which states that changes in network entropy rate, and the decay rate R of perturbations, are positively correlated:

$$\Delta H \Delta R > 0 \quad (3)$$

Notably, values of π_i for each node are calculated using the global network and each of them effectively contains information on global network topology and weights of all edges. Therefore, even at the level of individual nodes, network entropy rates should be considered global variables.

The relation between network entropy and cellular robustness has been confirmed experimentally in multiple studies. For example, it has been demonstrated that network entropy rate is increased in cancer cells, consistent with their higher level of robustness [23]. In *C. elegans* synthetic lethality was shown to be strongly correlated with a gene's local network entropy [21]. Finally, it has been established that sensitivity to a variety of cancer drugs is correlated with the local network entropy of their targets [22]. This relation constitutes the basis for our hypothesis that entropy metrics of the protein interaction network could be effective descriptors of drug response.

3 Resistance to MEK Inhibition in Cancer Cell Lines

3.1 Network Entropy Rates of Cell Cycle Proteins Correlate with Drug Resistance

To test our hypothesis, we have analyzed responses of cancer cell lines to a MEK-kinase inhibitor drug PD-0325901. Our goal was to find network entropy metrics correlated with drug sensitivity. We have selected 409 cell lines from the Cancer Cell Line Encyclopedia (CCLE [24]) with a range of sensitivities to PD-0325901. They included 146 “sensitive” lines ($IC_{50} < 1 \mu M$) and 273 “resistant” lines ($IC_{50} > 8 \mu M$). Each category contained lines representing various types of cancer. For each cell line we have computed network entropy metrics (node entropy and node entropy rates) for approximately 8,000 protein nodes that were part of the curated protein-protein interaction network obtained from HPRD database [25]. Normalized gene expression and drug sensitivity data for selected cell lines were obtained from the CCLE website [26]. Gene expression values for each of the selected cell lines were mapped onto the PPI network by matching gene IDs. Construction of the stochastic matrix for modeling random walk was based on the mass-action principle, as proposed in [22, 27]. The edges of the PPI network were assigned weights that are proportional to the product of normalized expression levels of genes mapped onto the adjacent nodes. The computation of the entropy metrics was performed with the R-package CompSR [22].

First, we have identified “differential entropy nodes” or DENs - network nodes with statistically higher values of entropy rates in resistant cell lines. To this end we have performed a t-test, followed by the multiple-testing correction. A mild adjusted p-value threshold of 0.05 was applied. Finally, we have selected a subset of nodes with the higher average values of entropy rates in the group of resistant cell lines. This procedure yielded 1024 DENs. For functional characterization of these nodes we used ReactomePA package [28] to perform enrichment analysis of ~ 500 human pathways by the proteins corresponding to these nodes. The results indicated that many of the top pathways by enrichment are related to the cell cycle (Fig. 1). We repeated this analysis using progressively lower thresholds of 500, 200 and 100 nM to define “sensitive” lines to ensure that results remain valid throughout the range of plasma concentrations of the drug observed in pharmacokinetics studies [29, 30]. We have observed that at all levels of sensitivity threshold functional characterization of differential entropy nodes remains practically unchanged with cell cycle related pathways dominating enrichment.

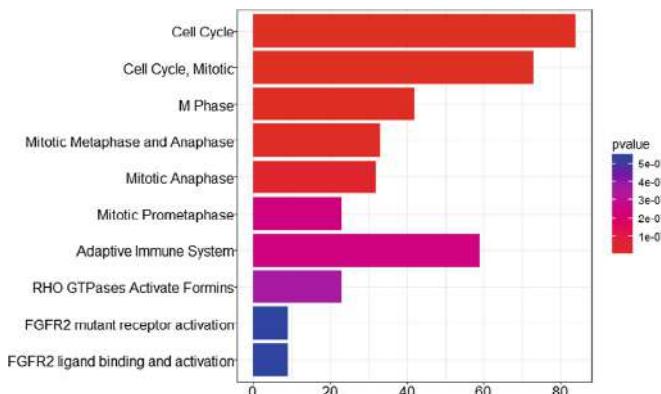


Fig. 1. Enrichment by differential entropy nodes (DENs) is dominated by pathways related to the cell cycle (adjusted p-value = 0.05).

For comparison, we have also performed functional analysis based on differential gene expression. The set of differentially expressed genes (DEGs) was generated using similar procedure, resulting in 2098 DEGs. Enrichment analysis confirms that the top pathways by enrichment are cell cycle associated processes. Next, we applied more stringent criteria for selecting significant DENs and DEGs by reducing adjusted p-value threshold. Comparison of pathway enrichment by the resulting sets shows striking differences between the functional properties of differentially expressed genes and nodes with differential entropy. Decreasing p-value threshold by 10-fold reduces the number of DENs by >50% (from 1024 to 446) but results only in minor changes in the corresponding functional profile. In contrast, even more modest 5-fold decrease in p-value threshold, leading to ~25% reduction in the number of selected DEGs (from 2098 to 1591), results in significant changes in the functional profile. To further investigate robustness of functional profiles of DENs and DEGs, we performed functional enrichment analysis using sliding threshold, varying the adjusted p-value in the range between 0.05 and 0.0025. For each value of the threshold we have compared composition of the top 10 pathways enriched by the corresponding sets of DENs and DEGs with top pathways obtained under the least stringent threshold (adjusted p-value = 0.05). In the case of DENs, applying increasingly selective threshold does not result in significant changes in the top 10 pathways until only approximately 350 significant nodes are retained. At that point functional picture changes dramatically and no longer resembles the profile of the larger sets. To the contrary, functional profile of DEGs changes gradually as p-value threshold becomes more stringent (Fig. 2).

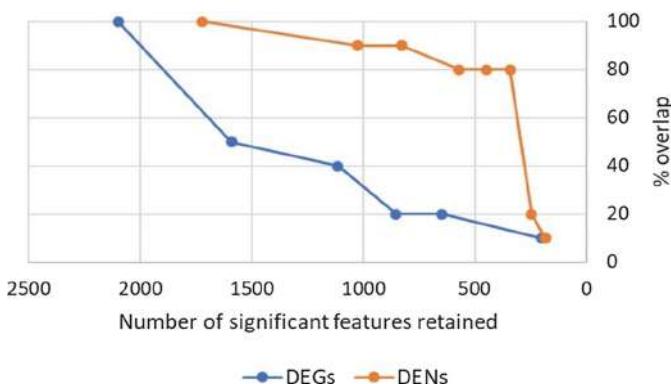


Fig. 2. Overlap of the top ten pathways by enrichment vs. the number or retained significant features for DENs and DEGs at different levels of adjusted p-value threshold. Overlap is a percentage of pathways that appear among the top ten at both current and the least stringent p-value threshold.

3.2 Entropy Rates Can Guide Combination Therapy to Overcome Resistance

The results above indicate that resistance to MEK inhibitors correlates with the higher than average network entropy rates of the proteins and processes related to the cell cycle. It has been shown that this resistance can be overcome by targeting cell cycle via the inhibition of CDK4/6 [7–10], suggesting that network entropy could guide combination therapy selection. To confirm this, we have measured synergy between MEK and CDK4/6 inhibitors in two cell lines with the high entropy of the cell cycle. First, we have evaluated pathway-level entropy rates. Using the fact that network entropy rate is an additive quantity (Eq. 2b), we computed them as sums of entropy rates of the individual network nodes that constitute each pathway. Next, we selected “G2/M transition” pathway (defined by the Reactome database [28]) as the representative core of the cell cycle and calculated Z-score for each cell line using the deviation of its G2/M transition pathway entropy rate from the average value across all cell lines. When cell lines are arranged by the resulting Z-scores, the high-entropy end is clearly dominated by the resistant lines, while the low-entropy end is dominated by the sensitive lines (Fig. 3). We have selected two high-entropy cell lines that were resistant to PD-0325901 and were readily available: NCI H460 (Z-score = 1.0) and Calu-1 (Z-score = 1.3) and treated both with the combination of trametinib (highly selective MEK inhibitor) and palbociclib (highly selective inhibitor of CDK4/6) with concentrations ranging from 1 nM to 10 μ M. Cells were incubated for 3 days in the continuous presence of drug or DMSO and viability was measured using CellTiter-Glo (Promega). Viability was calculated as a percentage of the DMSO treated cells. Resistance to MEK-inhibition was confirmed with 64% and 55% viability at 10 μ M of trametinib. Drug combinations proven to be highly synergistic (Fig. 4) with the synergy scores of 6.98 (NCI H460) and 3.16 (Calu-1). These results provide an early indication that network entropy could serve as a marker of response to combination therapies and can be used in guiding its selection for individual patients.

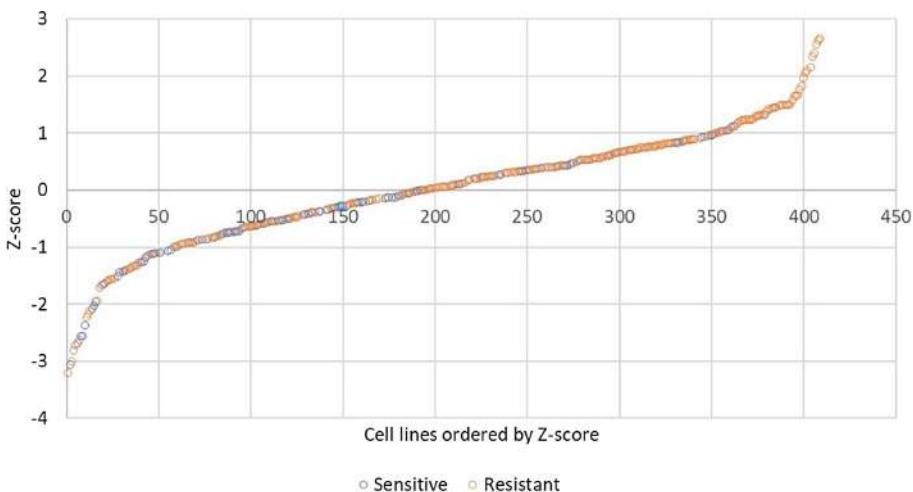


Fig. 3. Network entropy rate of “G2/M transition” pathway and drug resistance. Cell lines are arranged by the entropy rate of this pathway and are color-coded by drug sensitivity. The high-entropy end is dominated by the resistant cell lines while most of the sensitive lines are clustered near the low-entropy end.

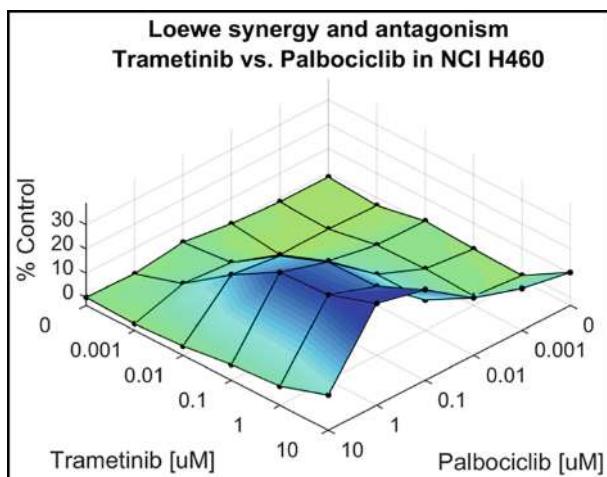


Fig. 4. Drug synergy analysis of a representative high-entropy lung cancer cell line (NCI H460) Treated with varying concentrations of the MEK inhibitor trametinib and CDK4/6 inhibitor palbociclib.

4 Discussion

According to multiple studies, most of biological functionality is carried out by network modules – highly interconnected subsets of the global protein network comprising anywhere from tens to hundreds of proteins [31–34]. Functional differences between cell lines ought to be reflected in the differences in expression of genes corresponding to relevant module(s). However, these differences are often small and, given the noise always present in gene expression, most of them individually may not be deemed statistically significant. This is exactly what we have observed while analyzing gene expression: functional characterization of differentially expressed genes is highly dependent on the choice of the p-value threshold (Fig. 2).

Unlike gene expression, network entropy rates of individual nodes depend on the collective properties of the entire network. At the same time, random walk process that is used to compute entropy rates tends to prioritize vertices that are central to the subsets of nodes, while also being highly specific to these subsets. Such prioritization is analogous to Google's PageRank algorithm [35] which prioritizes web pages that are highly relevant to specific search topics, while downgrading sites with a high number of non-specific web links. Cell cycle lies at the center of the complex network of proliferative signaling, which contains multiple paths converging on its regulation. The entropy rates of the cell cycle proteins should be highly sensitive to the topology and edge weights of this network module.

Positive correlation between entropy rate of the cell cycle proteins and IC₅₀ values for drug response observed in our study implies that in drug-resistant cells the network of proliferative signaling is denser with the higher number and/or weights of the edges. While some studies attribute resistance to MEK inhibitors to the presence of specific alternative signaling paths leading to re-activation of the cell cycle [36], our analysis suggests that this resistance is due to the overall resilience of the entire proliferative signaling network. Application of the drug alters signaling properties of one of the nodes, effectively generating a network perturbation. This perturbation can be “dissipated” by adjusting signaling via other edges and nodes. In cells where the network of proliferative signaling is densely connected there are many options for such adjustment, therefore dissipation is more effective, leading to drug resistance. Conversely, the cells in which this network is not as dense have fewer options to compensate for the perturbation and are more sensitive to the drug as a result. This implies that targeting individual proliferative signaling mechanisms in cancer cells with high network entropy of the cell cycle may not be the best strategy for overcoming drug resistance – a highly connected, robust network can always dissipate additional perturbations. Instead, downstream effector, in this case the cell cycle itself should be the focus of combination therapy. This is confirmed by the observed synergy between MEK and CDK4/6 inhibition in cells with high network entropy rate of the G2/M transition pathway.

Small differences in the levels of expression of genes in the proliferative signaling network module result in consistent differences in the entropy rates of its core effectors – cell cycle proteins. By comparing entropy metrics of these nodes across samples we effectively compare the state of the proliferative signaling network module across

different cell lines. In this sense entropy rates should be viewed as “mesoscopic” variables, reflecting collective properties of parts of the global PPI network to which they have specific connectivity. Functionally related nodes have specific connectivity with the same or very similar network neighborhood(s). Therefore, in the space of network entropy metrics representation of the functional differences becomes robust, as seen from our results. This property of entropy metrics makes them attractive candidates to use as features in machine learning models for classification of biological samples and prediction of clinical end-points. It has been previously shown that using “pathway scores” or “network module scores” instead of expression values or mutation status of individual genes can improve performance of such algorithms [37–39]. Currently most of such features are constructed by using *ad hoc* empirical procedures such as enrichment, mutation load, or average expression rank of genes in each pathway or module. In contrast, construction of network entropy metrics relies directly on theoretical considerations related to the stability of network states.

Acknowledgments. This work was supported in part by the NIH Cancer Center Support Grant to the Rogel Cancer Center at the University of Michigan (P30 CA046592-29).

References

1. Zeitouni, D., Pylayeva-Gupta, Y., Der, C., Bryant, K.: KRAS mutant pancreatic cancer: no lone path to an effective treatment. *Cancers (Basel)* **8**(4), 45 (2016)
2. Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**(7407), 330–337 (2012)
3. Ferrer, I., Zugazagoitia, J., Herbertz, S., John, W., Paz-Ares, L., Schmid-Bindert, G.: KRAS-mutant non-small cell lung cancer: from biology to therapy. *Lung Cancer* **124**, 53–64 (2018)
4. Baines, A., Xu, D., Der, C.: Inhibition of Ras for cancer treatment: the search continues. *Future Med. Chem.* **3**(14), 1787–1808 (2011)
5. Migliardi, G., Sassi, F., Torti, D., et al.: Inhibition of MEK and PI3 K/mTOR suppresses tumor growth but does not cause tumor regression in patient-derived xenografts of RAS-mutant colorectal carcinomas. *Clin. Cancer Res.* **18**(9), 2515–2525 (2012)
6. Infante, J., Fecher, L., Falchook, G., et al.: Safety, pharmacokinetic, pharmacodynamic, and efficacy data for the oral MEK inhibitor trametinib: a phase 1 dose-escalation trial. *Lancet Oncol.* **13**(8), 773–781 (2012)
7. Lee, M., Helms, T., Feng, N., et al.: Efficacy of the combination of MEK and CDK4/6 inhibitors in vitro and in vivo in KRAS mutant colorectal cancer models. *Oncotarget* **7**(26), 39595–39608 (2016)
8. Ziemke, E., Dosch, J., Maust, J., Shettigar, A., Sen, A., Welling, T., Hardiman, K., Sebolt-Leopold, J.: Sensitivity of KRAS-mutant colorectal cancers to combination therapy that cotargets MEK and CDK4/6. *Clin. Cancer Res.* **22**(2), 405–414 (2016)
9. Pek, M., Yatim, S., Chen, Y., Li, J., Gong, M., Jiang, X., Zhang, F., Zheng, J., Wu, X., Yu, Q.: Oncogenic KRAS-associated gene signature defines co-targeting of CDK4/6 and MEK as a viable therapeutic strategy in colorectal cancer. *Oncogene* **36**(35), 4975–4986 (2017)
10. Zhou, J., Zhang, S., Chen, X., Zheng, X., Yao, Y., Lu, G., Zhou, J.: Palbociclib, a selective CDK4/6 inhibitor, enhances the effect of selumetinib in RAS-driven non-small cell lung cancer. *Cancer Lett.* **408**, 130–137 (2017)

11. Remacle, F., Levine, R.: Statistical thermodynamics of transcription profiles in normal development and tumorigeneses in cohorts of patients. *Eur. Biophys. J.* **244**(8), 709–726 (2015)
12. Rietman, E., Platig, J., Tuszyński, J., Lakka Klement, G.: Thermodynamic measures of cancer: gibbs free energy and entropy of protein-protein interactions. *J. Biol. Phys.* **42**(3), 339–350 (2016)
13. Rietman, E., Scott, J., Tuszyński, J., Klement, G.: Personalized anticancer therapy selection using molecular landscape topology and thermodynamics. *Oncotarget* **8**(12), 18735–18745 (2017)
14. Atkins, P., De Paula, J.: Atkins' Physical Chemistry, 6th edn. Oxford University Press, Oxford (2006)
15. Dueck, H., Khaladkar, M., Kim, T., et al.: Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. *Genome Biol.* **16**(1), 122 (2015)
16. Eberwine, J., Kim, J.: Cellular deconstruction: finding meaning in individual cell variation. *Trends Cell Biol.* **25**(10), 569–578 (2015)
17. Marinov, G., Williams, B., McCue, K., et al.: From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res.* **24**(3), 496–510 (2014)
18. Shalek, A., Satija, R., Adiconis, X., et al.: Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**(7453), 236–240 (2013)
19. Davies, P., Demetrius, L., Tuszyński, J.: Cancer as a dynamical phase transition. *Theor. Biol. Med. Model.* **8**, 30 (2011)
20. Manke, T., Demetrius, L., Vingron, M.: An entropic characterization of protein interaction networks and cellular robustness. *J. R. Soc. Interface* **3**(11), 843–850 (2006)
21. Manke, T., Demetrius, L., Vingron, M.: Lethality and entropy of protein interaction networks. *Genome Inform.* **16**(1), 159–163 (2005)
22. Teschendorff, A., Sollich, P., Kuehn, R.: Signalling entropy: a novel network-theoretical framework for systems analysis and interpretation of functional omic data. *Methods* **67**(3), 282–293 (2014)
23. Teschendorff, A., Banerji, C., Severini, S., Kuehn, R., Sollich, P.: Increased signaling entropy in cancer requires the scale-free property of protein interaction networks. *Sci. Rep.* **5**, 9646 (2015)
24. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A., et al.: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012)
25. Keshava Prasad, T., Goel, R., Kandasamy, K., et al.: Human protein reference database—2009 update. *Nucl. Acids Res.* **37**, D767–D772 (2009)
26. Cancer Cell Line Encyclopedia. <https://portals.broadinstitute.org/cclle/home>
27. Teschendorff, A., Severini, S.: Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst. Biol.* **4**, 104 (2010)
28. Yu, G., He, Q.: ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.* **12**(2), 477–479 (2016)
29. LoRusso, P., Krishnamurthi, S., Rinehart, J., et al.: Phase I pharmacokinetic and pharmacodynamic study of the oral MAPK/ERK kinase inhibitor PD-0325901 in patients with advanced cancers. *Clin. Cancer Res.* **16**(6), 1924–1937 (2010)
30. Haura, E., Ricart, A., Larson, T., et al.: A phase II study of PD-0325901, an oral MEK inhibitor, in previously treated patients with advanced non-small cell lung cancer. *Clin. Cancer Res.* **16**(8), 2450–2457 (2010)
31. Vidal, M., Cusick, M., Barabási, A.: Interactome networks and human disease. *Cell* **144**(6), 986–998 (2011)

32. Tényi, Á., Cano, I., Marabita, F., et al.: Network modules uncover mechanisms of skeletal muscle dysfunction in COPD patients. *J. Transl. Med.* **16**(1), 34 (2018)
33. Yue, Z., Arora, I., Zhang, E., Laufer, V., Bridges, S., Chen, J.: Repositioning drugs by targeting network modules: a Parkinson's disease case study. *BMC Bioinform.* **18**(Suppl. 14), 532 (2017)
34. Zhang, S.: Comparisons of gene coexpression network modules in breast cancer and ovarian cancer. *BMC Syst. Biol.* **12**(Suppl 1), 8 (2018)
35. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **30**(1–7), 107–117 (1998)
36. Wee, S., Jagani, Z., Xiang, K., Loo, A., Dorsch, M., Yao, Y., Sellers, W., Lengauer, C., Stegmeier, F.: PI3K pathway activation mediates resistance to MEK inhibitors in KRAS mutant cancers. *Cancer Res.* **69**(10), 4286–4293 (2009)
37. Xu, Y., Dong, Q., Li, F., et al.: Identifying subpathway signatures for individualized anticancer drug response by integrating multi-omics data. *J. Transl. Med.* **17**(1), 255 (2019)
38. Wang, X., Sun, Z., Zimmermann, M., Bugrim, A., Kocher, J.: Predict drug sensitivity of cancer cells with pathway activity inference. *BMC Med. Genomics* **12**(Suppl. 1), 15 (2019)
39. Strunz, S., Wolkenhauer, O., de la Fuente, A.: Network-assisted disease classification and biomarker discovery. *Methods Mol. Biol.* **1386**, 353–374 (2016)



Computational Modelling of TNF α Pathway in Parkinson's Disease – A Systemic Perspective

Hemalatha Sasidharakurup, Lakshmi Nair, Kanishka Bhaskar,
and Shyam Diwakar^(✉)

Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham,
Amritapuri Campus, Clappana P.O., Kollam 690525, India
shyam@amrita.edu

Abstract. The paper aims developing a computational framework of signaling using the principles of biochemical systems theory as a model for Parkinson's disease. Several molecular interactions aided by TNF α , a proinflammatory cytokine play key roles in mediating glutamate excitotoxicity and neuroinflammation, resulting in neuronal cell death. In this paper, initial concentrations and rate constants were extracted from literature and simulations developed were based on systems of ordinary differential equations following first-order kinetics. In control or healthy conditions, a decrease in TNF α and neuronal cell death was predicted in simulations matching data from experiments, whereas in diseased condition, a drastic increase in levels of TNF α , glutamate, TNFR1 and ROS were observed similar to experimental data correlating diseased condition to augmented neuronal cell death. The study suggests toxic effects induced by TNF α in the substantia nigra may be attributed to Parkinson's disease conditions.

Keywords: Computational modelling · Parkinson's disease · TNF α · Glutamate · Neuroinflammation

1 Introduction

A systemic approach to the perception of human diseases can widen the horizons for novel drug discovery, elucidation of complex molecular networks of pathways involved in disease pathogenesis and provide a breeding ground for clinical treatment optimization involving computational modelling [1]. Parkinson's disease is considered one of the most common disorder afflicting about 6.1 million people with an annual cost of ₹10,000 per person in India solely for drugs [2, 3]. The key pathological observation of Parkinson's disease is the progressive loss of dopaminergic neurons from the substantia nigra pars compacta of the midbrain together identified with the presence of intraneuronal inclusions known as Lewy bodies [4]. From a systems theory

perspective, brain disorders involve molecular and signalling events relating to cellular, tissue and organ level functions. Given the complexity of molecular interactions in relating biological data to predictions, the potential application for a systems-modelling study pertaining to Parkinson's Disease (PD) is also to identify or design neuroprotective drugs.

Systems models have enabled reconstruction of PD associated complex molecular and genetic networks and may predict how alterations in these networks are organized into various physiological functions and pathological conditions [5]. A single cell network consists of hundreds of nodes that represent genes, proteins, other biomolecules and their respective kinetic reactions of multiple biological pathways involved in PD. By identifying disease-related pathways, it has been possible to predict genetic risk factors and novel biomarkers for PD. This systematic approach for interpreting disease mechanisms also helps building prediction models to analyse how these biochemical interactions change over the course of time and also under varying conditions [6].

Previous studies have been reported that TNF α signalling including glutamate excitotoxic pathway has an important role in loss of dopaminergic neurons in the substantia nigra pars compacta [7, 8]. TNF α can be focused as a biomarker for PD since several studies have been reported that anti TNF α therapy reduces the risk of developing PD [9, 10]. TNF α is also involved in rapid necrosis of tumours [11] and it is a pro-inflammatory cytokine involved in the innate immune response [12]. Recent studies have reported that neuroinflammation caused progressive degeneration of dopaminergic neurons with increased levels of TNF α [13] and increased serum levels of TNF α were also found in patients afflicted with PD [14]. Another study have shown that glutamate toxicity could result in death of dopaminergic neurons [15]. In this study, a computational model of mapping TNF α signalling in PD have been developed to understand how it mediates glutamate excitotoxicity and neuroinflammation leading to disease pathogenesis and progression. The model was developed using ODE (ordinary differential equation) based on biochemical systems theory (BST). The pathway information was extracted from literature survey with initial concentrations values and rate constants in both normal and diseased state from experimental techniques such as ion exchange chromatography, enzyme immunoassay, IMR assay etc. The concept of law of mass action has been applied to produce a time versus concentration plot to analyse how each species involved in the above mentioned pathway of TNF α signalling could lead to disease severity in case of PD on comparison with the control condition. This provides predictions on the early onset of PD from cellular level that can help in developing new therapeutics for PD.

2 TNF α Signalling Pathways in Parkinson's Disease

2.1 TNF α Signalling Linked Glutamate Cytotoxicity of Dopaminergic Neurons

At normal physiological conditions (in the absence of misfolded α -synuclein), infiltration of T cells and action of microglia results in more TNF α production and glutamate exocytosis were blocked thereby not creating a vicious cycle of TNF α production. In disease pathways, TNF α initiates glutamate mediated excitotoxicity which is induced by a series of events by T cells, microglia, astrocytes, and dying neurons under stress. During the neuroinflammatory process, infiltrated T cells release the cytokine IFN- γ which induces in microglia an increased TNF α production and release [16]. Microglia then phagocytoses the misfolded α -synuclein and subsequent production of TNF α from microglia by maintaining its concentration high [10]. TNF α then acts on the TNFR1 receptor on microglia leading to its own increased production [17]. Excess glutamate again binds on to mGluR2 receptors which in turn increase TNF α production [18]. The excess glutamate generated acts on NMDAR – a glutamate receptor on the dopaminergic neurons which results in an increase in calcium influx through activation of JNK signalling path, initiates the process of cell death [19]. As a result of excess calcium influx, calcium dependent proteases such as calpains are activated which in turn trigger Lewy body formation resulting in cell death [19]. Activated calpain result in the formation of cleaved Cdk5 which once again trigger death of dopaminergic neurons [20]. ATP released from nearby dying neurons are exocytosed by secretory vesicles and bind to P2X7R receptor on microglia with the help of proteins such as MEK, ERK and p38 again result in the production of TNF α [21, 22] (See Fig. 1).

2.2 TNF α Mediated Dopaminergic Cell Death via Glial Pathway in PD

Under normal physiological function of the CNS, microglia which is the major source of TNF α during neuroinflammation is inhibited by neurons and astrocytes [23]. Astrocytes releases TGF β and IL-10 which will inhibit microglial activation [24]. TNF α released by astrocytes and microglia cells bind to TNFR1 receptor in the dopaminergic neurons and results in the activation of caspase 3, 8, and cytochrome c and results in cell death [25]. Cytosolic cytochrome c complexed with Apaf-1 and subsequently triggers the sequential activation of caspase 9 & 3 and leads to cell death [26]. Excessive uptake of neuronal α -synuclein by astrocytes can lead to accumulation of protein aggregation in astrocytes, and cause an upregulation of IL-1 α , IL-1 β and IL-6, followed by the release of TNF α and IL-6 [27]. TNF α activates the NF- κ B signaling pathway leading to the production of a large variety of pro-inflammatory genes, including TNF α , IL-1 β , and NO [28]. ROS production could activate p38 MAPK pathway which ultimately leads to cell death [29] (See Fig. 2).

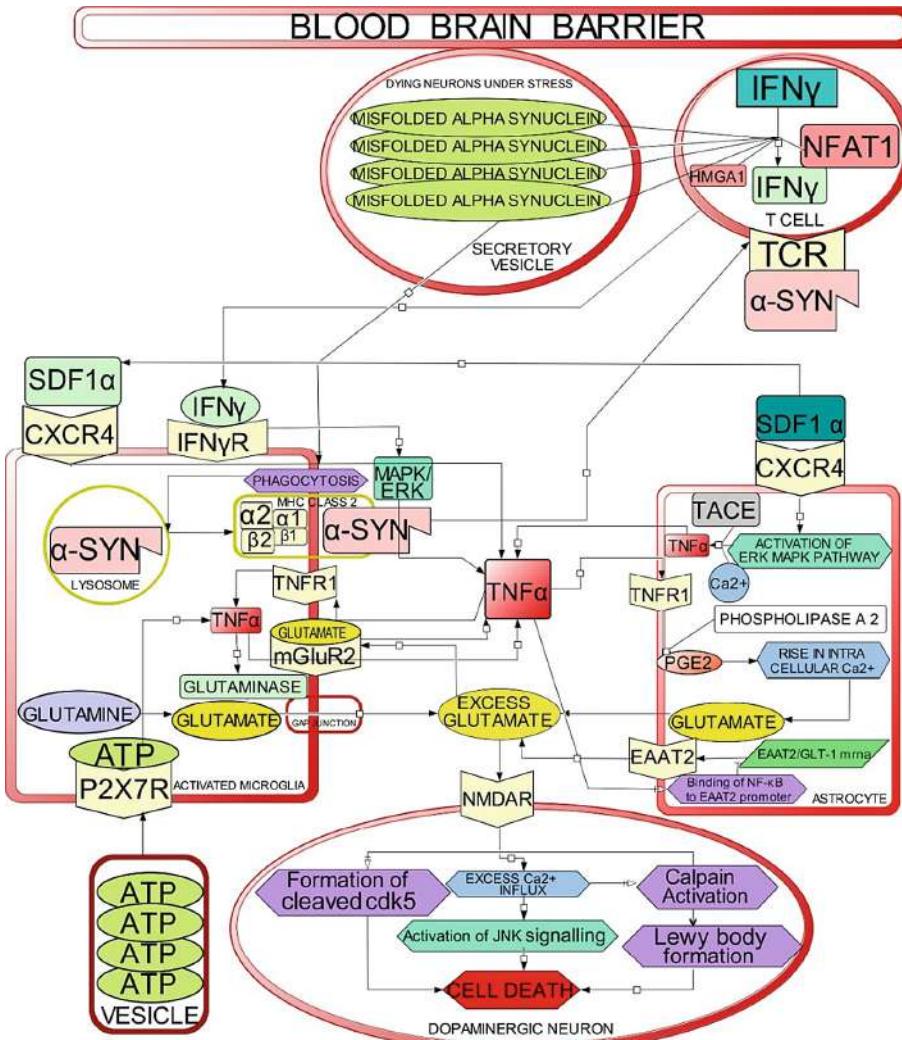


Fig. 1. TNF- α induced glutamate excitotoxic pathway in disease state

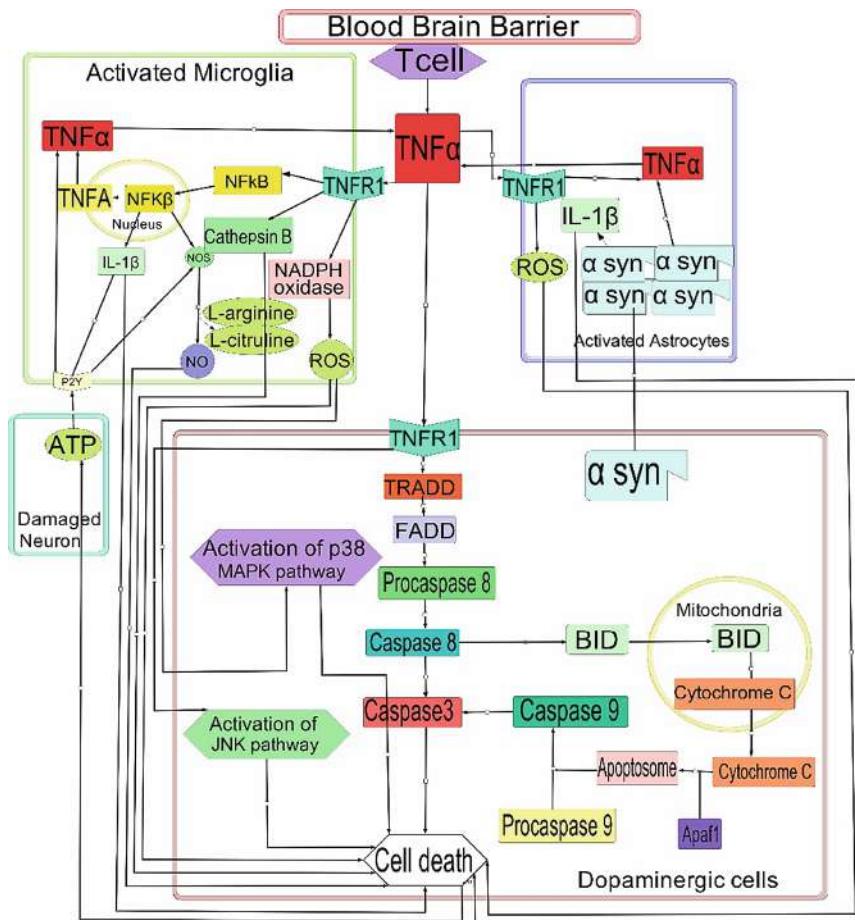


Fig. 2. TNF- α pathway during neuroinflammation in diseased conditions

3 Methods

The biochemical networks of TNF α signalling, especially in the substantia nigra pars compacta of the midbrain, have been modelled by using BST [30]. Experimental data to model both normal and diseased conditions were extracted from literature. Concentration values of both diseased and normal condition and their rate constants were extracted from the previous experimental studies (See Table 1). All the single reactions in the pathway were then converted into mathematical equations using BST to study the behavior of the system (See main relationships in Table 2).

Table 1. Fitting models to experimental data

Name	Experiment	Region	Model (in vitro)	Concentration		References
				Control	Diseased	
TNF α	EIA assay	CSF	Humans	22.3 \pm 9.5 pg/ml	96.3 \pm 9.1 pg/ml	Mogi et al. 1993
Glutamate	IEX chromatography	Blood plasma	Humans	34.1 \pm 11.3 μ mol/L	71.7 \pm 8.5 μ mol/L	Yasuo et al. 1992
Ca $^{2+}$	Fluorometric analysis	SNpc	Sprague-Dawley rats	0.080 μ M	100 μ M	Yumi et al. 2002
TNFR1	ELISA	Blood serum	Humans	438.9 \pm 171.9 pg/ml	558.5 \pm 246.3 pg/ml	Paula et al. 2009
ROS	Fluorescent assay	Brain Cortex	Guinea Pig	12.7 \pm 0.17 pmol/min/mg	32.0 \pm 2.2 pmol/min/mg	Tretter et al. 2004
α -synuclein	ELISA	Blood	Humans	68.19–645.57 pg/ml	55.2–1294.9 pg/ml	Oczkowska 2014

Table 2. Concentration of keynote species from published literature

Reaction	Equation	Parameters	Reference
TNF α production via ERK MAPK pathway	$dx(1) = s56 * s59 * s58 * k1$	s56-Activated ERK MAPK pathway s59-TACE s58-Ca $^{2+}$ k1-Rate constant	Bezziet et al. [18]
Glutamate production-enzyme glutaminase	$dx(2) = s43 * s45 * k2$	s43-glutamine s45-glutaminase k2-Rate constant	Takeuchi et al. [17]
Lewy body mediated cell death	$dx(3) = s89 * k3$	s89-Lewy body k3-Rate constant	Zheng et al. 2010
Calcium production	$dx(4) = [s72 * k4] - [s73 * k5]$	s72-NMDAR s73-Ca $^{2+}$ influx k4-Rate constant k5-Rate constant	Izumi et al. [15]
TNFR1 activation by TNF α	$dx(5) = s2 * k6$	s2 = TNF- α k6 = rate constant	Wang et al. 2015
α -synuclein aggregation	$dx(6) = s14 * k7$	s14 = α -synuclein k7 = rate constant	Hirsch et al. 2003
ROS triggered cell death	$dx(7) = s7 * k8$	s7 = ROS k8 = rate constant	Miller et al. 2009

Generalized Mass Action Kinetics

$$\dot{x}_i = \sum_{j=1}^{N_i} a_{ij} \prod_{k=1}^d x_k^{g_{ijk}} \quad (1)$$

with $i = 1, \dots, d$. Here, x_i represents the concentration of a reactant and \dot{x}_i represent the time derivative of x_i . The parameters a_{ij} and g_{ijk} represents the rate constants and kinetic orders respectively.

Michaelis-Menten Equation. Rate of an equation v_p can be given as

$$v_p = \frac{V_{\max} S}{K_M + S} \quad (2)$$

Here, V_{\max} represents Maximal velocity, K_M denotes Michealis constant and S represents concentration of the substrate. Michaelis Constant, K_M is the substrate concentration at which the rate is half of the maximal velocity. For an example, mathematical representation of TNFR1 receptor activation by TNF- α is denoted as $dx/(1) = s_2 * k_1$, where s_2 is TNF- α and k_1 is rate constant.

4 Results

4.1 TNF α -Linked Glutamate Excitotoxic Pathway

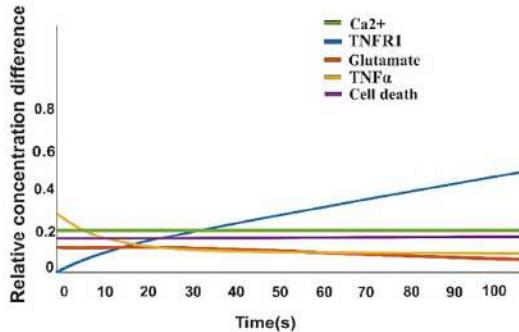
Results from control conditions showed a constant level of calcium, followed by slightly increased level of TNF α in the initial stage and then it maintained a constant level. Glutamate also maintained at a constant level. During normal physiological condition, astrocytes and neurons release certain factors which will inhibit the microglial activation, which is the major source of TNF- α . This results in the decreased level of TNF- α and its receptor in brain. It has been observed that the level of TNF α was high in diseased condition. On comparing cell death in both control and disease, an increase in dopaminergic neuronal cell death have been observed from the analysis of diseased state (See Fig. 3a, b).

4.2 TNF α -Mediated Dopaminergic Cell Death via Glial Pathway in PD

During control condition, it was observed that the level of TNF- α and its surface receptor, TNFR1 is considerably low when compared to the diseased state. The cell death rate is also decreased when compared to the diseased state (Fig. 4a). During the diseased state, there is an elevation of TNF- α and TNFR1. Activation of microglial

cells and astrocytes results in the increased production of TNF- α and other pro-inflammatory cytokines in the PD brain. Results also have shown a high concentration level of ROS. The relative amount of α -synuclein and cell death have been also increased when compared to disease condition (Fig. 4a, b).

3a.



3b.

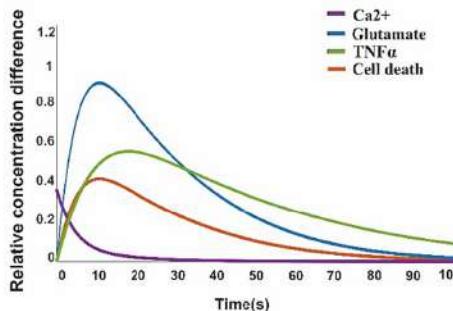
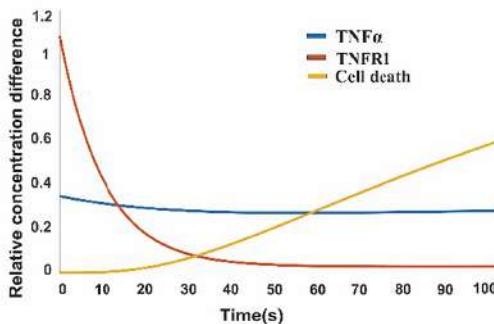


Fig. 3. TNF α signaling including glutamate cytotoxicity (a) Decrease in TNF α levels with a constant level of glutamate and calcium release in control condition. (b) Increased level of TNF α resulting in excess glutamate mediated dopaminergic neuronal cell death in disease state.

5 Discussion

The objective of this study was to develop a computational model of TNF α signalling including pathways such as glutamate excitotoxicity and neuroinflammation related to PD. The biochemical pathways of both normal and diseased condition have been modelled using BST and rate equations. In normal condition, the results have shown an increased level of TNF α in the initial stage and then it maintained a constant level compared to diseased condition. The prediction from results supports several experimental data reporting that under normal physiological state, during the absence of misfolded α -synuclein, T-cell infiltration is blocked and the production of TNF α is comparatively low [31, 32]. Since there is no release of glutamate from microglia due to high levels of TNF α , cell death via glutamate excitotoxicity is prevented in normal condition. In disease state, the concentration level of TNF α was high compared to normal. This is due to the presence of misfolded α -synuclein. As a result of number of autocrine loops and microglial activation, release of glutamate from both astrocyte and microglia lead to glutamate mediated cytotoxicity of dopaminergic neurons. Misfolded α -synuclein, the key pathological condition in PD keeps the T cells in an activated state thus resulting in more TNF α production and subsequently more glutamate production potentiating glutamate toxicity. In the case of TNF α induced neuroinflammation, the results have shown in normal condition that levels of TNF α and TNFR1 is considerably low when compared to the diseased state. Under normal physiological function of the CNS, microglia which is the major source of TNF α during neuroinflammation is inhibited by neurons and astrocytes c-FLIP protein can inhibit apoptosis by competitively dimerizing with caspase 8 and inhibiting its activation resulting in less cell death. In disease condition, simulations predicts elevation in the concentration of TNF α together with its receptor TNFR1. Simulations also have shown that ROS is increased in the diseased condition. An increase in the level of α -synuclein have been also observed. The relative cell death in diseased state have been observed high when compared to normal state. Several studies have reported that TNF- α activated microglial cells and astrocytes can produce a variety of noxious compounds, including ROS and also proinflammatory cytokines like TNF α , which results in the increase in the level of TNF- α in the diseased state [33]. TNF α results in the neuroinflammation plays an important role in PD, by directly or indirectly contributing to the degeneration of dopaminergic neurons.

4a.



4b.

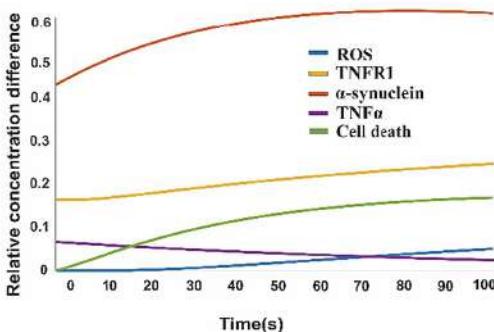


Fig. 4. TNF α -mediated cell death via glial pathway. (a) Decreased level of TNF- α and TNFR1 receptor in the normal condition. (b) Elevated level of TNF- α and TNFR1 receptor in the pathogenesis of PD.

6 Conclusion

The modelled biochemical networks helps to analyse the potential role of TNF- α in dopaminergic neurodegeneration leading to disease pathogenesis and subsequently to its progression. This model can be used to find different therapeutic strategies and targets that could be aimed in bringing a novel biomarker or an all-time cure for such diseases.

Acknowledgements. This work derives direction and ideas from the Chancellor of Amrita University, Sri Mata Amritanandamayi Devi. This work was partially funded by Department of Science and Technology Grant DST/CSRI/2017/31, Government of India and by Embracing the World Research-for-a-Cause initiative.

References

- Ji, Z., Yan, K., Li, W., Hu, H., Zhu, X.: Mathematical and computational modeling in complex biological systems. *Biomed. Res. Int.* **2017**, 1–16 (2017)
- Ray Dorsey, E., et al.: Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **17**, 939–953 (2018)
- Ragothaman, M., Govindappa, S.T., Rattihalli, R., Subbakrishna, D.K., Muthane, U.B.: Direct costs of managing Parkinson's disease in india: concerns in a developing country. *Mov. Disord.* **21**, 1755–1758 (2006)
- Perez, R.G., et al.: A role for α -synuclein in the regulation of dopamine biosynthesis. *J. Neurosci.* **22**, 3090–3099 (2002)
- Sasidharakurup, H., Melethadathil, N., Nair, B., Diwakar, S.: A systems model of Parkinson's disease using biochemical systems theory. *Omics J. Integr. Biol.* **21**, 454–464 (2017)
- Stayte, S., Vissel, B.: Advances in non-dopaminergic treatments for Parkinson's disease. *Front. Neurosci.* **8**, 113 (2014)
- Lindenau, J.D., Altmann, V., Schumacher-Schuh, A.F., Rieder, C.R., Hutz, M.H.: Tumor necrosis factor alpha polymorphisms are associated with Parkinson's disease age at onset. *Neurosci. Lett.* **658**, 133–136 (2017)
- Nagatsu, T., Sawada, M.: Biochemistry of postmortem brains in Parkinson's disease: historical overview and future prospects. *J. Neural Transm. Suppl.* 113–120 (2007). <https://doi.org/10.1007/978-3-211-73574-9-14>
- Peter, I., et al.: Anti-tumor necrosis factor therapy and incidence of Parkinson disease among patients with inflammatory bowel disease. *JAMA Neurol.* **75**, 939 (2018)
- Harms, A.S., et al.: Delayed dominant-negative TNF gene therapy halts progressive loss of nigral dopaminergic neurons in a rat model of Parkinson's disease. *Mol. Ther.* **19**, 46–52 (2011)
- Wood, L.B., Winslow, A.R., Strasser, S.D., Wood, L., Israel, B.: Systems biology of neurodegenerative diseases graphical abstract HHS public access. *Integr. Biol. (Camb.)* **7**, 758–775 (2015)
- Clark, I.: How TNF was recognized as a key mechanism of disease. *Cytokine Growth Factor Rev.* **18**, 335–343 (2007)
- Parent, A., Parent, M., Charara, A.: Glutamatergic inputs to midbrain dopaminergic neurons in primates. *Parkinsonism Relat. Disord.* **5**, 193–201 (1999)
- Kouchaki, E., et al.: Increased serum levels of TNF- α and decreased serum levels of IL-27 in patients with Parkinson disease and their correlation with disease severity. *Clin. Neurol. Neurosurg.* **166**, 76–79 (2018)
- Izumi, Y., et al.: Vulnerability to glutamate toxicity of dopaminergic neurons is dependent on endogenous dopamine and MAPK activation. *J. Neurochem.* **110**, 745–755 (2009)
- Hanisch, U.-K.: Microglia as a source and target of cytokines. *Glia* **40**, 140–155 (2002)
- Takeuchi, H., et al.: Tumor necrosis factor- α induces neurotoxicity via glutamate release from hemichannels of activated microglia in an autocrine manner. *J. Biol. Chem.* **281**, 21362–21368 (2006)
- Bezzi, P., et al.: CXCR18-activated astrocyte glutamate release via TNF α : amplification by microglia triggers neurotoxicity. *Nat. Neurosci.* **4**, 702–710 (2001)
- Dufy, B.M., et al.: Calpain-cleavage of α -synuclein: connecting proteolytic processing to disease-linked aggregation. *Am. J. Pathol.* **170**, 1725–1738 (2007)

20. Smith, P.D., et al.: Cyclin-dependent kinase 5 is a mediator of dopaminergic neuron loss in a mouse model of Parkinson's disease. *Proc. Natl. Acad. Sci.* **100**, 13650–13655 (2003)
21. Inoue, K.: Microglial activation by purines and pyrimidines. *Glia* **40**, 156–163 (2002)
22. Negro, S., et al.: ATP released by injured neurons activates Schwann cells. *Front. Cell. Neurosci.* **10**, 134 (2016)
23. Welser-Alves, J.V., Milner, R.: Microglia are the major source of TNF- α and TGF- β 1 in postnatal glial cultures; regulation by cytokines, lipopolysaccharide, and vitronectin. *Neurochem. Int.* **63**, 47–53 (2013)
24. Vincent, V.A.M., Tilders, F.J.H., Dam, A.V.A.N.: Inhibition of endotoxin-induced nitric oxide synthase production in microglial cells by the presence of astroglial cells: a role for transforming growth factor β . *Glia* **198**, 190–198 (1997)
25. Cabezas, R., et al.: Astrocytes role in Parkinson: a double-edged sword (2013). <https://doi.org/10.5772/54305>
26. Junn, E., Mouradian, M.M.: Apoptotic signaling in dopamine-induced cell death: the role of oxidative stress, p38 mitogen-activated protein kinase, cytochrome c and caspases. *J. Neurochem.* **78**, 374–383 (2001)
27. Hirsch, E.C.: Glial cells and Parkinson's disease. *J. Neurol.* **247**, 58–62 (2000)
28. Qian, L., Flood, P.M.: Microglial cells and Parkinson's disease. *Immunol. Res.* **41**, 155–164 (2008)
29. He, J., Zhong, W., Zhang, M., Zhang, R., Hu, W.: P38 mitogen-activated protein kinase and Parkinson's disease. *Transl. Neurosci.* **9**, 147 (2018)
30. Knorre, W.: M. A. Savageau, Biochemical Systems Analysis. A Study of Function and Design in Molecular Biology. 396 S., 115 Abb., 14 Tab. Reading, Mass. 1976. Addison-Wesley Pbl. Co./Advanced Book Program. £ 26,50. Z. Allg. Mikrobiol. **19**, 149–150 (2007)
31. Olmos, G., Lladó, J.: Tumor necrosis factor alpha: a link between neuroinflammation and excitotoxicity. *Mediat. Inflamm.* **2014** (2014)
32. Leal, M.C., Casabona, J.C., Puntel, M., Pitossi, F.: Interleukin-1 β and TNF- α : reliable targets for protective therapies in Parkinson's disease? *Front. Cell. Neurosci.* **7**, 53 (2013)
33. Fischer, R., Maier, O.: Interrelation of oxidative stress and inflammation in neurodegenerative disease: role of TNF. *Oxidative Med. Cell. Longev.* **2015**, 1–18 (2015)



Understanding the Progression of Congestive Heart Failure of Type 2 Diabetes Patient Using Disease Network and Hospital Claim Data

Md Ekramul Hossain^(✉), Arif Khan, and Shahadat Uddin

Complex Systems Research Group, Faculty of Engineering,
The University of Sydney, Sydney, NSW, Australia

{ekramul.hossain, arif.khan,
shahadat.uddin}@sydney.edu.au

Abstract. Chronic diseases have increasingly become common and caused most of the burden of ill health in most countries. They have large impacts on quality of life, social and economic conditions. These diseases bring several health risks to those patients suffering from more than one chronic disease at one time (also known as comorbidity of chronic disease). Due to this, governments and healthcare service providers are concerned about the burden of comorbidity of chronic diseases. Understanding the progression of comorbidities can provide vital information for the prevention and better management of chronic diseases. The routinely collected hospital claim data contain semantic information about patients' health in the form of disease codes. Therefore, these data can be used to understand the progression of chronic disease comorbidities. Most studies in this field are focused on understanding the progression of one chronic disease rather than multiple chronic diseases. In this study, we aim to understand the progression of multiple chronic diseases, i.e., comorbidities that occur when patients of one chronic condition progress towards another. Based on the prevalence of chronic diseases within the Australian population, we have particularly focused on the comorbidity progression of congestive heart failure (CHF) for type 2 diabetes (T2D) patients. In this study, we propose a research framework to understand and represent the progression of CHF in patients with T2D using graph theory and social network analysis. We used hospital claim data drawn from the Australian healthcare context. We constructed two baseline disease networks from two cohorts (i.e., patients with both T2D and CHF and patients with only T2D). A final weighted disease network from two cohorts was then generated by giving more weights to the prevalent comorbidities in patients with T2D and CHF compared to the patients with only T2D. The results show that chronic pulmonary disease, cardiac arrhythmias, valvular disease and renal failure occurred frequently during the progression of CHF for T2D patients. In addition, the final disease network shows the highest transition between electrolyte disorders and renal failure. This indicates that these two diseases may be potential risk factors for the progression towards CHF in patients with T2D for this population cohort. Thus, the proposed network representation can help the healthcare provider to understand high-risk diseases and progression pattern between recurrence of T2D and CHF. Also, it can help in the efficient

management of healthcare resources. The proposed framework could be useful for stakeholders including governments and health insurers to adopt appropriate preventive health management program for the patients at high risk of developing multiple chronic diseases.

Keywords: Chronic disease · Comorbidity · Hospital claim data · Graph theory · Social network analysis

1 Introduction

Chronic diseases, such as diabetes, Alzheimer's disease, chronic obstructive pulmonary disease, and congestive heart failure, usually progress slowly over time and they are generally not cured completely [1]. The prevalence of chronic diseases has been growing over time. For example, about half of all Australians have a chronic disease, and around 20% have at least two chronic diseases according to data released by the Australian Institute of Health and Welfare (AIHW) [2]. 'Comorbidity' is a related term in this context which refers to the co-occurrence of different diseases, generally complex and often chronic, in the same patient [3]. Sometimes comorbidity of chronic diseases can occur simply by chance, but often they share common risk factors between them. For example, ageing is a risk factor for developing chronic disease comorbidity. However, the prevalence of congestive heart failure (CHF) increases with age and comorbidities such as obesity, chronic pulmonary disease, hypertension, and type 2 diabetes (T2D) [4]. In clinical practice, T2D and CHF are common companions. The patients with T2D have over twice the risk of occurrence CHF than people without diabetes [5, 6]. This is partly because of common risk factors between CHF and T2D, including obesity, advanced age, hypertension, chronic kidney disease, and coronary heart disease. Alongside the projected increase in prevalence, comorbidities of T2D and CHF exert a significant social and health burden, as well as associate with higher healthcare costs. Therefore, effective preventive measures are needed to address the expected increased burden of chronic disease comorbidities. In particular, an area where research attention is needed is in patients with both T2D and CHF as the data show an incremental risk of death and hospitalisation for CHF compared to patients with CHF without T2D [7]. However, conventional methods of conducting studies and regular monitoring of a large population are expensive in terms of available clinical resources as well as economic conditions. An alternative data source collected from hospital admission and discharge data is available that include patients' health information in the form of standardised ICD (International Classification of Diseases) codes [8]. Analysis of this hospital claim data using data mining and social network analysis can help us to understand the comorbidity of multiple chronic diseases.

In this study, a research framework is proposed to understand the progression of CHF in patients with T2D using a disease-based network model. This network model considers the disease history of patients and analyses the pattern of disease progression over a long time. The disease network is treated as 'Baseline Disease Network' that can reveal the comorbidities and progression of one chronic disease to another chronic disease. To implement the model, this study used real hospital claim data in the

Australian healthcare context. The proposed method presented here is an improved and detailed extension of our previous study for understanding the comorbidity of multiple chronic diseases [9].

Significant research has been done in the related field of understanding the progression of chronic disease comorbidities. Rule-based scoring is one of the earliest models which is based on the empirical and clinical understanding of symptoms and diseases comorbidities [10, 11]. In this model, the severity of a patient is determined by setting scores to socio-demographic and geographic risk factors, physiologically observable syndromes, and presence of comorbidities. In 1987, the Charlson Comorbidity Index was proposed to predict the 10-year mortality for a patient using rule-based scoring model [10]. Another widely used rule-based scoring model is APACHE-II (Acute Physiology and Chronic Health Evaluation-II) which is used to assess intensive care unit (ICU) patients' health condition in the first 24 h of admission [11]. The results of a group of diagnostic tests are considered as scores that are also used to assess or make prognosis. For example, Ewing and Clarke proposed five tests (known as Ewing's battery test) to assess the risk of cardiovascular disease in diabetes [12]. However, these scoring models are derived from empirical observation and do not test for a large diverse population with multiple comorbidities. Although they work well in the specific healthcare setting. Based on the A1C and clinical characteristics of type 2 diabetes patients, a diabetes-specific equation was proposed to understand the disease progression and estimate the 5-year risk of cardiovascular disease [13].

Nowadays, administrative data have been used in healthcare research and clinical decision making such as treatment, diagnosis, understanding disease progression and predicting disease risks [14, 15]. These data are generated when a patient visits a physician, admits to a hospital, undergoes diagnosis tests and purchases medicines at a pharmacy. In 2001, Nichols and his colleagues proposed a research framework to estimate the prevalence and incidence of CHF in patients with T2D using electronic medical data. They also identified risk factors for diabetes-associated CHF using multiple logistic regression models [16]. Later, they updated their study for estimating the CHF incidence rate in T2D and identifying risk factors for developing CHF in patients with T2D over 6 years of follow-up [17]. Various data mining and machine learning methods used healthcare data in different healthcare research [18, 19]. For instance, collaborative filtering methods were proposed for understanding disease progression and predicting disease risk utilising administrative healthcare data [18, 20]. Bayesian network [21], a combination of graph theory and probability theory, has been used to understand the comorbidity of multiple chronic diseases [22]. A risk prediction model was developed to predict the risk of progression to chronic obstructive pulmonary disease in asthma patients using electronic health data [23]. Here, Bayesian network was used to construct the proposed model.

Coordination of entities at a different level of healthcare organisational structure can be conceptualised, studied and quantified using graph/network analysis or Social Network Analysis (SNA) [24, 25]. In this context, SNA can be defined as a set of entities, such as physicians, diseases and hospitals, with some relationships or interactions between them. Recently, hospital claim data have been widely used to demonstrate SNA based approaches [26, 27]. The main goal of this network-based approach is to understand relations between healthcare entities [24] and improving

collaboration efficiency among physicians [28]. SNA and traditional network analysis techniques were applied on electronic healthcare data of CHF patients to explore the patterns of service delivery for improved care coordination [29]. Recently, Khan et al. [15] developed a research framework to understand the progression of type 2 diabetes using graph theory and SNA. The proposed model was applied to administrative claim data in the Australian healthcare context. The study focused on understanding single chronic disease (e.g., T2D) rather than the progression of multiple chronic diseases. To the best of our knowledge, there is very little work used SNA and hospital claim data to develop a research framework to understand the progression of CHF in patients with T2D, which is the main goal of this study.

The rest of the paper is organised as follows. In Sect. 2, this study introduces the methodology, including the selection of study cohort, ICD code grouping and procedures of the proposed framework. Next, the results and discussions are represented in Sect. 3. Finally, Sect. 4 presents the conclusions with the future research direction.

2 Methods

This section describes the generation process of the baseline disease network, and it includes a brief description on how to select the study cohort as well as ICD codes.

2.1 Data Source

The hospital claim dataset used in this study is collected from a private healthcare fund in Australia to evaluate the proposed framework for understanding the progression of CHF in patients with T2D. It contained medical history for about 124,000 patients who received healthcare services between the year 1995 and 2018. The medical history included coded information about patients (e.g., patient ID, age, sex and location), hospital admission (e.g., admission date, discharge date, episode ID and funding source) and clinical details (e.g., diagnosis and procedure type, ICD codes and DRG codes). The patients' health condition was represented in the form of ICD codes present in the admission data of each admission episode. Both T2D and CHF patient data were considered over the full period of the research dataset. In order to collect the research dataset, a systematic process of filtering was applied to the original hospital claim data. Some of the filtering criteria are – (a) selecting patients having at least one admission with valid ICD codes (b) excluding duplicate records and (c) excluding ICD codes related to physical injuries, fever and vomiting. To explore the risk of CHF for diabetic patients, this study used diagnose-related ICD codes as the primary data item.

2.2 Study Cohort

After pre-processing the research dataset, we focus to define our study cohorts. There are two cohorts selected in this study. One is the cohort of patients who were first diagnosed with T2D and then diagnosed with CHF (referred to as $Cohort_{T2D\&CHF}$). Another is the cohort of patients who were diagnosed with T2D but not diagnosed with CHF at any stage of their entire admission history (referred to as $Cohort_{T2D}$). To select

the cohort, this study looked for the presence of ICD codes of the particular disease in the admission history of the patients. There are well-defined ICD codes for both T2D and CHF shown in Table 1. For $Cohort_{T2D\&CHF}$, we searched for patients who have ICD codes for both T2D and CHF. The search was followed in a way so that it resulted from only those patients who were diagnosed with T2D in an earlier admission and with CHF at any subsequent admission. The patients for $Cohort_{T2D}$ were selected by a search criterion in where the patients have one or more ICD codes for T2D in their entire admissions but no ICD codes for CHF at any stage of their admissions.

Table 1. ICD codes for CHF and T2D.

Disease name	ICD codes (ICD-10-AM)	ICD codes (ICD-9-AM)
CHF	I09.9, I1.0, I13.0, I13.2, I25.5, I42.0, I42.5–I42.9, P29.0	398.91, 402.11, 402.91, 404.11, 404.13, 404.91, 404.93
T2D	E11.0, E11.1, E11.2–E11.9	25000, 25002, 25010, 25012, 25020, 25022, 25030, 25032, 25040, 25042, 25050, 25052, 25060, 25062', 25070', 25072', 25080, 25082, 25090, 25092

2.3 ICD Code Grouping

The dataset used in this study contains the disease codes that are encoded in both ICD-9-AM and ICD-10-AM format. Each version has more than 20,000 unique codes [30], many of them are likely to be present in the dataset. If we consider all codes to construct the baseline disease network, it would be difficult to analyse and visualise the results. To solve this problem, the disease codes are grouped into comorbidities so that each node of the baseline disease network represents one of the comorbidities. Thus, the overall number of nodes is reduced into a reasonably small number of nodes. In this context, comorbidity represents a group of diseases or health conditions such as T2D, CHF and renal failure. In the literature, there are several well-established lists of comorbidities known as comorbidity index. Some of them are- Charlson comorbidities [10], Elixhauser comorbidities [14] and Charlson/Deyo comorbidities [31]. Elixhauser comorbidity index [14] was specially developed for measuring comorbidity using administrative data. This study adopted the Elixhauser index for the lists of comorbidities for the particular chronic disease.

2.4 Baseline Disease Network

This study uses several concepts from graph-theory to implement the proposed research framework. The definition of the graph-based concepts is introduced in this section.

The disease network represents the health trajectory of a large number of patients. It is referred to as individual disease network when it represents the medical history of an individual patient. The health trajectory shows the patient's disease transition from one

disease to another during subsequent admissions in the healthcare centre over time. Individual disease network shown in the middle part of Fig. 1 is essentially a graph where each node denotes the disease and the edge between two nodes indicates that these two nodes tend to occur sequentially. The edge is directional meaning that the disease of the patient found in the earlier admission is the source node and the disease that the patient has in a latter admission is considered as the target node of the network. If there are multiple diseases in any admission, then all possible disease pairs are considered in this study. In addition, when the patient has more than one disease in the same admission, all possible disease pairs are shown as bi-directional edges between them of the same admission. The additional information of node of the disease network called frequency that refers to the number of times the diseases have occurred for all the chronic disease patients considering all admissions. Similarly, the edge of the disease network has an attribute called weight that refers to the numbers of times two diseases have occurred simultaneously or in consecutive admissions.

The admission histories of the selected two cohorts are used to develop the baseline disease network. This study constructed two of such networks from two selected cohorts. The first network referred to as $N_{T2D\&CHF}$ is generated from $Cohort_{T2D\&CHF}$, i.e., who are first diagnosed with T2D and then diagnosed with CHF. The second network referred to as N_{T2D} is generated from $Cohort_{T2D}$, i.e., who are diagnosed with T2D but not diagnosed with CHF. These two networks represent the health trajectories of the patients of $Cohort_{T2D\&CHF}$ and $Cohort_{T2D}$, respectively. The baseline disease network is generated by merging the individual disease networks of the patients of respective cohorts. In this way, the attributes of the node and edge of individual disease networks are summed up. In general, the construction process of the baseline disease network is shown in Fig. 1.

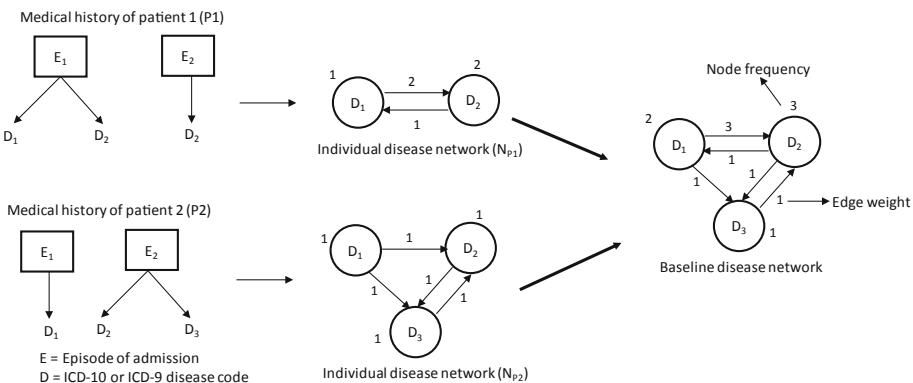


Fig. 1. Flow of generating the baseline disease network. The individual disease networks are constructed from medical history of the corresponding patients and are then aggregated to generate baseline disease network.

2.5 Final Disease Network Through Attribution Adjustment

After generating the two baseline disease networks from two cohorts, we merge them into one final network referred to as the final disease network (N_F). In this study, attribution theory is followed for generating the final disease network. Generally, attribution theory is the process of understanding the factors that are responsible for an event. These factors are used to predict the future occurrence of that event [32]. The baseline disease networks $N_{T2D\&CHF}$ and N_{T2D} give us a scenario of attribution principle. If we find a disease that has a higher node frequency in $N_{T2D\&CHF}$, it does not mean that this disease will be the risk factor for the developing of CHF in patients with T2D. Because that particular disease may be the higher prevalence in the diabetes patient cohort, N_{T2D} . However, instead of finding for more prevalent comorbidities in $N_{T2D\&CHF}$, this study focuses on looking for more prevalent comorbidities in $N_{T2D\&CHF}$ that are less prevalent in N_{T2D} . Thus, we look for more exclusive comorbidities or diseases in $N_{T2D\&CHF}$ compared to N_{T2D} and in the process adjust for the attribution effect.

After applying attribution adjustment, the nodes and edges of the final disease network (N_F) are generated by merging the nodes and edges of $N_{T2D\&CHF}$ and N_{T2D} . The frequency of any node in N_F is calculated by finding its relative frequency increment in $N_{T2D\&CHF}$ compared to N_{T2D} . Similarly, the weight of edges is calculated.

2.6 Procedure of the Proposed Framework

The input to the model is the hospital claim data collected from an Australian private healthcare fund. After pre-processing and filtering the dataset, this study selected two study cohorts, $Cohort_{T2D\&CHF}$ and $Cohort_{T2D}$. The disease information of the patients in the dataset is stored in the form of ICD codes. For $Cohort_{T2D\&CHF}$, the admission histories for T2D (first diagnosed at a certain time) and CHF (diagnosed after being diagnosed with T2D) are identified based on ICD codes. Then, all other ICD codes (related to comorbidities from Elixhauser index) between the two services for a patient (when he or she was first diagnosed with T2D and CHF, respectively) are considered to generate the individual disease network. These individual networks are then aggregated to construct the baseline disease network, $N_{T2D\&CHF}$ for patients who were first diagnosed with T2D and then diagnosed with CHF. The similar process is applied on the dataset to generate the baseline disease network, N_{T2D} for patients who were diagnosed with T2D but not diagnosed with CHF. Next, the two baseline disease networks, $N_{T2D\&CHF}$ and N_{T2D} are merged by applying attribution principle and thus the final disease network (N_F) is created. Finally, SNA and graph theory are applied on N_F to understand the disease progression of CHF in T2D. The complete workflow of the proposed model is illustrated in Fig. 2.

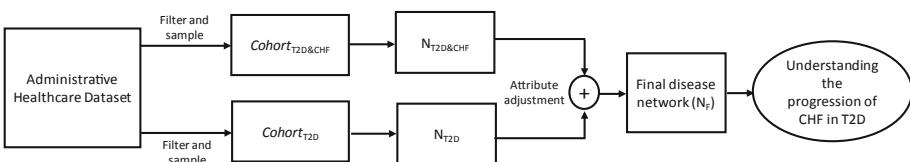


Fig. 2. Block diagram of the proposed framework to understand the progression of CHF in patients with T2D.

3 Results and Discussions

After pre-processing the research dataset, we identified the T2D and CHF patients by looking for the corresponding ICD codes (Table 1) in the admission history of the patients. We found a total 3,908 and 656 patients containing T2D and CHF related ICD codes, respectively. Then a total number of 100 patients is selected for the $Cohort_{T2D\&CHF}$ in where the patients from that cohort first diagnosed with T2D and then diagnosed with CHF at a later stage. This study needs an equal number of patients for each cohort to generate the final disease network. Thus, a total number of 100 patients is randomly selected for the $Cohort_{T2D}$ in where the patients diagnosed with T2D but not diagnosed with CHF in his/her entire life. Next, two baseline disease networks ($N_{T2D\&CHF}$ and N_{T2D}) are generated from two cohorts. The admission data for the cohorts were coded in both ICD-9-AM and ICD-10-AM as the dataset was based on an Australian healthcare context collected from a private insurance company in Australia. In this study, the Elixhauser comorbidity index is used to select the comorbidity lists for the two selected cohorts. The ICD codes of the Elixhauser comorbidity lists are only considered to generate the baseline disease networks. Therefore, the ICD codes using Elixhauser index are available in ICD-9-CM and ICD-10-CM format whereas our data are coded in ICD-9-AM and ICD-10-AM format. However, we used a translation table from the study of Quan et al. [33] and manual verification process to convert the Elixhauser comorbidity codes to ICD-9-AM and ICD-10-AM. The adapted Elixhauser index [34] included 31 comorbidities. As this study focus is understanding the progression of CHF in patients with T2D, we removed 3 comorbidities related to T2D and CHF. Thus, the comorbidity lists reduced into 28 groups. Each comorbidity or disease group has the corresponding ICD codes. If a patient' admission data contain one or more of those ICD codes, the patient is considered to have the corresponding comorbidity(s).

Table 2 shows the most prevalent comorbidities or diseases from the two baseline disease networks. For the network $N_{T2D\&CHF}$, the diseases or comorbidity conditions corresponding to the ICD codes were diagnosed after the diagnosis of T2D at a certain time but before the diagnosis of CHF, whereas the diseases for the network N_{T2D} corresponding to the ICD codes were diagnosed before the diagnosis of T2D. In general, the statistics of the Table 2 refers that the patients with both T2D and CHF have more comorbidities than patients with T2D. In addition, there is a difference between the two baseline disease networks in term of the most occurring comorbidity conditions. The $N_{T2D\&CHF}$ baseline disease network shows significantly high prevalence of cardiac arrhythmias, chronic pulmonary disease, hypertension and kidney disease. These diseases are the risk factors of diabetic patients that are associated with developing CHF. This is consistent with the studies in the literature [35, 36]. However, the frequency of node of the two baseline disease networks can give comparative insights about the prevalence of comorbidity conditions between the patients with both T2D and CHF and the patients with only T2D.

Table 2. Top-10 most prevalent comorbidities for patients with both T2D and CHF, and patients with only T2D. The prevalence refers the number of admissions that have ICD codes related to those comorbidities.

Comorbidity conditions for $N_{T2D\&CHF}$	Prevalence	Comorbidity conditions for N_{T2D}	Prevalence
Cardiac arrhythmias	56	Solid tumour without metastasis	43
Chronic pulmonary disease	40	Renal failure	30
Renal failure	39	Peptic ulcer disease excluding bleeding	16
Solid tumour without metastasis	37	Hypertension	16
Fluid and electrolyte disorders	36	Liver disease	15
Peptic ulcer disease excluding bleeding	35	Lymphoma	12
Peripheral vascular disorders	28	Cardiac arrhythmias	10
Deficiency anaemia	13	Psychoses	7
Hypertension	12	Obesity	5
Pulmonary circulation disorders	12	Drug abuse	5

In the next step, the final disease network, N_F derived from the two baseline disease networks ($N_{T2D\&CHF}$ and N_{T2D}) shows the unique characteristics of CHF progression in patients with T2D. By applying attribution adjustment, N_F network assigned a higher weight to comorbidity conditions (i.e., node frequency) and their progression (i.e., edge weight) for those that are more prevalent in patients with both T2D and CHF compared to the patients with T2D only. The node frequency and edge weight of N_F are normalised to the range of 0 to 1. Figure 3 shows the top 10 comorbidities or diseases of N_F with their normalised scores. The score 1 for ‘chronic pulmonary disease’ indicates that this disease was exclusive to the patients with both T2D and CHF. The score more than 0.4 but less than 1 for any disease (such as valvular disease and cardiac arrhythmias) indicates that this disease was the most prevalent in the $N_{T2D\&CHF}$ and had small prevalent in the N_{T2D} . The other comorbidities like depression and solid tumour without metastasis gained scores less than 0.4. This refers to the little differentiation of occurring the comorbidities between the two baseline disease networks. Figure 3 suggests that the patients with T2D have certain risk factors like chronic pulmonary disease, valvular disease, cardiac arrhythmias and renal failure that are associated with developing CHF. This observation is also consistent with the study of Nichols et al. [16] and Zhang et at. [37].

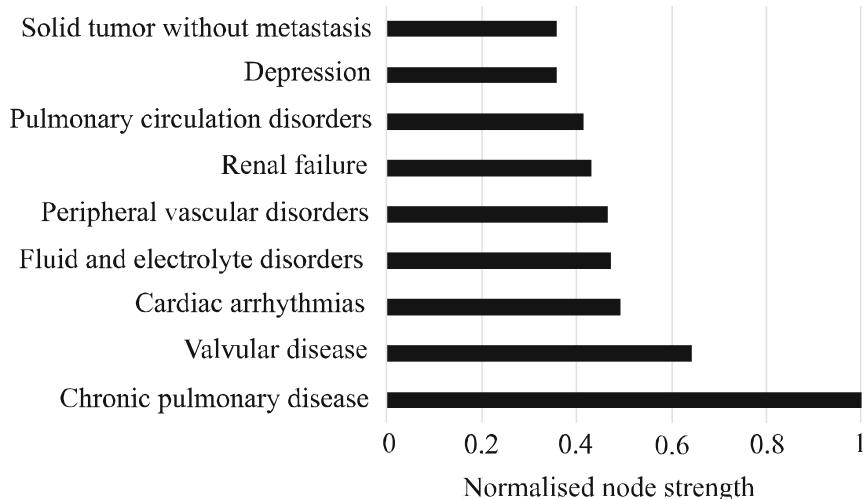


Fig. 3. Top-10 comorbidity conditions that attributed most for CHF in patients with T2D.

The pair of most prevalent disease progression of developing CHF for type-2 diabetic patients are given in Table 3. The prevalent transitions show a lot of comorbidity conditions or diseases that are associated with the progress towards the CHF in patients with T2D. In Table 3, the highest transition which is from fluid and electrolyte disorders to renal failure refers to two different body system problems. Fluid and electrolyte disorders is the deficiency or excess in key minerals (e.g., calcium and phosphorous) and electrolyte imbalances (e.g., sodium and potassium). The patients with T2D develop a constellation of fluid and electrolyte disorders [38]. Also, potassium disorders can lead to the development of cardiovascular disease [39]. Whereas, renal failure is chronic kidney disease (CKD). There is a strong comorbid relationship between CKD and CHF for a diabetic patient [37, 40]. Therefore, the transition between “fluid and electrolyte disorders” and “renal failure” may be a potential risk factor for the progression towards CHF in patients with T2D. In addition, the normalised weight of this transition is 1 indicates that this is exclusive in the patients with both T2D and CHF compared to the patients with T2D. The other top transitions shown in Table 3 indicate the contribution to developing the CHF for diabetic patients.

Figure 4 shows the graph representation of the final disease network (N_F) that represents the overall health trajectory of the two cohorts in terms of comorbidity progression over time. The visualisation and subsequent analysis of the network were done in social network analysis software, Gephi [41]. The nodes in the figure represent the comorbidity conditions or diseases. The size of the nodes and labels are proportional to the prevalence of corresponding comorbidity condition. Cardiac arrhythmias, fluid and electrolyte disorders, chronic pulmonary disease and renal failure are dominating

Table 3. Top-10 most prevalent transitions between comorbidities in the final disease network (N_F).

Initial condition	Next condition	Normalised weight
Fluid and electrolyte disorders	Renal failure	1
Cardiac arrhythmias	Peptic ulcer disease excluding bleeding	0.6
Obesity	Cardiac arrhythmias	0.55
Cardiac arrhythmias	Chronic pulmonary disease	0.5
Cardiac arrhythmias	Renal failure	0.5
Cardiac arrhythmias	Fluid and electrolyte disorders	0.45
Fluid and electrolyte disorders	Other neurological disorders	0.4
Pulmonary circulation disorders	Renal failure	0.4
Weight loss	Cardiac arrhythmias	0.33
Chronic pulmonary disease	Obesity	0.33

the final disease network. It reflects that these comorbidity conditions are the risk factors to progress towards the CHF in patients with T2D. The network shows a large number of edges that represent the transition from one disease to another disease. The thickness of the edge is proportional to its weight. In the final disease network, the highest thickness of the edge between “fluid and electrolyte disorders” and “renal failure” indicates that these two diseases are the risk factors to the development of CHF in patients with T2D.

Finally, this study performed a network comparison of the three baseline disease networks. Several social network-based measures are calculated on these networks and the results are shown in Table 4. Table 4 shows that the total number of nodes for $N_{T2D\&CHF}$ is higher than the total number of nodes for N_{T2D} . This indicates the higher comorbidity conditions of the patients with both T2D and CHF than the patients with T2D only. Also, the edge count in $N_{T2D\&CHF}$ is the thrice of the edges in N_{T2D} . This indicates that patients with both T2D and CHF have relatively more admissions and more transitions between comorbidity conditions in subsequent admissions. Also, this may indicate the more exclusive comorbidity conditions in subsequent admissions. The high graph density for the network, $N_{T2D\&CHF}$ also strengthens this fact. The graph density of $N_{T2D\&CHF}$ is higher than the graph density of N_{T2D} . This suggests that the patients with both T2D and CHF present a higher admission burden and complex progression structure over subsequent admissions. The remaining two measures (e.g., average clustering coefficient and average path length) do not show important features in the present context.

Table 4. Network analysis of three baseline networks.

Criteria	N _{T2D&CHF}	N _{T2D}	N _F
Number of nodes	23	16	26
Number of edges	122	39	122
Graph density	0.241	0.16	0.188
Network diameter	4	5	4
Average clustering co-efficient	0.406	0.344	0.362
Average path length	1.925	2.371	1.925

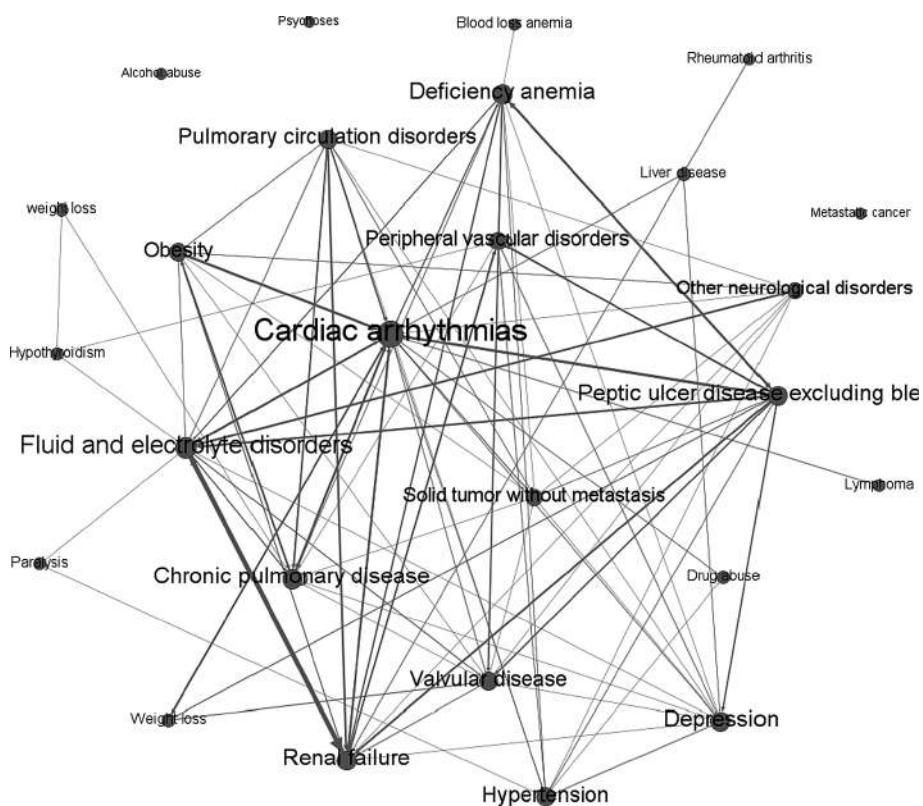


Fig. 4. Final disease network after attribute adjustment to understand the progression of CHF in patients with T2D. Node size and labels are proportional to the prevalence of corresponding comorbidity condition. The thickness of an edge between two comorbidity conditions is proportional to its weight.

4 Conclusions

This study introduced a research framework to understand the comorbidity of two chronic diseases (T2D leading to the development of CHF). For this, the proposed framework applied graph theory and social network analysis on hospital claim data. The final disease network constructed from two cohorts (i.e., patients with both T2D and CHF and patients with T2D only) can represent the overall health trajectory of the patients of cohorts in terms of comorbidity progression over time. By analysis the network and its attributes, this study showed some risk factors (i.e., chronic pulmonary disease, renal failure and cardiac arrhythmias) that are associated to the development of CHF in patients with T2D. Thus, the proposed network representation can help the healthcare provider to understand high-risk diseases and progression pattern between recurrence of T2D and CHF. Also, it can help to manage the healthcare resources efficiently. As future work, the knowledge generated from the final disease network of comorbidity of chronic diseases can be utilised to develop predictive models for future chronic disease. This can be implemented by comparing the baseline disease network with the disease progression network of individual test patients. If the test patient's network mostly matches with the baseline disease network, the patient will progress on that chronic disease pathway.

References

1. Australian Institute of Health and Welfare, Chronic Diseases (2017). [www document]. <http://www.aihw.gov.au/chronic-diseases>. Accessed 03 Aug 2017
2. Australian Institute of Health and Welfare, 1 in 5 Australians affected by multiple chronic diseases (2015). [www document]. <https://www.aihw.gov.au/news-media/media-releases/2015/august/1-in-5-australians-affected-by-multiple-chronic-di>. Accessed 03 Aug 2017
3. Capobianco, E.: Comorbidity: a multidimensional approach. Trends Mol. Med. **19**(9), 515–521 (2013)
4. Ponikowski, P., et al.: 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur. J. Heart Fail. **18**(8), 891–975 (2016)
5. Thrainsdottir, I.S., et al.: The association between glucose abnormalities and heart failure in the population-based Reykjavik study. Diabetes Care **28**(3), 612–616 (2005)
6. Dei Cas, A., et al.: Impact of diabetes on epidemiology, treatment, and outcomes of patients with heart failure. JACC: Heart Fail. **3**(2), 136–145 (2015)
7. MacDonald, M.R., et al.: Impact of diabetes on outcomes in patients with low and preserved ejection fraction heart failure: an analysis of the Candesartan in Heart failure: assessment of Reduction in Mortality and morbidity (CHARM) programme. Eur. Heart J. **29**(11), 1377–1385 (2008)
8. World Health Organisaion | International Classifications of Diseases (ICD) (2019). <https://www.who.int/classifications/icd/en/>. Accessed 22 May 2019

9. Hossain, M.E., Uddin, S.: Understanding the comorbidity of multiple chronic diseases using a network approach. In: Proceedings of the Australasian Computer Science Week Multiconference. ACM (2019)
10. Charlson, M.E., et al.: A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J. Chronic Dis.* **40**(5), 373–383 (1987)
11. Wong, D.T., Knaus, W.A.: Predicting outcome in critical care: the current status of the APACHE prognostic scoring system. *Can. J. Anesth./Journal canadien d'anesthésie* **38**(3), 374–383 (1991)
12. Ewing, D.J., et al.: The value of cardiovascular autonomic function tests: 10 years experience in diabetes. *Diabetes Care* **8**(5), 491–498 (1985)
13. Cederholm, J., et al.: Risk prediction of cardiovascular disease in type 2 diabetes: a risk equation from the Swedish National Diabetes Register. *Diabetes Care* **31**(10), 2038–2043 (2008)
14. Elixhauser, A., et al.: Comorbidity measures for use with administrative data. *Med. Care* **36**(1), 8–27 (1998)
15. Khan, A., Uddin, S., Srinivasan, U.: Comorbidity network for chronic disease: a novel approach to understand type 2 diabetes progression. *Int. J. Med. Inform.* **115**, 1–9 (2018)
16. Nichols, G.A., et al.: Congestive heart failure in type 2 diabetes: prevalence, incidence, and risk factors. *Diabetes Care* **24**(9), 1614–1619 (2001)
17. Nichols, G.A., et al.: The incidence of congestive heart failure in type 2 diabetes: an update. *Diabetes Care* **27**(8), 1879–1884 (2004)
18. Davis, D.A., et al.: Time to CARE: a collaborative engine for practical disease prediction. *Data Min. Knowl. Disc.* **20**(3), 388–415 (2010)
19. Gupta, S., et al.: Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open* **4**(3), e004007 (2014)
20. Davis, D.A., et al.: Predicting individual disease risk based on medical history. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. ACM (2008)
21. Jensen, F.V.: An Introduction to Bayesian Networks, vol. 210. UCL Press, London (1996)
22. Faruqui, S.H.A., et al.: Mining patterns of comorbidity evolution in patients with multiple chronic conditions using unsupervised multi-level temporal Bayesian network. *PLoS ONE* **13**(7), e0199768 (2018)
23. Himes, B.E., et al.: Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. *J. Am. Med. Inform. Assoc.* **16**(3), 371–379 (2009)
24. Anderson, J.G.: Evaluation in health informatics: social network analysis. *Comput. Biol. Med.* **32**(3), 179–193 (2002)
25. DuGoff, E.H., et al.: A scoping review of patient-sharing network studies using administrative data. *Transl. behav. Med.* **8**(4), 598–625 (2018)
26. Soulakis, N.D., et al.: Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. *J. Am. Med. Inform. Assoc.* **22**(2), 299–311 (2015)
27. Uddin, S., et al.: A study of physician collaborations through social network and exponential random graph. *BMC Health Serv. Res.* **13**(1), 234 (2013)
28. Uddin, S., Khan, A., Piraveenan, M.: Administrative claim data to learn about effective healthcare collaboration and coordination through social network. In: 2015 48th Hawaii International Conference on System Sciences (HICSS). IEEE (2015)
29. Merrill, J.A., et al.: Transition networks in a cohort of patients with congestive heart failure. *Appl. Clin. Inform.* **6**(03), 548–564 (2015)
30. ACCD. Australian Consortium for Classification Development (2019). [www document]. <https://www.accd.net.au/Icd10.aspx>. Accessed 12 June 2019

31. Deyo, R.A., Cherkin, D.C., Cioł, M.A.: Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J. Clin. Epidemiol.* **45**(6), 613–619 (1992)
32. Moskowitz, G.B.: Social Cognition: Understanding Self and Others. Guilford Publications (2013)
33. Quan, H., et al.: Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* 1130–1139 (2005)
34. Garland, A., et al.: The epidemiology and outcomes of critical illness in Manitoba (2011). Accessed 1 Dec 2011
35. Tong, B., Stevenson, C.: Comorbidity of cardiovascular disease, diabetes and chronic kidney disease in Australia. Australian Institute of Health and Welfare (2007)
36. Huo, X., et al.: Risk of non-fatal cardiovascular diseases in early-onset versus late-onset type 2 diabetes in China: a cross-sectional study. *Lancet Diabetes Endocrinol.* **4**(2), 115–124 (2016)
37. Zhang, J., Gong, J., Barnes, L.: HCNN: heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records. In: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). IEEE (2017)
38. Liamis, G., et al.: Diabetes mellitus and electrolyte disorders. *World J. Clin. Cases: WJCC* **2**(10), 488 (2014)
39. Barbosa, A., Sztajnbok, J.: Fluid and electrolyte disorders. *Jornal de pediatria* **75**, S223–S233 (1999)
40. Gansevoort, R.T., et al.: Chronic kidney disease and cardiovascular risk: epidemiology, mechanisms, and prevention. *Lancet* **382**(9889), 339–352 (2013)
41. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. *Icwsrm* **8**(2009), 361–362 (2009)



Networks of Function and Shared Ancestry Provide Insights into Diversification of Histone Fold Domain in the Plant Kingdom

Amish Kumar¹ and Gitanjali Yadav^{1,2()}

¹ National Institute of Plant Genome Research, New Delhi 110067, India
gy246@cam.ac.uk

² Department of Plant Sciences, University of Cambridge,
Cambridge CB23EA, UK

Abstract. The Histone Fold Motif (HFM) of core histone proteins is one of the most highly conserved signature motifs in living organisms. Despite significant variation in sequence over millions of years of evolution, the HFM retains a distinctive structural fold that has diversified into several non-histone protein families. We have identified over 4000 HFM containing proteins in plants that are not histones, raising the question of why the family has expanded so considerably in the plant kingdom. We find that a majority of non-histone HMFs are playing regulatory roles, and that they are distributed widely within and across taxonomic groups. In this work we explore the relationships between the HFM of non-histones and that of their ancestral core histone forerunners, using a network approach. Networks of core histones and non-histone counterparts were superimposed with additional layers of complexity, like functional annotations, sub cellular locations, taxonomy and shared ancestry. HFM networks of model plants rice and *Arabidopsis* were investigated in terms of gene expression, interactions with other proteins as well as regulatory potential, to gain insights into diversification events during evolution that are not immediately evident from phylogenetic trees or raw data alone. Taken together, the networks elucidate diverse paths of evolution of the histone fold motif, leading to sub-functionalization and neo-functionalization of the HFM.

Keywords: Histone fold motif · Plant kingdom · Co-expression networks · Gene regulatory networks · Protein-protein interaction networks

1 Introduction

1.1 The Histone Fold Motif (HFM)

Histones are one of the most evolutionary conserved and ubiquitous proteins among all eukaryotes, fundamental to formation of DNA compaction units called ‘nucleosomes’, which, in turn are known to organize genomic DNA into chromatin (Baxevanis et al. 1995). Five major families of histones exist: H1/H5 and H2A, H2B, H3, and H4, of which the latter four are known as the core histones, and are known to bind to each other as dimers in very specific combinations; H2A with H2B, and H3 with H4.

Despite differences in their DNA sequence, these four core histone subunits are relatively similar in 3-dimensional structure and are highly conserved through evolution, all featuring a ‘helix turn helix turn helix’ motif known as the Histone Fold Motif (HFM). The HFM is found in all four core histones, and it serves as the recognition site between pairing partners of core histones in the two combinations mentioned above. Without histones, the unwound DNA in most living organisms would be very long and impossible to confine within cells, swelling into a length to width ratio of more than 10 million to 1.

1.2 Non-histone HFM Containing Proteins

Apart from core histones, the HFM has been detected in a large number of proteins of multicellular organisms, and such non-histone proteins are often regulatory in nature, including several families of transcription factors, co-activators and repressors of gene expression. However, even though the non-histone HFM proteins have lost the ability to compact DNA, many of them appear to have retained the ability to form dimers, as well as to bind with DNA. Among the most well studied families known to have the HFM, are the NF-Y and TAF families. TAFs represent TATA Binding Protein Associated Factors, which are activators of gene expression, playing important roles in transcriptional initiation and regulation, and several TAFs are known to form histone-like dimers (Hoffmann et al. 1996; Xie et al. 1996). NF-Ys represent Nuclear Factor-Y family, which in turn has three subunits, namely the NF-YA, NF-YB and NF-YC, respectively, all of which are necessary for DNA binding. Of these three sub-units, NF-YB and NF-YC are known to form a heterodimer very similar to the histone H2A-H2B complex (Romier et al. 2003). Other than these two major transcription factor families, the HFM is also found in a transcription repressor known as Negative cofactor 2 (NC2), composed of two subunits α and β , also known as Down regulator 1 down regulator associated protein 1 (DR1 and DRAP1). Very little is known about this family, even though the interaction of DR1 and DRAP1 is reported to be similar to the H2A-H2B dimer (Kamada et al. 2001).

1.3 HFMs in Plants – A Network Approach

In the plant kingdom, only the NF-Y class of functional non-histone HFM proteins are well characterized, whereas other known classes like TAFs, Dr1/DrAp1, H3-like Centromeric proteins have been investigated inadequately, and still other families like the Chromatin accessibility complex, DNA polymerase epsilon subunits, and Centromeric Protein-S are very poorly known (Kumar and Yadav 2016). In order to overcome this gap in knowledge, we report here a large scale network analysis of about 4000 non-histone HFM containing proteins in the plant kingdom, that were identified recently in our group (Kumar 2019). A network approach has already been established as one of the most effective ways of reducing data dimensionality especially in large datasets like genes and proteins, and to gain insights at the systems level. Our initial networks were constructed across all 64 plants to bring out aspects of lineage, major functional classifications and shared ancestry between core histone domains and their non-histone HFM containing counterparts. The individual HFM domains and their

ancestral associations with core histone subunits were superimposed with taxonomic information and functional annotations in order to get a comprehensive view of cross-taxon relationships. This was followed by a detailed network analysis of HFM diversification within two model plants, namely *Oryza sativa* (rice) and *Arabidopsis thaliana*. This stage of the analysis involved constructions and analysis of multiple sub-networks of gene expression, co-expression across various stages and plant tissues, protein-protein interactions, and finally, gene regulatory networks of the non-histone HFM domains. Topological analyses of all networks enabled identification of complex diversification events that have taken place among the non-histone HFM containing proteins within and across the two major taxa, resulting in a huge diversity of new functions in the present day HFM domains. Most importantly, the network analysis brought out relationships between the domains that were not evident from phylogenetic studies alone.

2 Methodology

2.1 HFM Identification and Annotation

Plant core histone protein sequences were obtained by taxonomic filtering of the UNIPROT database (<http://www.uniprot.org/>). This dataset was used to train hidden markov models to accurately identify class specific HFMs in new sequences, and these profile HMMs were then used to scan complete proteome data for 64 plant species as described in Kumar (2019). The results were filtered for significant homologs based on individual sequence e-value. All hits scoring better than default e-value (e-value 0.001) were considered as HFM containing proteins. Functional annotation of uncharacterized sequences among all identified HFM containing proteins was done by Blast2GO tool (Conesa et al. 2005). All identified HFM containing proteins were grouped based on their functional annotation, followed by assignment to the respective ancestral core histone class, based on best e-value for each HFM domain in the four HMMER outputs. For assignment of ancestral core histone subunits, redundant sequences were removed from each HMMER scan, by retaining only the most significant homolog based on e-value, and these were assigned to the respective core histone class. For example, if a sequence was identified by both H2A and H2B profiles, then it would be assigned to parental class of H2A or H2B based on which profile ranked higher in HMMER score.

2.2 Species Specific HFM Gene and Protein Datasets

Gene expression of selected non-histone HFM genes data analysis was performed on the GENEVESTIGATOR 7.2.0 (Zimmermann et al. 2004) platform. In *Arabidopsis* 5825 wild type Affymetrix ATH1 Genome array data and in Rice 1906 wild type Affymetric Rice Genome array data was used for this. Genes were clustered based on PCC of their normalized relative expression. Physical interaction networks of non-histone HFM proteins were generated using protein-protein interaction data in STRING (<https://string-db.org>). Following this, all proteins were subjected to Gene Ontology (GO) enrichment in

order to identify the pathways/processes in which the HFM are involved. GO enrichment analysis was done using BiNGO (Maere et al. 2005) and Cluego (Bindea et al. 2009) applications of cytoscape. To understand the regulation of non-histone HFM proteins, a region spanning 1.7kB was extracted as promoter element, encompassing 1500 bp upstream and 200 bp downstream from the annotated transcription start site (TSS) of each gene using custom perl scripts. Cis-regulatory elements were identified on each promoter region by scanning through position weight matrices of unique collection of DNA-binding models, available from the Match TM program in the TRANSFAC ® professional suite. The resulting total occurrence of distinct regulatory sites on each non-histone HFM promoter were considered as observed frequencies (O). This was compared with data from a control-set of promoters, comprising of random genes of both the plants. Total occurrence of distinct regulatory sites on these random promoters was normalised and treated as expected frequency (E). The observed and expected frequency of each distinct regulatory site was later checked for significance of enrichment at $p < 0.05$ using the chi-square (χ^2) test. Only the significant hits were used for construction of a gene regulatory network (GRN).

2.3 Network Construction and Topological Analyses

All networks were constructed in structured information file (SIF) format and incorporated into R-CRAN version 3.6.1, and visualized in Cytoscape version 3.6 (Shannon et al. 2003), followed by topological analyses. Plots of network parameters were constructed using NetworkAnalyser package of Cytoscape. Perturbation analysis on high scoring nodes was performed using our in-house webserver NEXCADE (Yadav and Babu 2012).

3 Results

3.1 Annotation of Identified HFM Domains in Plants

We have identified about 4000 non-histone HFM containing domains of plant origin using profile hidden markov models (Kumar 2019), which have been used in this work for exploring the extent to which diversification has occurred in novel HFM sequences across 64 complete plant proteomes representing all major taxa of green plants. As described in methods, these non-histone HFM proteins were assigned to four parental histone classes based on homology to parental core histone subunits, and Fig. 1 depicts this assignment as a Bar chart.

As can be seen in this Figure, the H2B-HFM class was found to have highest number of non-histone protein homologs (1650), followed by H2A-HFM (1193) and H4-HFM (914). The least number of non-histone HFM proteins (161) were found to match with H3-HFM. Functional annotation of the newly identified 3918 non-histone HFM proteins in the plant kingdom revealed their presence in various protein sub-families, many of which are transcription factors or activators or co-regulators, as shown in Table 1, where all new annotations are ordered by the parental core histone class. As can be seen from this table, H2B subunits have evolved and diversified into a

large group of non-histones proteins, whereas H3 subunit has shown minimal evolution into non-histones, evident also from Fig. 1.

As shown in Table 1, we find not only the most commonly known NF-Y, TAF and Dr1/DRAP1 classes of transcription factors in our data, but a huge expansion of the HFM to cover several other groups of non-histone proteins as well. More than half (52%) of the identified non-histone HFM proteins in plants belong to these three large multi gene families, namely the NF-Ys, followed by several classes of TAFs (23%) and then Dr1/DrAp1 (10%) as shown in Table 1. The additional families identified in the analysis have very rarely been reported/characterized or studied in plants, if at all. These represent the Chromatin accessibility complex (CHRAC), DNA polymerase epsilon subunit 3 (DPE-3) and DNA polymerase epsilon subunit C (DPE-C), Centromeric histone 3 (CENH3), Centromeric proteins S (CEN-S) and the Bromodomain transcription factor, listed in Table 1.

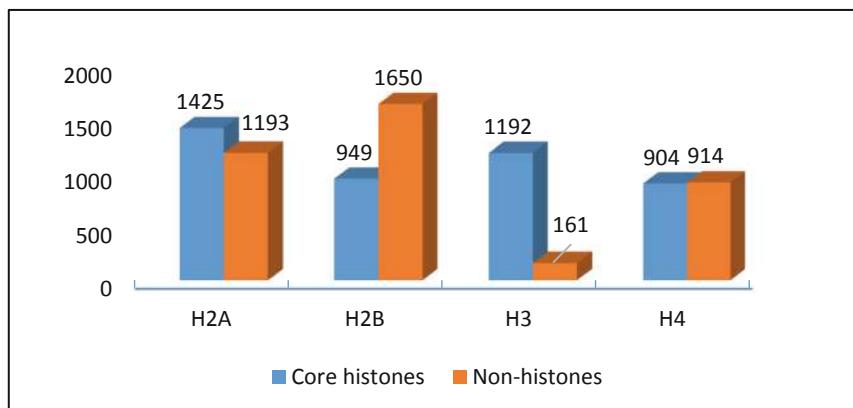


Fig. 1. Bar chart showing histones and non-histone sequence proportions in HMM search.

All identified non-histone HFM containing proteins were assigned to four parental groups of core histones, namely H2A, H2B, H3 and H4, based on e-value of the HFM in each profile HMM. Within most of these families, we were able to identify partnering pairs of domains with homology to H2A and H2B subunits of core histones, known to form dimers. For example, majority of NF-YB homologs could be assigned to H2B class of histones, while members of the subunit NF-YC were identified as H2A homologs. Similarly, Down regulator 1 (DR1) and Down Regulator associated protein 1 (DrAp1), assigned were assayed to H2B and H2A classes of HFM respectively. It is interesting to note that both NF-YB/NF-YC and Dr1/DrAP1 constitute cognate pairing partners (Bellorini et al. 1997; Kumar and Yadav 2016), just as in case of H2A/H2B, suggesting that pairwise protein-protein interactions are maintained between NF-YB and C, as well as between Dr1 and DrAP1 classes of non-histone proteins, mediated via the conserved HFM.

In contrast to the previous pairwise interactors, we did not find any domains that may be interacting with counterparts of DPE or CHRACs. Members of CHRAC and DPE-C functional classes of non-histone HFM proteins were found to be homologous to H2A-HFM, with 23 CHRAC and 71 DPE-C members being identified in this study, as shown in Table 1. Other non-histone proteins with homology to H2A-HFM belong to DNA repair proteins, Calmodulin binding proteins, Neurofilament heavy polypeptide like proteins. These are in very small numbers and assigned as ‘others’ in Table 1.

Table 1. Assignment of parental histone class to plant HFM domains

Core histone group	Non-histone protein class	Non-histones within group	Total
H2A	NF-YC	847	1193
	DrAp1	228	
	CHRAC	23	
	DPE-C	71	
	Others	24	
H2B	NF-YB	1199	1650
	Dr1	153	
	TAF12	237	
	Others	61	
H3	CENH3	91	161
	Others	70	
H4	CEN-S	61	914
	DPE-3	87	
	TAFs	659	
	Bromodomain	18	
	Others	89	

Surprisingly, even though TAF subunits were found in significant numbers, we could not identify pairing partnerships amongst these. TAFs remain largely unexplored in plants but there is significant literature on this family in yeast, drosophila and humans. TAFs have also been reported in TBP less acetyl-transferase complex and SAGA complexes (Gangloff et al. 2000). Majority of TAFs were identified as homologs of H4, including TAF6, TAF8, TAF9, TAF10, and TAF11, and none could be found matching the H3, the known pairing partner of H4. In vitro, TAFs are known to form dimers similar to the histone heterodimers, and five such dimers have been identified among TAFs, namely TAF3-10, TAF6-9, TAF4-12, TAF8-10 and TAF11-13 (Gangloff et al. 2000; Werten et al. 2002). However, we could not find significant HFM homology in any member of TAF13 and TAF4 in the plant kingdom. Furthermore, H4 ancestral core histone subunit could be assigned to all TAFs classes, with the exception of TAF12, which was closely related to H2B-HFM as shown in Table 1. We could not identify any TAFs subunits that could be assigned to the H2A-HFM, as potential partners of TAF12.

In order to resolve parental classes for NF-Y and DRAPs which are often wrongly annotated as being one and the same, we performed detailed phylogenetic and network analyses. The investigation was also meant to identify potential partners among the large TAF family, as well as to resolve the true ancestry of TAF12, if not H2B, as described above. In the last section, we have further attempted to resolve the issue of several DR1/DRAP1 s being mis-annotated in literature as NF-Ys (Petroni et al. 2012; Hackenberg et al. 2012) and this was done using a combination of functional networks, superimposed with interactome, transcriptome and annotation.

3.2 Network Analyses

The complete HFM data described above was collated to generate networks of associations between present day HFM domains and their four ancestral core histone subunits as shown in Fig. 2, where nodes represent core histone subunits and major functional groups of non-histone HFMs as described in earlier sections, across six taxonomic groups in the plant kingdom, moving from the evolutionary hierarchy of lower plants to higher plants, namely algae, ferns, pines and finally the flowering plants. Edges represent ancestry of the HFM domains, and the major plant taxa are depicted stemming from the identified HFM proteins, namely Chlorophytes, Tracheophytes, Gymnosperms, Amboreales, Dicots and Monocots. Node colors and labels distinguish the plant taxa for HFM domains, while edge colors represent the four ancestral core histone classes.

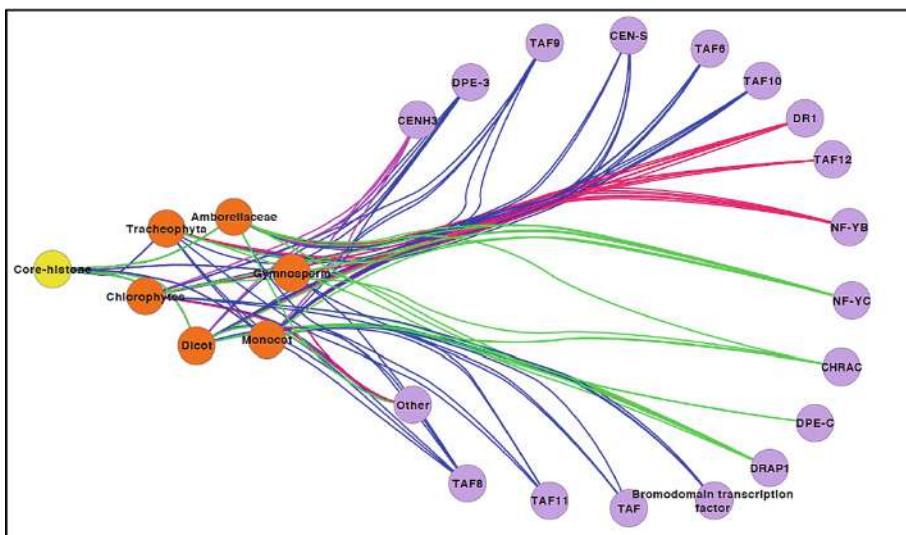


Fig. 2. Network representation of the association between core histone subunits (yellow) and present day HFM functions (purple) across major taxonomic groupings (orange) in the plant kingdom. Edges represent ancestry to H2A (green), H2B (red), H3 (pink) or H4 (blue).

The network illustrates how all four core histones have diversified into a large number of non-histone functions in present day HFM containing non-histones, and that each functional group stems from one of the four core histone classes. Overall, the inferences from Table 1 manifest clearly from the network; diversification has taken place extensively in H2A, H2B and H4 classes of core histones, but not in H3. This is also evident from the high topological coefficient of all functional group nodes, and is based on the wide occurrence and conservation of histones in all forms of life. Chlorophytes are much lower on the taxonomic hierarchy, but this group has a higher degree than other lower plant groups in the network, despite having lower numbers of non-histone HFM domains, suggesting greater diversification of the limited domains into various present day functions.

In order to better understand relationships between individual HFM domains and their core histone counterparts, a more detailed network was constructed to incorporate all taxa and each HFM within each family, and this is illustrated in Fig. 3. As expected, the network is fully connected with a diameter of six and characteristic pathlength of 3.3. It has 8478 nodes including all newly identified HFM domains along with ancestry assignment to core histones, classified according to major taxonomic groups.

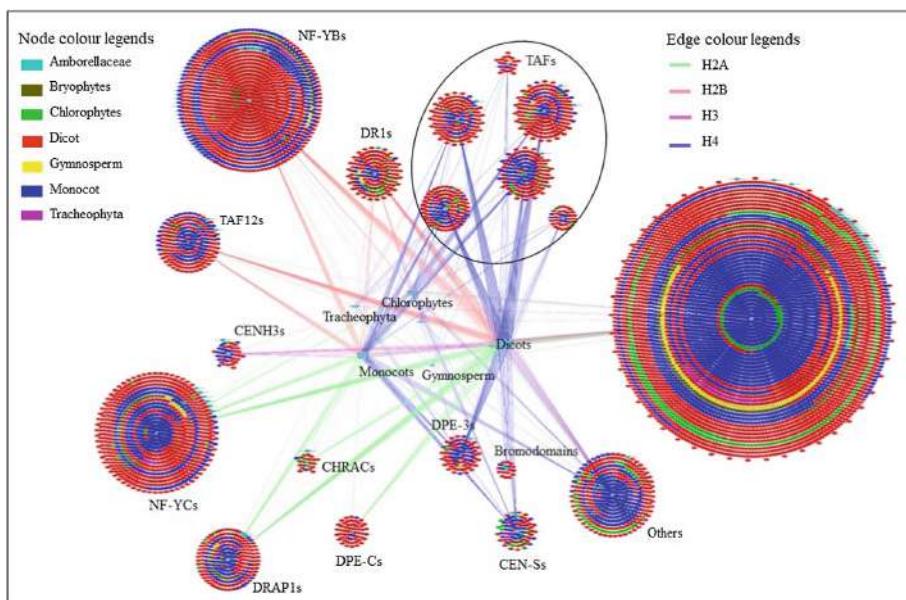


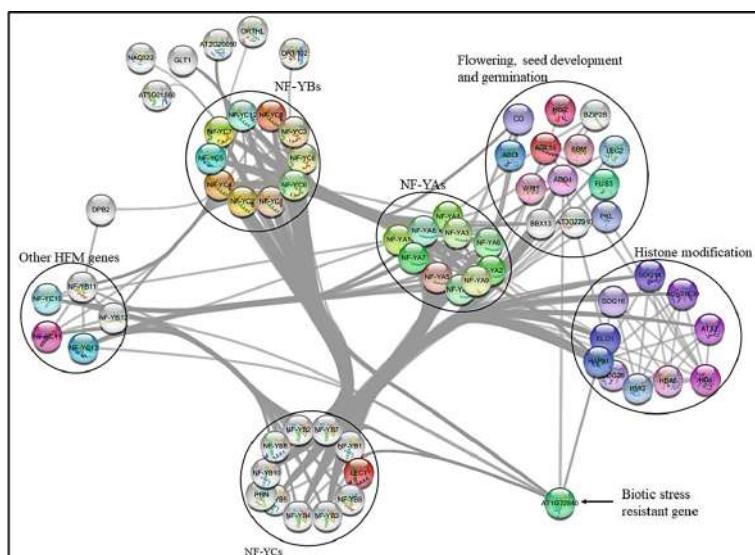
Fig. 3. Network representation of all identified HFM proteins. Node colors represents the plant taxa of the HFM protein while edge colors represents ancestral core histone class.

As can be seen in the detailed network in Fig. 3, diversification events during evolution of the HFM become clearer within each functional group. Non-histone HFM genes belong to several functional families, including the well-studied NF-Ys and TAF groups, but also the less investigated DR1/DRAPs, CENs and DPE groups. Most

functions have evolved in each taxonomic group, and the clusters of NF-YB/NF-YC, and DRAP-1/DR1 are present in comparable numbers within respective taxa to suggest existence of cognate pairing partners among present day proteins, a feature that remains unexplored for most HFM s in the plant kingdom, dissected further in the next section.

3.3 Species Specific Networks of Function

Ancestry networks in the previous section indicated existence of interacting partners within several classes of non-histone HFM s, and here we describe species specific functional networks that support our predictions on partner pairing preferences. For predicting cognate partners, we have combined structural bioinformatics and gene expression information. For example, we identified 10 NF-YBs and 10 NF-YCs in model plant *Arabidopsis*, and these can theoretically interact in 100 possible combinations. Each combination was scored for interface residues at 3.5 Ang distance, as per the core H2A-H2B dimer structure, and correlated mutations were identified using a multiple sequence alignment. The top scoring combinations of domains were then filtered by whether they are co-expressed in the same sub cellular location, and only the pairs of domains that fulfilled this criteria were considered as putative partners. The predictions were then tested using physical interactions reported for the same species, and superimposed with annotation using networks. Figure 4 shows the protein-protein interaction (PPI) network for NF-Y group in *Arabidopsis* revealing all ten NF-YB and NF-YCs.



Two interesting features were observed in the PPI network; one being the strong interaction of the NF-Y complex with histone modification proteins including several copies of histone deacetylases and histone methyl transferases. Second was the interaction of NF-Ys with ‘other’ non-histone HFM genes (see left most cluster in Fig. 4). These genes are interacting with all NF-Y components and were therefore misinterpreted as being NF-Ys themselves in earlier studies. However, our analysis revealed these to be DPE-C, DPE-3, DR1 and DRAP1 gene families. This resolution of ancestry was also supported by comprehensive phylogenetics (data not shown). The network further elucidates how various NF-Ys are involved in various developmental and cellular processes, including flowering seed germination and stress responses. Several HFM domains appear to have retained ancestral histone like features, as in case of chromatin remodeling and DNA repair and interaction with histone modifying enzymes, and the networks make it possible to explore the extent of divergence within and across HFM families. Pairwise dimerization potential has been well known in the

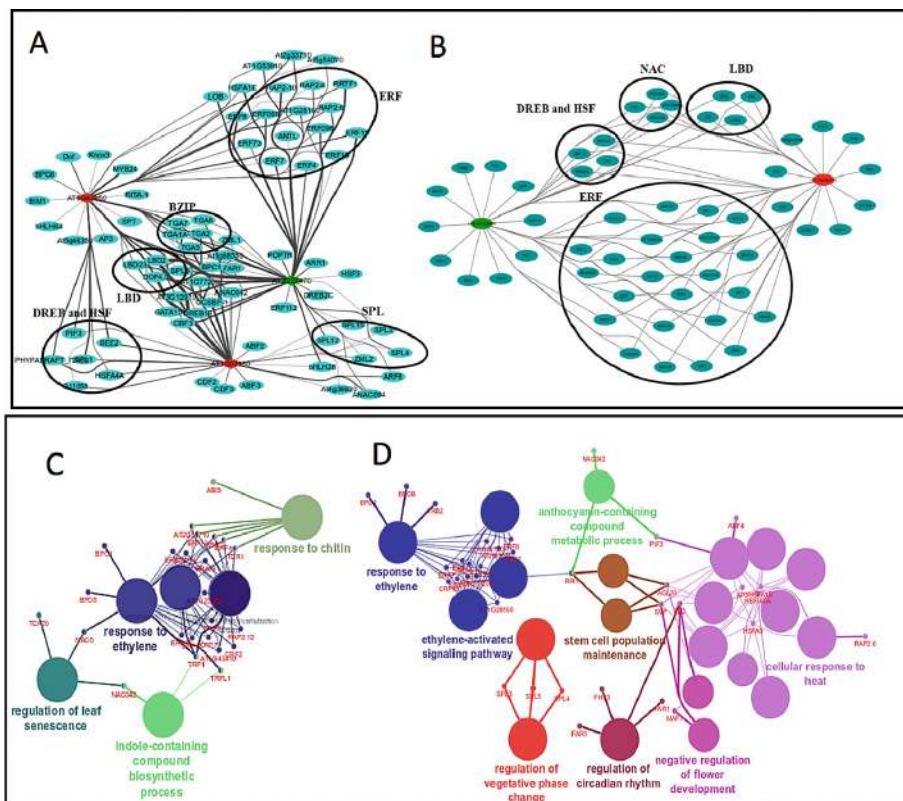


Fig. 5. Upper panel: Gene Regulatory network of DPE-3 (green) and DPE-C (red) in Arabidopsis (A) and Rice (B) Transcription factor families are marked in circles. Lower panels: Gene ontology networks of DR1/DRAP1 in (C) Rice and (D) Arabidopsis, where large nodes (and node colors) represent GO terms, while small nodes are transcriptional regulators.

NF-Y family, but has not been explored fully for the TAF family in plants while for the other families like CEN3 and DRAPs, interaction data has not even begun to be explored. To overcome this gap, we constructed and investigated detailed molecular networks for each HFM-based family in rice and Arabidopsis, and studied each one in detail. The representative ontology networks and gene regulatory networks (GRN) of TAF, DRAPs and DPE-C families of non-histone HFMs are depicted in separate panels of Fig. 5.

Panels A and B of Fig. 5 depict the GRNs of DPE-3 and DPE-C, and both share common transcription regulators suggesting pairwise pairing, core regulation and participation in common activities, as was observed in expression heat maps and PPI networks. In rice, both DPE-3 and DPE-C have single copy gene while in Arabidopsis there are two copies of DPE-C and one copy of DPE-3. The GRN in Fig. 5A suggests partial redundancy of both copies of DPE-C in Arabidopsis and it is possible that the distinct functions may still be evolving. DR1 and DRAP1 are known to interact with each other and form an active functional complex (Song et al. 2002) and we constructed the Gene Regulatory Networks of both of these HFM domains, by extracting promoter regions (1500 bp upstream and 200 bp downstream to the transcription start site), which were then subjected to *cis*-regulatory element prediction, followed by selection of true positive transcription factors, as described in methods. DR1 and DRAP1 promoters were found to be under the potential control of 79 transcription regulators in both Arabidopsis and rice, with GO enrichment networks depicted in Figs. 5C and D. The network suggests involvements of DR1-DRAP1 in hormone response, maintenance of meristematic cell number, regulation of vegetative phase change, circadian rhythm, flower development and heat stress. Heat responsive role of DR1-DRAP1 is also supported by expression heat maps, GRN and the PPI networks.

4 Conclusion

The Histone Fold Motif (HFM) is a conserved 70–90 amino acid sequence that has been variously identified in a large number of non-histone proteins in several organisms. In this work we report a network oriented analysis of about 4000 HFM containing non-histone proteins in 64 plant species. Annotation of the HFMs revealed several functional classes, some of which have been reported for the first time in this study, as homologs of HFM, for example CEN-S, DPE-3 and DPE-C. Visualization of the complete HFM dataset, including core histones and non-histone counterparts, their shared ancestry and taxonomic associations superimposed with functional annotations revealed extensive diversification and neo-functionalization of the HFM. In order to explore potential pairing preferences among non histone HFMs, all eight classes of protein families were investigated in detail in rice and Arabidopsis, two model plants representing the monocot and dicot lineages respectively.

Conceptualization of genetic and physical interaction data in the form of interconnected graphs or networks has provided insights into a more complete understanding of the non-histone HFMs. Mainly, three levels of analyses were conducted; (a) Transcriptome network analysis to understand the spatial and temporal expression levels as well as co-expression patterns of non-histone HFM genes, (b) Interactome

network analysis to understand the cross-talk of non-histone HFM proteins with other proteins in the plant, and (c) Regulatory data analysis to construct gene regulatory networks which elucidated potential pathways or processes in which the HFM containing non-histone domains are involved. These functional networks provided additional evidence for pairwise pairing partnerships, resolution of mis-annotations, and insights into divergence of HFMs into distinct, and often exclusive pathways.

Inferences made from one of the three networks (transcriptome/regulome/ interactome) were corroborated by observations in the corresponding networks of the other two kinds, elucidating several aspects of HFM diversification, pairwise recognition, and avenues for resolving mis-annotations that are being further investigated in our group.

Author Contribution, Funding and Acknowledgement. GY concieved the idea and AK planned and executed the work. This work was funded by the DBT-BTISNET and DST-SERB grant, AK obtained fellowship grant from UGC-CSIR and NIPGR. GY is funded by the DBT-Cambridge Lectureship and the GCRF TIGR2ESS Grant ID [BB/P027970/1TIGR2ESS]. Authors thank Director, NIPGR for support.

References

- Baxevanis, A.D., Arents, G., Moudrianakis, E.N., Landsman, D.: A variety of DNA-binding and multimeric proteins contain the histone fold motif. *Nucleic Acids Res.* **23**, 2685–2691 (1995). <https://doi.org/10.1093/nar/23.14.2685>
- Bellorini, M., Lee, D.K., Dantonel, J.C., Zemzoumi, K., Roeder, R.G., Tora, L., Mantovani, R.: CCAAT binding NF-Y-TBP interactions: NF-YB and NF-YC require short domains adjacent to their histone fold motifs for association with TBP basic residues. *Nucleic Acids Res.* **25**, 2174–2181 (1997). <https://doi.org/10.1093/nar/25.11.2174>
- Binda, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W. H., Pagès, F., Trajanoski, Z., Galon, J.: ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**(8), 1091–1093 (2009). <https://doi.org/10.1093/bioinformatics/btp101>
- Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., Robles, M.: Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**(18), 3674–3676 (2005). <https://doi.org/10.1093/bioinformatics/bti610>
- Gangloff, Y.G., Werten, S., Romier, C., Carré, L., Poch, O., Moras, D., Davidson, I.: The human TFIID components TAF(II)135 and TAF(II)20 and the yeast SAGA components ADA1 and TAF(II)68 heterodimerize to form histone-like pairs. *Mol. Cell. Biol.* **20**, 340–351 (2000). <https://doi.org/10.1128/MCB.20.1.340-351.2000>
- Hackenberg, D., Wu, Y., Voigt, A., Adams, R., Schramm, P., Grimm, B.: Studies on differential nuclear translocation mechanism and assembly of the three subunits of the arabidopsis thaliana transcription factor NF-Y. *Mol. Plant* **5**(4), 876–888 (2012). <https://doi.org/10.1093/mp/ssr107>
- Hoffmann, A., Chiang, C.-M., Oelgeschläger, T., Xie, X., Burley, S.K., Nakatani, Y., Roeder, R. G.: A histone octamer-like structure within TFIID. *Nature* **380**, 356–358 (1996)
- Kamada, K., Shu, F., Chen, H., Malik, S., Stelzer, G., Roeder, R.G., Meisterernst, M., Burley, S. K.: Crystal structure of Negative Cofactor 2 recognizing the TBP-DNA transcription complex. *Cell* **106**(1), 71–81 (2001). [https://doi.org/10.1016/S0092-8674\(01\)00417-2](https://doi.org/10.1016/S0092-8674(01)00417-2)

- Kumar, A.: Structural and functional evolution of plant gene families encoding proteins with histone fold motif (HFM). Ph.D. thesis (Submitted to JNU, India) (2019)
- Kumar, A., Yadav, G.: Diversification of the histone fold motif in plants: evolution of new functional roles. *Defence Life Sci. J.* **1**, 63–68 (2016). <https://doi.org/10.14429/dlsj.1.10061>
- Maere, S., Heymans, K., Kuiper, M.: BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* **21**(16), 3448–3449 (2005)
- Petroni, K., Kumimoto, R.W., Gnesutta, N., Calvenzani, V., Fornari, M., Tonelli, C., Holt, B.F., Mantovani, R.: The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. *Plant Cell* **24**(12), 4777–4792 (2012). <https://doi.org/10.1105/tpc.112.105734>
- Romier, C., Cocchiarella, F., Mantovani, R., Moras, D.: The NF-YB/NF-YC structure gives insight into DNA binding and transcription regulation by CCAAT factor NF-Y. *J. Biol. Chem.* **278**, 1336–1345 (2003). <https://doi.org/10.1074/jbc.M209635200>
- 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). <https://doi.org/10.1101/gr.1239303>
- Song, W., Solimeo, H., Rupert, R.a., Yadav, N.S., Zhu, Q.: Functional dissection of a Rice Dr1 / DrAp1 transcriptional repression complex. *Plant cell* **14**(January), 181–195 (2002). <https://doi.org/10.1105/tpc.010320>
- Werten, S., Mitschler, A., Romier, C., Gangloff, Y.-G., Thuault, S., Davidson, I., Moras, D.: Crystal structure of a subcomplex of human transcription factor TFIID formed by TATA binding protein-associated factors hTAF4 (hTAFII135) and hTAF12 (hTAFII20). *J. Biol. Chem.* **277**, 45502–45509 (2002). <https://doi.org/10.1074/jbc.M206587200>
- Xie, X., Kokubo, T., Cohen, S.L., Mirza, U.A., Hoffmann, A., Chait, B.T., Roeder, R.G., Nakatani, Y., Burley, S.K.: Structural similarity between TAFs and the heterotetrameric core of the histone octamer. *Nature* **380**, 316–322 (1996). Nature Publishing Group
- Yadav, G., Babu, S.: NEXCADE: perturbation analysis for complex networks. *PloS one* **7**(8), e41827 (2012)
- Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., Grussem, W.: GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. *Plant Physiol.* **136**(1), 2621–2632 (2004)



In-silico Gene Annotation Prediction Using the Co-expression Network Structure

Miguel Romero^(✉), Jorge Finke, Mauricio Quimbaya, and Camilo Rocha

Pontificia Universidad Javeriana Cali, Cali, Colombia

{miguel.romero,jfinke,maquimbaya,camilo.rocha}@javerianacali.edu.co

Abstract. Identifying which genes are involved in particular biological processes is relevant to understand the structure and function of a genome. A number of techniques have been proposed that aim to annotate genes, i.e., identify unknown biological associations between biological processes and genes. The ultimate goal of these techniques is to narrow down the search for promising candidates to carry out further studies through in-vivo experiments. This paper presents an approach for the in-silico prediction of functional gene annotations. It uses existing knowledge body of gene annotations of a given genome and the topological properties of its gene co-expression network, to train a supervised machine learning model that is designed to discover unknown annotations. The approach is applied to *Oryza Sativa Japonica* (a variety of rice). Our results show that the topological properties help in obtaining a more precise prediction for annotating genes.

Keywords: Co-expression network · Topological properties · *Oryza Sativa Japonica* · Machine learning · Functional gene annotation

1 Introduction

Available genome data has grown exponentially in the last decade, mainly due to the development of new technologies, including gene expression profiles generated with RNA sequencing [17]. Intuitively, genes are said to co-express whenever they are active simultaneously, indicating that they are associated to the same biological processes. Co-expression networks have been used widely to predict biological information (specific biological functions and processes) based on the interactions of the genes [16, 22, 24, 25]. The working hypothesis of correlated expression implying a relevant biological relationship has resulted in a promising strategy to perform functional genome annotation.

Co-expression networks are generally, represented as undirected, weighted graphs built from empirical data. Vertices denote genes and edges indicate a weighted relationship about their co-expression. Since co-expression networks include all correlated expression patterns between genes, a detailed analysis of the network topology –in addition to node-to-node relationships– may provides

insights into the network structure and organization. An approach based on co-expression networks ultimately provides additional information to build up novel biological hypotheses. It remains an open challenge to develop models that combine ideas from network theory and machine learning, and take advantage of the co-expression network structure for predicting functional gene annotations.

This paper presents an approach for predicting gene annotations based on the topological properties of the gene co-expression network of a given genome. The main idea is to combine the co-expression information available for the genome, the topological properties, and the body of known annotations (experimentally verified). The goal is to predict unknown annotations for genes. By taking advantage of the co-expression network structure, this approach aims to exploit additional information for the prediction that helps to establish strong functional associations between genes and the biological processes in which they are involved.

The proposed approach is showcased to predict gene annotations for the *Oryza Sativa Japonica* species, a variety of rice. The rice co-expression network is built from information available at ATTED-II [4] and a body of annotations gathered from RAP-DB [19]. The supervised machine learning technique XGBoost is used for the prediction of 141 functional gene annotations. For each annotation, a model with and without topological measures are trained. Their performance is compared to identify how topological measures can improve the annotation prediction in terms of precision. The experiments show that there are promising candidates to carry out further studies through in-vivo experiments, i.e., there exists set of genes that are consistently predicted to have a given annotation.

The remainder of the paper is organized as follows. Section 2 presents an overview on gene annotation and techniques used to perform functional genome annotation. Section 3 describes co-expression networks and a network-based approach for predicting gene annotation. Section 4 presents a case study for the *Oryza Sativa Japonica* species. Section 5 draws some conclusions and presents future research directions.

2 Gene Annotation

The goal of gene annotation is to determine the structural organization of a genome and discover sets of gene functions, i.e., the locations of genes and coding regions in a genome that determine what genes do [18, 26]. Once a genome is sequenced, it is annotated to understand its structure and how it encodes biological function. Though several organism have been completely sequenced, genome annotation remains a significant challenge, mainly due to its extreme combinatorial nature.

Genome annotation focuses on two complementary processes. First, the genome structure is defined, that is genes are identified and intergenic regions are characterized from specific sequences that are associated with genomic structures (particular promoter motifs or repetitive signatures). Second, putative functions of genes are assigned to establish gene and, as a whole, genome functional characterizations [10]. While genomic structure can be determined by the detection

of specific genomic elements inserted in the sequence itself, genome functional annotation is more laborious. It generally depends on several annotation strategies that combine alignment-based information with experimental evidence associated with gene functional predictions. Often extensive in-vivo experimentation is required to gain certainty on processes associated to genes [28]. The rapid accumulation nowadays of genome-wide data describing both, genome sequences and functional properties of genes, has facilitated the novel development of integrative approaches to target genome annotation.

Global analysis of similarity in gene expression patterns has been used to infer specific regulatory networks by analysis of gene co-expression analysis. Different techniques and tools, mostly supported by statistical inference, have been proposed to suggest putative biological processes to genes whose functional annotation is partially or completely unknown [14, 25].

3 Prediction Based on Co-expression Network Structure

Here gene co-expression networks are represented as undirected graphs where each vertex identifies a gene and an edge the level of co-expression between two genes.

Definition 1. Let V a set of genes, E a set of edges that connect pairs of genes and w a weight function. A (weighted) gene co-expression network is a weighted graph $G = (V, E, w : E \rightarrow \mathbb{R}_{\geq 0})$.

The set of genes V in a co-expression network is particular to the genome under study. The correlation of expression profiles between each pair of genes is measured, commonly, with help of the Pearson correlation coefficient. Every pair of genes is assigned and ranked according to a relationship measure, and a threshold is used as a cut-off measure to determine E . The weight function w denotes how strongly co-expressed are each pair of genes in V . For any pair of genes $u, v \in V$, $w(u, v)$ is usually inversely proportional to the measure of *mutual rank* (MR) between genes u and v . Note that a value of 0 would be assigned to the strongest connections [13].

There are gene co-expression network databases containing several expression profiles obtained from cDNA microarrays and RNA sequencing. Each profile indicates how gene expression is perturbed when the subject organism is exposed to multiple types of stress (for example, to biotic and abiotic stresses). The correlation of expression for a set of genes under multiple conditions may suggest their functional relation, thus offering information on how genes can be related in terms of biological function.

In an annotated gene co-expression network, each gene is associated with the collection of biological functions to which it is related (e.g., through in-vivo experiments).

Definition 2. Let A be a set of biological functions. An annotated gene co-expression network is a gene co-expression network $G = (V, E, w)$ complemented with an annotation function $\phi : V \mapsto 2^A$.

The network-based approach to gene annotation proposed in this paper can now be explained in detail. Given an annotated co-expression network $G = (V, E, w)$ with annotation function ϕ , the goal is to use the information represented by ϕ together with topological properties of G to obtain a function $\psi : S \mapsto 2^A$. Function ψ predicts associations between annotations and genes based on a supervised machine learning technique.

The overall success of this approach is evaluated in two complementary ways. On the one hand, this approach would be successful if higher precision is achieved for a suggestion $a \in \psi(v)$ to annotate gene $v \in V$, when compared to other approaches (e.g., to a suggestion in which only the information represented by ϕ and the edge structure of G are taken into account). On the other hand, this approach would also be successful if genes $v \in V$ are found for which $\psi(v) \setminus \phi(v) \neq \{\}$, meaning that new annotation suggestions have been found for the candidate gene v . This latter situation is desirable in practice to reduce time and costs associated to laboratory experimentation. In particular, for a biological function $a \in \psi(v)$, with $\psi(v) \setminus \phi(v) \neq \{\}$, laboratory experimentation can then increase the focus on specific biotic and abiotic stresses to see if gene v is actually associated to the biological function a in the genome under study.

4 In-silico Experimentation with *Oryza Sativa Japonica*

This section describes an in-silico experimentation case study of gene annotation prediction for *Oryza Sativa Japonica* (Osa). It follows the network-based approach proposed in Sect. 3, and explains how the gene co-expression network is built, how it is initially annotated, and how –with the help of topological properties– machine learning techniques are used to improve gene annotation.

4.1 The Co-expression Network and Gene Annotations

The co-expression information used in this paper is taken from the ATTED-II database [4, 12, 15]. The gene co-expression network $G = (V, E, w)$ comprises 19 665 vertices (genes) and 553 125 edges. The weight function $w : E \mapsto \mathbb{R}_{\geq 0}$ measures the mutual rank (MR) between any pair of genes; it assigns smaller values to stronger links. A MR threshold of 100 is used as the cut-off measure for G , i.e., E contains edges e that satisfy $w(e) \leq 100$.

The annotation information for G is taken from the RAP-DB [19] database, a comprehensive set of gene annotations for the genome of rice. Among these annotations, there are 899 for molecular function (i.e., molecular activities of individual gene products), 187 for cellular components (i.e., location of the active gene products), and 633 for biological processes (i.e., pathways to which a gene contributes). It is important to note that genes may be associated to several annotations in each category. Since this work is mainly focused on pathways and large processes, only biological process annotations are considered. The annotation function $\phi : V \rightarrow 2^A$ for G associates $|A| = 615$ annotations to $|V'| = 7478$ genes, where $V' \subseteq V$ is the set of genes associated to at least one biological process.

4.2 Topological Properties

Given the co-expression network $G = (V, E, w)$, properties of its network structure are computed for gene annotation prediction. The topological measures considered for each gene u are the following:

- degree: number of edges incident to u ;
- eccentricity: maximum shortest distance from u to any vertex in its connected component;
- clustering coefficient: ratio between the number of triangles (3-loops) that pass through u and the maximum number of 3-loops that could pass through it;
- closeness centrality: the reciprocal of the average shortest path length from u ;
- betweenness centrality: the amount of control that u has over the interactions of other nodes in the network;
- neighborhood connectivity: the average connectivity of all neighbors of u ;
- topological coefficient: the extent to which u shares neighbors with other nodes.

These measures were computed with the help of Cytoscape [21], an open source platform for visualizing and analyzing molecular interaction networks and biological pathways.

4.3 Supervised Training

Two models are trained for predicting gene annotations, one per biological function. Namely, one in which the topological measures of G are used and another one in which they are not. The next paragraphs describe how these models are built, trained, and evaluated.

The dataset summarizes data for the 19 665 genes, 615 annotations, and 7 topological measures. It comprises 19 665 rows and 222 columns. For these experiments, the dataset is heavily imbalanced since 77% of the annotations are related to less than 10 genes each one. In order to counter such an imbalance, two decisions are made. First, only annotations associated with at least 10 genes are considered for prediction, reducing the number of annotations from 615 to 141. Second, the Synthetic Minority Over-sampling TTechnique (SMOTE) is used to over-sample the minority class to potentially improve the performance of a classifier without loss of data [7]. This technique derives the new samples of the minority class from interpolation rather than extrapolation, in order to avoid over-fitting problems.

A supervised machine learning technique for the annotation prediction is used. In particular, the XGBoost implementation of gradient boosted decision trees is used [8]. This technique has recently been dominating applied machine learning competitions for structured or tabular data, and it has implementations in many programming languages, including C++, Java, and Python. In the experiments presented in this section, a Python implementation was used.

Finally, k -fold cross validation is used in the training of the two models with $k = 50$. The number of k is determined for statistical significance in the false positive analysis prediction. The performance of the models is compared using the accuracy, F1-score, and AUC ROC measures.

4.4 Annotation Prediction

Figure 1 presents a summary of the accuracy achieved by the two trained models for predicting gene annotations. In particular, the results are depicted for 32 different annotations. The annotations are sorted in descending order by the

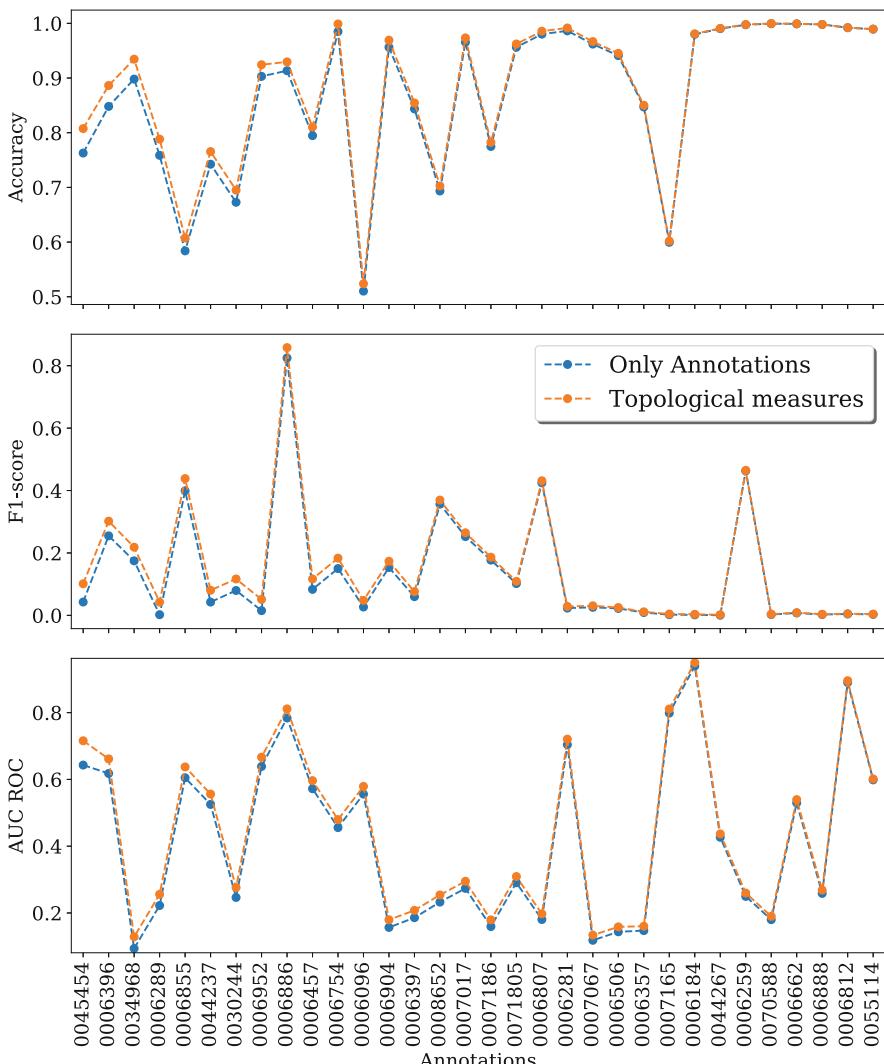


Fig. 1. Performance accuracy, F1-score, and AUC ROC measures for the prediction of 32 annotations with the two trained models (with and without topological measures).

Table 1. Number of genes most frequently annotated as false positives for the 32 annotations by the model trained with topological measures. The ‘Max FP’ column summarizes the number of times (out of a total of 50) such an annotation is suggested for a gene, while the ‘FP’ column identifies the number of genes that are consistently given such an annotation.

ID	Biological process	# Genes	Max FP	# FP
0006807	Nitrogen compound metabolic process	15	41	1
0006289	Nucleotide-excision repair	20	46	1
0006397	mRNA processing	17	48	1
0007017	Microtubule-based process	18	49	1
0070588	Calcium ion transmembrane transport	10	36	1
0006184	GTP catabolic process	49	47	1
0044267	Cellular protein metabolic process	25	49	1
0007186	G-protein coupled receptor protein signaling	11	50	1
0006281	DNA repair	62	50	2
0006754	ATP biosynthetic process	24	49	3
0006904	Vesicle docking involved in exocytosis	11	50	4
0055114	Oxidation-reduction process	870	47	5
0006886	Intracellular protein transport	135	50	19
0006855	Drug transmembrane transport	32	50	21
0006662	Glycerol ether metabolic process	28	50	27
0006888	ER to Golgi vesicle-mediated transport	16	50	29
0006259	DNA metabolic process	15	50	32
0007067	Mitosis	11	48	33
0008652	Cellular amino acid biosynthetic process	18	50	52
0030244	Cellulose biosynthetic process	23	50	64
0034968	Histone lysine methylation	11	50	93
0006812	Cation transport	62	50	96
0045454	Cell redox homeostasis	83	49	103
0006506	GPI anchor biosynthetic process	12	50	284
0007165	Signal transduction	104	50	370
0071805	Potassium ion transmembrane transport	24	50	570
0006357	Regulation of transcription from RNA polymerase	12	50	1199
0006396	RNA processing	58	50	1212
0044237	Cellular metabolic process	75	50	1318
0006457	Protein folding	162	50	2358
0006952	Defense response	133	50	2679
0006096	Glycolysis	50	50	2875

performance difference between the models. This plot shows that the model trained with the additional information of the topological measures can be more reliable in these cases. The results for the remaining annotations are omitted,

but in these other cases the additional information provided by the network structure does not result in a better prediction performance.

Table 1 presents details about the prediction made by the model trained with the topological data of the co-expression network. The annotations listed in this table correspond to the same 32 annotations included in Fig. 1. For each annotation, the table includes its gene ontology term (ID), its associated biological process, and the total number of genes known to be associated with it in the co-expression network. A false positive analysis is applied to the annotation predictions: the idea is to identify the genes that tend to be classified as a false positive because they are the candidate genes on which lab experimentation can focus on. The ‘Max FP’ column summarizes the number of times (out of a total of 50) such an annotation is suggested for a gene, while the ‘FP’ column identifies the number of genes that are consistently given such an annotation.

Note that the annotations in Table 1 are sorted in ascending order by the number of genes most frequently classified as false positive. This set of genes is considerably small for the first 12 annotations of the table and therefore can be seen as good candidates for experimental verification. For example, the only gene associated to the nitrogen compound metabolic process (0006807) is proline dehydrogenase, identified as Os10g0550900, which is related to the functional annotation proline catabolic process to glutamate (0010133).

5 Related Work and Conclusion

Complex network structure has been widely used for the enrichment of analysis techniques from different perspectives and in different domains. A modest summary of the enormous body of work for, mainly, biological predictions is presented next.

The application of networks in biology has grown exceptionally in the last decade due to the large amount of molecular interaction data available [5]. There are two main types of biological networks that are a current focus of research. The first group is of molecular networks. It includes protein interaction networks where proteins are represented as vertices that are connected by physical interactions, metabolic networks where metabolites are vertices connected by co-appearance in biochemical reactions, regulatory networks where connections are regulatory relationships between transcription factors and genes, and RNA networks. The second group is of genetic networks. It includes co-expression networks, in which genes are vertices that are connected by similar expression patterns, such as the ones studied in the present work. This latter group has been used to identify the function of a large set of genes and their role in specific biological processes in different species [5, 25]. In particular, the co-expression networks have helped to address the problem of identifying the role of the genes in biological systems. This work takes a step forward in exploiting the rich structural property of a network with the goal of increasing annotation prediction precision.

Studying the link prediction problem is one of the most common applications of the topological properties of networks. Tan et al. [23] examine the role of

network topology in predicting missing links from the perspective of information theory. They present a practical approach based on the mutual information of network structures. Naaman et al. [11] show that edges with similar network topology, as defined by a combination of network measures having similar signs, can be used to predict edge sign based on correlation measures on the network topology.

In biological networks, the topological properties have been used for the prediction of interactions between genes or proteins. Lobato et al. [2,3] use the topology of biological interactomes for the prediction of interactions in biological networks. Santolini et al. [20] use biochemical networks, with experimentally measured kinetic parameters, to predict the impact of perturbation patterns in biological interactome networks. They approximate perturbation patterns using increasingly accurate topological models. Benstead-Hume et al. [6] explore computational approaches to identify genes that have become essential in individual cancer cell lines. They use machine learning techniques, the protein-protein interaction network, and the network topology to classify genes that can be essential to human cancer processes.

Within the broader picture of network-based analysis techniques, there is some recent work in other domains. For instance, Abeysinghe et al. [1] study the topological properties of real-world electricity distribution networks at the medium voltage level by employing the techniques from complex networks analysis and graph theory. Jiang et al. [9] use large urban street networks for topological analysis and show that these networks have the small-world property, but do not exhibit scale-freeness. Zhang et al. [27] use topological properties to better understand bus networks in large cities to optimize the bus lines and transfers.

This paper presented a network-based approach for annotation prediction of genes. It uses the information of the co-expression network of the genome under study to build a predictive model using machine learning techniques. When trained with the topological measures as part of the data set, the model is shown by a series of in-silico experiments to improve the accuracy, F1-score, and AUC ROC in comparison to a model trained without the topological measures of the co-expression network. Each pair of models is trained for predicting a particular gene annotation. By measuring the number of genes most frequently classified as false positive by the prediction model, a small number of genes is identified for 12 biological processes in *Oryza Sativa Japonica*: these genes are good candidates for experimental verification.

As usual, significant work remains to be done. A next step is to perform experimental evaluation in the laboratory to validate the in-silico predictions. Also, more in-silico experimentation can be used to predict gene annotations in other species, such as sugar cane.

Acknowledgments. The authors would like to thank the anonymous referees for their helpful comments. This work was funded by the OMICAS program: Optimización Multiescala In-silico de Cultivos Agrícolas Sostenibles (Infraestructura y Validación en Arroz y Caña de Azúcar), sponsored within the Colombian Scientific Ecosystem by

The World Bank, Colciencias, Icetex, the Colombian Ministry of Education and the Colombian Ministry of Industry and Turism under Grant FP44842-217-2018.

References

1. Abeysinghe, S., Wu, J., Sooriyabandara, M., Abeysekera, M., Xu, T., Wang, C.: Topological properties of medium voltage electricity distribution networks. *Appl. Energy* **210**, 1101–1112 (2018)
2. Alanis Lobato, G.: Exploitation of complex network topology for link prediction in biological interactomes (2014)
3. Alanis-Lobato, G., Cannistraci, C.V., Ravasi, T.: Exploitation of genetic interaction network topology for the prediction of epistatic behavior. *Genomics* **102**(4), 202–208 (2013)
4. Aoki, Y., Okamura, Y., Tadaka, S., Kinoshita, K., Obayashi, T.: ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. *Plant Cell Physiol.* **57**(1) (2016)
5. Barabási, A.-L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**(1), 56–68 (2011)
6. Benstead-Hume, G., Wooller, S.K., Dias, S., Woodbine, L., Carr, A.M., Pearl, F.M.G.: Biological network topology features predict gene dependencies in cancer cell lines. *Systems Biology* (2019, preprint)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
8. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
9. Jiang, B., Claramunt, C.: Topological analysis of urban street networks. *Environ. Plan.* **31**(1), 151–162 (2004)
10. Mudge, J.M., Harrow, J.: The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.* **17**(12), 758–772 (2016)
11. Naaman, R., Cohen, K., Louzoun, Y.: Edge sign prediction based on a combination of network structural topology and sign propagation. *J. Complex Netw.* **7**(1), 54–66 (2019)
12. Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., Kinoshita, K.: ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. *Plant Cell Physiol.* **59**(1) (2018)
13. Obayashi, T., Kinoshita, K.: Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.* **16**(5), 249–260 (2009)
14. Obayashi, T., Kinoshita, K.: COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Res.* **39**(Database), D1016–D1022 (2011)
15. Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Aoki, Y., Shirota, M., Kinoshita, K.: ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. *Plant Cell Physiol.* **55**(1) (2014)
16. Oti, M., van Reeuwijk, J., Huynen, M.A., Brunner, H.G.: Conserved co-expression for candidate disease gene prioritization. *BMC Bioinform.* **9**(1), 208 (2008)
17. Ranganathan, S., Gribskov, M.R., Nakai, K., Schönbach, C.: Encyclopedia of Bioinformatics and Computational Biology (2019). OCLC: 1052465484

18. Rust, A.G., Mongin, E., Birney, E.: Genome annotation techniques: new approaches and challenges. *Drug Discov. Today* **7**(11), S70–S76 (2002)
19. Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C., Iwamoto, M., Abe, T., Yamada, Y., Muto, A., Inokuchi, H., Ikemura, T., Matsumoto, T., Sasaki, T., Itoh, T.: Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* **54**(2) (2013)
20. Santolini, M., Barabási, A.-L.: Predicting perturbation patterns from the topology of biological networks. *Proc. Natl. Acad. Sci.* **115**(27), E6375–E6383 (2018)
21. Shannon, P.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**(11), 2498–2504 (2003)
22. Stuart, J.M.: A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302**(5643), 249–255 (2003)
23. Tan, F., Xia, Y., Zhu, B.: Link prediction in complex networks: a mutual information perspective. *PLoS ONE* **9**(9), e107056 (2014)
24. van Dam, S., Võsa, U., van der Graaf, A., Franke, L., de Magalhães, J.P.: Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings Bioinform.* (2017)
25. Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L., Van de Peer, Y.: Unraveling transcriptional control in arabidopsis using cis-regulatory elements and coexpression networks. *Plant Physiol.* **150**(2), 535–546 (2009)
26. Yandell, M., Ence, D.: A beginner’s guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**(5), 329–342 (2012)
27. Zhang, H., Zhao, P., Gao, J., Yao, X.-M.: The analysis of the properties of bus network topology in Beijing basing on complex networks. *Math. Problems Eng.* **1–6**, 2013 (2013)
28. Zhou, Y., Young, J.A., Santrosyan, A., Chen, K., Yan, S.F., Winzeler, E.A.: In silico gene function prediction using ontology-based pattern identification. *Bioinformatics* **21**(7), 1237–1245 (2005)

Network Neuroscience



Linear Graph Convolutional Model for Diagnosing Brain Disorders

Zarina Rakhimberdina^(✉) and Tsuyoshi Murata

Tokyo Institute of Technology, Tokyo, Japan
zarina.rakhimberdina@net.c.titech.ac.jp, murata@c.titech.ac.jp
<http://www.net.c.titech.ac.jp>

Abstract. Deep learning models find an increasing application in the diagnosis of brain disorders. Designed for large scale datasets, deep neural networks (DNNs) achieve state-of-the-art classification performance on a number of functional magnetic resonance imaging (fMRI) data. While utilizing DNNs might improve the performance, the complexity of the learning function decreases the interpretability of the model. Moreover, DNNs require considerably more time to train compared to their linear predecessors. In this paper, we re-examine the use of deep graph neural networks for graph-based disease prediction in favor of simpler linear models. We present a simplified linear model, which is more than 10 times faster to train than the previous DNN counterparts. We test our model on three fMRI datasets and show that it achieves comparable or superior performance to the state-of-the-art methods.

Keywords: Graph Convolutional Network · Brain functional connectivity

1 Introduction

Deep learning is revolutionizing data analysis tools in many domains [9, 12] including neuroscience. In particular, deep neural networks (DNNs) have recently gained significant attention in the analysis of brain connectivity patterns acquired through functional magnetic resonance imaging (fMRI). fMRI technique is based on the fact that neural activity and blood flow in the brain are correlated: activation of a region in the brain requires an increase in blood flow to that region [3]. In this way, fMRI captures patterns of brain activations under various physical factors, cognitive states or brain disorders [1, 3, 4].

Graphs are widely used to model complex interaction patterns extracted from imaging data [2, 3]. The diversity of graph-based models proposed earlier in the literature can be categorized into two classes, based on the way the nodes are defined. The first class of models (Fig. 1(a)) uses graphs to describe structural or functional connectivity of the human brain on an individual level [3]. In other words, nodes represent brain regions and edges represent functional correlations between time series of those regions [7, 8, 15]. As a result, each constructed graph

corresponds to one subject and further analysis is performed using graph comparison metrics. The second class of models (Fig. 1(b)) involves the construction of a population graph, a structure composed of the entire set of human subjects and representing interindividual connectivity between them. In this approach, each subject is modeled as a node with corresponding brain-connectivity data, and each edge is defined based on the similarity between subjects' phenotypic features (age, gender, handedness, etc.) [13]. This property of incorporating both imaging and non-imaging features made population graph models effective for brain disorder classification [13, 14].

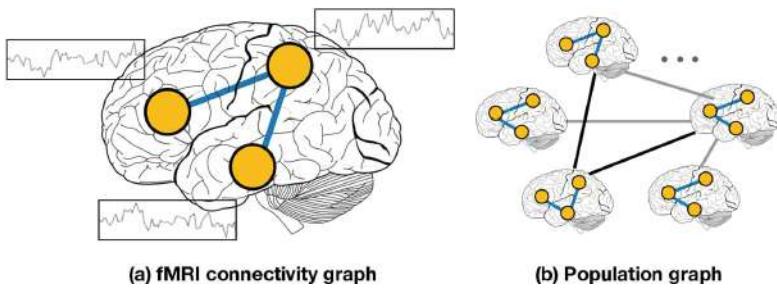


Fig. 1. Graph-based approaches of modeling with fMRI data.

Over the past few years, Graph Convolutional Network (GCN) has been exploited as a powerful method for processing graph structure. The strength of GCN originates from combining local node-level context and global neighborhood-level context [5, 10]. Similar to standard 2D image convolution [11], which uses a rectangular filter over the values of the neighboring pixels, graph convolution aims to aggregate the neighborhood information of nodes in a graph using graph Laplacian [10]. [14] were among the first to propose an application of Graph Convolutional Networks (GCN) on population-based brain disorder diagnosis task. In [14], authors combined imaging (fMRI scans) and non-imaging (phenotypic data) data by representing subjects as nodes, and pairwise similarities between them as edges. Further, the authors formulated the task of subject classification as node labeling over the population graph.

Despite the improved performance of the GCN model, the application of such deep neural network introduces unnecessary complexity for relatively small population graphs (less than 1,000 subjects) considered in this and the related works. More specifically, due to its multilayer non-linear structure, GCN is required to learn more parameters corresponding to each layer. This results in GCN's reduced performance in terms of time and test accuracy. Another limitation of GCN is overfitting to high-dimensional input, which requires additional data preprocessing [13]. To overcome this limitation, [13] employed several dimensionality reduction methods including ridge classifier, PCA, MLP, and Autoencoder. While the dimensionality reduction improves the performance of multilayer

models, it also exposes their limitation of using handcrafted features restraining the model from learning optimal feature representation from raw data.

In this paper, we propose a simplified and efficient framework, which is based on the concept of simple graph convolutions (SGC) for population-based disease prediction. By using a linear counterpart of the earlier GCN model, we propose a simple yet powerful graph convolutional model, which does not require intermediate preprocessing pipelines. Through an extensive evaluation over three fMRI databases on a binary classification task of diagnosing a brain disorder, we show that the linear model proposed in this work achieves comparable or superior performance to the state-of-the-art graph model proposed by [13]. In particular, the following are the contributions of this work:

- We introduce the novel concept of linear graph convolutions or simple graph convolutions (SGC)[17] for brain disorder classification. To the best of our knowledge, this is the first work to propose a linear end-to-end alternative to the multilayer GCN model used in earlier works.
- We propose a new straightforward graph construction pipeline by defining the weights of the edges as a hamming distance between subjects' phenotypic features.
- Further, we experimentally show the ability of our proposed model to generalize well across different datasets by testing on three publicly available fMRI databases: ABIDE, COBRE, and ADHD-200.
- Our model outperforms GCN-based baselines on a binary classification task both in terms of accuracy and computational time. The model achieves an accuracy of 68.56% on the ABIDE dataset and state-of-the-art performance on COBRE and ADHD-200 datasets (80.55% and 74.35% respectively). In terms of training time, our model provides up to 50 times speedup in computation.

The performance of the proposed linear model with the new edge weighting method is evaluated by comparing the classification accuracy results with the existing baseline methods¹.

2 Datasets and Preprocessing

In this section, we describe three publicly available fMRI datasets used in our experiments (see Table 1). Evaluation on multiple datasets aims to demonstrate the generalizability of our model to different databases, which are acquired at multiple locations using diverse fMRI protocols. In all of the three datasets, we pursue a common goal: to classify healthy control subjects and pathological subjects.

The Autism Brain Imaging Data Exchange (ABIDE)² database combines structural and functional MRI data of 1,112 subjects acquired at 20 imaging

¹ Source code at <https://github.com/zarina-aniraz/linear-graph-convolution>.

² <http://preprocessed-connectomes-project.org/abide/>.

sites. To achieve a fair comparison with the state-of-the-art graph convolutional network, we use the subset of 871 subjects and fMRI preprocessing described in [13]. The dataset consists of 403 patients with Autism Spectrum Disorder (ASD) and 468 healthy individuals. We use both imaging and non-imaging data provided in ABIDE. Imaging data from fMRI scans is represented by a connectivity matrix between 111 regions of interests (ROIs) in the brain (we refer to [13] for more details). Non-imaging data corresponds to phenotypic features, such as gender, age, and imaging site.

The Center for Biomedical Research Excellence (COBRE)³ dataset provides anatomical and functional MR data of 147 subjects. After fMRI prepossessing and extracting ROI-to-ROI correlation matrices with CONN functional connectivity toolbox [16], 71 patients diagnosed with Schizophrenia and 74 healthy subjects were selected from each category. Phenotypic information accompanying the dataset includes diagnosis, age, gender, and handedness.

The ADHD-200⁴ database, dedicated to a study of the Attention Deficit Hyperactivity Disorder (ADHD), contains 973 resting-state fMRIs aggregated across 8 independent imaging sites. 585 of scans were gathered from typically developing individuals and 388 from patients with ADHD. To obtain a balanced dataset, we selected a subset of 714 fMRI scans with an equal proportion from both subject categories. In addition to resting-state fMRI, each participant's data includes individual phenotypic data. Similarly to ABIDE, fMRI scans were acquired from multiple sites.

For each of the datasets, we used imaging data, which is fMRI connectivity matrices, and non-imaging data defined by phenotypic measures to construct a population graph. Phenotypic features were used to measure the similarity between subjects and to construct weighted edges, while imaging data was used as nodes' feature vectors. Graph construction is discussed in more detail in the Sect. 3.1.

Table 1. The statistics of the three datasets used in this work.

	ABIDE	COBRE	ADHD-200
Total subjects	871	145	714
Patients	403	71	357
Healthy controls	468	74	357
Female/Male	144/727	37/108	241/473
Age range	6.47–58	18–65	–
Age mean	16.94	36.90	–
Handedness (Right/Left)	–	131/14	694/20
Sites	20	1	7/8

³ http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html.

⁴ http://fcon_1000.projects.nitrc.org/indi/adhd200/index.html.

3 Methods

3.1 Network Construction

In this section, we define population graph construction using a set of healthy controls and pathological subjects. We construct an undirected weighted graph $G = (V, E, \mathcal{E})$, where the set of nodes $V = \{v_1, \dots, v_n\}$ corresponds to a set of healthy and pathological subjects. The set of edges $E \subseteq V \times V$ corresponds to links between these nodes, and $\mathcal{E} : E \mapsto \mathbb{R}$ is a function which assigns a weight to each edge $e \in E$. In addition, each node v_i is associated with a D -dimensional feature vector \mathbf{f}_i extracted from fMRI imaging data. A feature matrix $\mathbf{F} \in \mathbb{R}^{n \times D}$ consists of stacked feature vectors of n nodes in the graph. Our goal is to predict subject classes (0 - pathological or 1 - healthy) using node classification task. We use the population graph as an input to the proposed linear graph convolution model, which performs binary classification on each node of the graph.

To ensure a fair comparison with baseline models, we follow the graph construction pipeline proposed in [13] with two major improvements. First, we design an end-to-end model that eliminates the need for any intermediary feature extraction pipelines. We use the original D -dimensional feature vectors that represent fMRI connectivity matrices of each subject. This step was possible due to the linearity of the convolutional model: it allows efficient computations using original high dimensional feature matrices without extra memory or time requirements (more detail on the model in Sect. 3.2).

Our second improvement concerns edge construction and edge weight assignment. We design an intuitive and straightforward edge weighting function for categorical values of phenotypic features based on Hamming distance⁵. The edge weighting function \mathcal{E} for any two nodes v_i and v_j in the graph is defined as follows:

$$\mathcal{E}(v_i, v_j) = m - \|\mathbf{f}_i - \mathbf{f}_j\|_H \quad (1)$$

where m is the length of the phenotypic feature vector and $\|\mathbf{f}_i - \mathbf{f}_j\|_H$ is hamming distance between phenotypic feature vectors \mathbf{f}_i and \mathbf{f}_j . The smaller the Hamming distance is between the feature vectors of two nodes $\|\mathbf{f}_i - \mathbf{f}_j\|_H$, the greater weight is assigned to the edge between them. For example, given two subjects v_1 and v_2 with feature vectors of $[0, 0, 3]$ and $[1, 0, 5]$, where column values correspond to gender (0 - male, 1 - female), handedness (0 - right, 1 - left) and site number, the weight assigned to the edge (v_1, v_2) between the subjects is $\mathcal{E}(v_i, v_j) = 3 - \|1, 0, 2\|_H = 1$. Note that for age feature, numerical values are first converted to corresponding categorical values, i.e age groups. To construct the population graph for ABIDE, COBRE, and ADHD-200, we use available phenotypic features. For ABIDE, they include gender, age group, and acquisition site. For COBRE, the features are gender, age group, and handedness, and the ADHD-200 graph is constructed using gender, handedness and acquisition site features.

⁵ The Hamming norm $\|\mathbf{v}\|_H$ is defined as the number of non-zero entries of vector \mathbf{v} .

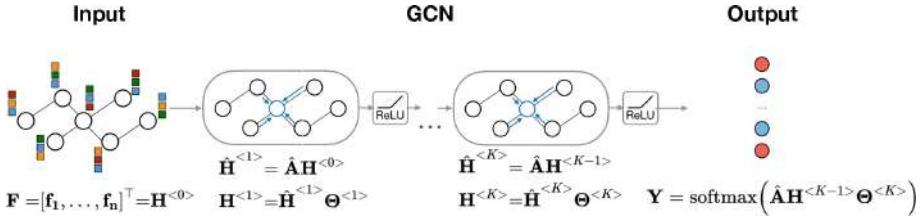


Fig. 2. Schematic pipeline for GCN.

3.2 Subject Classification Using Graph Neural Networks

Multilayer Graph Convolutional Network (GCN) is considered to be the state-of-the-art method for learning population graph representations [6, 13]. Inspired by the convolutional neural network in the image processing domain, GCN adopts similar neural network architecture by stacking feature propagation layers followed by a non-linear activation function. However, the application of a multi-layer neural network results in greater complexity due to non-linear transformations [6, 17].

In this work, we exploit a simple linear graph convolutional model proposed in [17] for brain disorder classification. The simplicity of the linear graph convolutional model is achieved by removing non-linear activations between layers, and aggregating weight matrices corresponding to each layer into a single matrix. Advantages of using the linear model called SGC [17], which stands for Simple Graph Convolutions, include natural interpretability, scalability for large graphs and computational speedup [17]. The interpretability of the model is achieved due to the linear logistic regression, i.e. the association between features and the output is modeled linearly. As far as we know, this is the first work to exploit the concept of linear graph convolution in the context of brain disorder classification. To demonstrate the simplicity of our method, we first describe the state-of-the-art GCN-based computational framework proposed by [13] and later describe our approach.

Graph Convolutional Networks. In the context of population-based brain disorder diagnosis, Graph Convolutional Network (GCN) takes adjacency matrix representation \mathbf{A} of population graph $G = (V, E, \mathcal{E})$ as an input. The element a_{ij} of symmetric adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ corresponds to edge weight between nodes v_i and v_j . The graph is annotated with feature matrix $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_n]^\top$, where each row $\mathbf{f}_i \in \mathbb{R}^D$ corresponds to a D -dimensional fMRI connectivity of node v_i . The feature matrix \mathbf{F} is used as an input feature matrix to the first layer of GCN, i.e. $\mathbf{H}^{<0>} = \mathbf{F}$. The output of the model is binary matrix $\mathbf{Y} \in \mathbb{R}^{n \times C}$, where each row \mathbf{y}_i denotes the probability of node v_i belonging to one of the $C = 2$ classes, healthy or pathological (see Fig. 2).

Convolutional operation on a graph structure in layer l is performed by multiplying adjacency matrix \mathbf{A} and the input feature matrix $\mathbf{H}^{<k-1>}$ from the

previous layer $k - 1$: $\mathbf{H}^{<k>} = \mathbf{A}\mathbf{H}^{<k-1>}$. One multiplication operation $\mathbf{A}\mathbf{H}^{<k>}$ averages feature vectors of nodes within the one-hop neighborhood. By stacking K layers, GCN aims to aggregate K-hop neighborhood information from node features. In addition to performing graph convolution, each layer k linearly transforms hidden feature representation by a weight matrix $\Theta^{<k>}$. Then, a non-linear activation such as ReLU is applied to each layer, resulting in the following forward propagation formula for layer k :

$$\mathbf{H}^{<k>} = \text{ReLU}(\hat{\mathbf{A}}\mathbf{H}^{<k-1>}\Theta^{<k>}), \quad (2)$$

where $\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A} + \mathbf{I})\mathbf{D}^{-\frac{1}{2}}$ is normalized adjacency matrix containing self-loops and \mathbf{D} is a degree matrix of \mathbf{A} . Finally, class labels for each node in the graph are computed using the final K -th softmax layer:

$$\mathbf{Y}_{GCN} = \text{softmax}(\hat{\mathbf{A}}\mathbf{H}^{<K-1>}\Theta^{<K>}) \quad (3)$$

The computational complexity of two-layer GCN is linear in the number of graph edges $\mathcal{O}(|E|DMC)$, where M is hidden layer dimensionality. Memory requirement using a sparse representation for $\hat{\mathbf{A}}$ is linear in the number of edges $\mathcal{O}(|E|)$.

Simplified Graph Convolutions (SGC). SGC is the simplest formulation of a graph convolutional model which was introduced to better understand and explain the mechanisms of GCN. [17] also showed that the multilayer Graph Convolutional Neural (GCN) network primarily benefits from local averaging achieved by a graph convolution operation. Therefore, after removing non-linear layers from the Eq. (2) and collapsing the repeated multiplications with normalized adjacency matrix $\hat{\mathbf{A}}$ into a single matrix $\hat{\mathbf{A}}^K$, the resulting linear model becomes as follows:

$$\mathbf{Y}_{SGC} = \text{softmax}(\hat{\mathbf{A}}^K \mathbf{F} \Theta), \quad (4)$$

with K -hop convolutional component $\hat{\mathbf{A}}^K \mathbf{F}$ (where $\mathbf{F} = \mathbf{H}^{<0>}$) multiplied by logistic regression classifier parametrized by weight matrix $\Theta = \prod_{k=1}^K \Theta^{<k>}$.

In this work, we utilize linear architecture of SGC for performing efficient computations on the constructed population graphs (see Fig. 3). For each of the datasets, we first compute normalized adjacency matrix $\hat{\mathbf{A}}$ (with added self-loops) of the population graph $G = (V, E, \mathcal{E})$. We further annotate each node v_i in the graph with feature vector \mathbf{f}_i representing fMRI connectivity. For simplicity of further calculations, we stack individual feature vectors into a single feature matrix $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_n]^\top$. Adjacency matrix $\hat{\mathbf{A}}$ and feature matrix \mathbf{F} become the inputs to SGC model that is further trained to learn output matrix $\mathbf{Y} \in \mathbb{R}^{n \times 2}$. The computational speedup and memory efficiency of SGC over GCN is achieved by precomputing $\hat{\mathbf{A}}^K \mathbf{F}$, which minimizes memory consumption as the model learns only a single weight matrix Θ . The training of the model reduces to binary logistic regression performed with the Adam optimization algorithm.

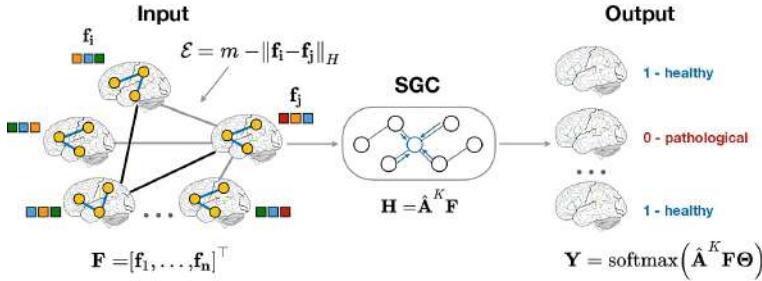


Fig. 3. Overview of classification pipelines based on SGC. SGC takes adjacency matrix representation $\hat{\mathbf{A}}$ of population graph $G = (V, E, \mathcal{E})$ as an input.

4 Results

4.1 Comparison with Baseline Models

To verify the proposed graph construction process and efficiency of the SGC-based model, we compare our method with several baselines. The first two baselines are our PyTorch implementations of the GCN-based population model and the state-of-the-art ChebyGCN-based model, proposed by [13]. These baselines were introduced to test the performance of the SGC model over other graph convolutional models. Additionally, we introduce three baselines: *Graph_random*, *Graph_identity*, and *Graph_no_features* to evaluate graph construction and the importance of subjects' imaging and non-imaging features in the classification task. Underlying these three baselines is the same SGC model running on the same population set. More specifically, *Graph_random* and *Graph_no_features* baselines are designed to test the meaningfulness of the proposed edge construction method based on Hamming distance. In *Graph_random*, edges are randomly rewired so that the edge density of the original graph is preserved. The *Graph_identity* baseline is introduced to test the meaningfulness of fMRI features regardless of non-imaging phenotypic data. This means that each node is still associated with fMRI feature vector, however, the identity matrix is used in place of adjacency matrix to represent a completely disconnected graph, i.e. no two nodes are connected. On the other hand, *Graph_no_features* baseline is implemented to test the performance of the model with respect to graph structure acquired using phenotypic features.

We chose 10 fold cross-validation to make a fair comparison with the state-of-the-art method [13]. For GCN_Cheby implementation on ABIDE, the model parameters were chosen according to [13]. For other cases, hyperparameters for GCN_Cheby, GCN, and SGC were tuned using grid search. For the ABIDE dataset, we trained the model on 2,000 epochs with Adam optimizer, learning rate = 0.1 and propagation step $K = 2$. For COBRE and ADHD-200 datasets, we trained both models on 100 epochs using Adam optimizer with $K = 1$ and learning rate = 0.1 and 0.05 respectively. All models were trained on NVIDIA GeForce GTX TITAN X GPU with 12 GB memory size. The accuracy values

of all baselines across three fMRI datasets using stratified cross-validation are presented in Table 2.

Table 2. Test accuracy(%) and training time per one fold (seconds) of baselines and the proposed model. Numbers are averaged over 10 runs. ChebyGCN denotes our PyTorch implementation of [13]. The best results are highlighted in bold.

Dataset	Model	Test accuracy	Time
ABIDE	ChebyGCN	67.53 ± 4.55	58.22 ± 0.56
	GCN	65.90 ± 4.95	54.24 ± 0.22
	<i>Graph_no_features</i>	50.50 ± 3.79	1.15 ± 0.02
	<i>Graph_random</i>	67.84 ± 4.80	1.29 ± 0.05
	<i>Graph_identity</i>	65.34 ± 5.30	1.28 ± 0.04
	Ours	68.56 ± 4.33	1.28 ± 0.04
COBRE	ChebyGCN	76.54 ± 11.50	2.45 ± 0.04
	GCN	73.02 ± 10.58	0.68 ± 0.09
	<i>Graph_no_features</i>	58.70 ± 8.82	0.07 ± 0.03
	<i>Graph_random</i>	70.26 ± 10.90	0.07 ± 0.03
	<i>Graph_identity</i>	73.82 ± 11.42	0.07 ± 0.03
	Ours	80.55 ± 10.88	0.07 ± 0.03
ADHD-200	ChebyGCN	71.57 ± 7.38	0.63 ± 0.03
	GCN	62.43 ± 6.53	8.28 ± 0.07
	<i>Graph_no_features</i>	70.72 ± 6.00	0.07 ± 0.03
	<i>Graph_random</i>	73.50 ± 6.37	0.35 ± 0.03
	<i>Graph_identity</i>	73.8175 ± 2.70	0.52 ± 0.03
	Ours	74.35 ± 4.76	0.34 ± 0.03

4.2 Evaluation

Based on the results presented in Table 2, we conclude that our model is competitive with other models and baselines. For example, the classification accuracy of our linear model on the ABIDE dataset is comparable to the performance of ChebyGCN based method proposed by [13] (68.56% vs. 67.53%). For COBRE and ADHD-200 datasets, our SGC based model significantly outperforms both ChebyGCN and GCN models. Moreover, the linear model achieves the best performance when compared to *Graph_random*, *Graph_identity*, and *Graph_no_features* baselines, which justifies the usefulness of the constructed population graphs based on phenotypic features. Consistently higher accuracy of our model over other baselines across three datasets shows its robustness of capturing the population graph structure and extracting useful features.

To investigate the parameter sensitivity of our linear graph convolutional model with respect to the number of feature propagation steps K , we vary K

from 1 to 5. In Fig. 4, we present 10-fold cross-validation results for our best performing model based on the graph constructed using subjects phenotypic features and weighted edge function. For the ABIDE dataset, the performance of the model is relatively stable with the top accuracy of 68.56% when $K = 2$. For COBRE, the best average classification accuracy 80.55% is achieved with $K = 1$, significantly reducing as K increases. The best classification accuracy for ADHD-200, 74.35%, is achieved with $K = 1$. For all three datasets, higher values of K lead to a decrease in accuracy. This is due to a small graph size and small diameter, which for all population graphs is equal to 2. Therefore, a higher number of propagation steps causes output features to converge to the same value.

In addition to accuracy improvement, by using a simplified linear graph convolutional model we were also able to achieve a better performance in terms of training time. In the last column of Table 2, we present the training time for each of the models. Our model shows up to 50 times of a speedup on the ABIDE dataset compared to the ChebyGCN model. We attribute the performance boost of our SGC-based model to a significantly lower number of parameters the model has to learn and linearity of the learning function. This makes our population-based model lightweight and allows it to run on CPU. In addition, the model is able to scale up to larger datasets, which is especially important given the increasing volume of the fMRI imaging data.

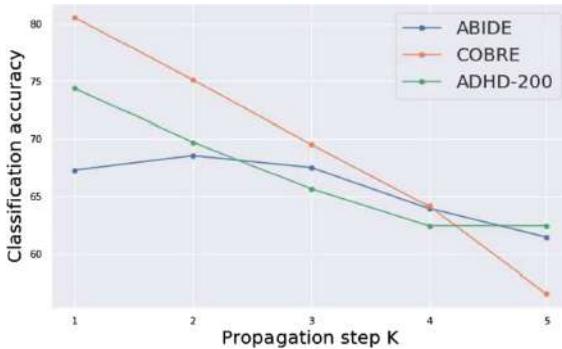


Fig. 4. Parameter analysis with respect to propagation step K on ABIDE, COBRE, and ADHD-200 datasets.

5 Conclusion and Future Work

In this work, we proposed a linear model for the problem of population-based disease prediction that utilizes the novel concept of simplified graph convolutions. We improved upon previous works by designing a simple linear end-to-end graph convolutional model. We evaluated our method on three neuroimaging fMRI datasets with different clinical settings to verify the model's ability to generalize

across datasets. The experiments we conducted demonstrate the comparable or superior performance of the proposed linear SGC-based model, confirming our initial assumption of unnecessary complication of graph convolutional models. In addition to increasing the quality of prediction on COBRE and ADHD-200 datasets, the use of the linear model has a clear impact on decreasing its computational time. Due to the improved performance, efficiency, and scalability, we encourage a simple linear model like SGC to be used as a baseline for comparison with other graph-based models involving deep neural network architecture.

One of the potential extensions of this work will be an automatic phenotypic feature selection for edge construction. The edge construction strategy introduced in this work can be further improved by designing a model capable of learning the edge weights. This can be achieved by introducing self-attention weights on imaging and non-imaging features. This framework could help to discover the features which can further improve classification accuracy.

Acknowledgement. This work was supported by JSPS Grant-in-Aid for Scientific Research (B)(Grant Number 17H01785) and JST CREST (Grant Number JPMJCR1687).

References

1. Bassett, D.S., Bullmore, E.T.: Human brain networks in health and disease. *Curr. Opin. Neurol.* **22**(4), 340 (2009)
2. Bassett, D.S., Zurn, P., Gold, J.I.: On the nature and use of models in network neuroscience. *Nat. Rev. Neurosci.* **19**(9), 566 (2018)
3. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**(3), 186–198 (2009)
4. Bullmore, E., Sporns, O.: The economy of brain network organization. *Nat. Rev. Neurosci.* **13**(5), 336 (2012)
5. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Advances in Neural Information Processing Systems, pp. 3844–3852 (2016)
6. He, T., Kong, R., Holmes, A.J., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T.: Is deep learning better than kernel regression for functional connectivity prediction of fluid intelligence? pp. 1–4 (2018)
7. Hsieh, T.H., Sun, M.J., Liang, S.F.: Diagnosis of schizophrenia patients based on brain network complexity analysis of resting-state fMRI. In: The 15th International Conference on Biomedical Engineering, pp. 203–206. Springer (2014)
8. Ji, C., Maurits, N.M., Roerdink, J.B.T.M.: Comparison of brain connectivity networks using local structure analysis, pp. 639–651 (2018)
9. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. *Science* **349**(6245), 255–260 (2015)
10. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: 5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings, Toulon, France, 24–26 April 2017 (2017)
11. 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)

12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436 (2015)
13. Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D.: Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med. Image Anal.* **48**, 117–130 (2018)
14. Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Moreno, R.G., Glocker, B., Rueckert, D.: Spectral graph convolutions for population-based disease prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 177–185. Springer (2017)
15. Ventresca, M.: Using algorithmic complexity to differentiate cognitive states in fMRI. In: International Conference on Complex Networks and their Applications, pp. 663–674. Springer (2018)
16. Whitfield-Gabrieli, S., Nieto-Castanon, A.: Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connectivity* **2**(3), 125–141 (2012)
17. Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., Weinberger, K.: Simplifying graph convolutional networks **97**, 6861–6871 (2019)



Adaptive Network Modeling for Criterial Causation

Jan Treur^(✉)

Social AI Group, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

j.treur@vu.nl

https://www.researchgate.net/profile/Jan_Treur

Abstract. Propagation of activation of neurons depends on settings of a number of intrinsic characteristics of the network of neurons, such as synaptic connection strengths and excitability thresholds for neurons. These settings serve as criteria on the incoming signals for a neuron to get activated. As part of the plasticity of the neural processing these network characteristics also change over time. Such changes can be slow compared to propagation of activation, like in learning from a number of experiences, but they can also be fast, like in memory formation. From the informational perspective on the criteria, this can be considered a form of information formation, and the firing of neurons as driven by this information. This is called criterial causation. In this paper, an adaptive network model is presented modeling such criterial causation. Moreover, it is shown how criterial causation in the brain relates to the more general temporal factorisation principle for the world's dynamics.

1 Introduction

Neural processing is much more than propagation of activation of neurons; e.g., [25]. Such propagation depends on settings for a number of intrinsic characteristics of the network of neurons, such as synaptic connection strengths and excitability thresholds for neurons. These settings form a configuration in the brain that serves as a set of criteria on the incoming patterns of signals for a neuron to get activated; by Tse [25, 26] this is called *criterial causation*. As put forward by Tse, the criteria can be considered a form of *information* realised in the concerning brain configuration: ‘physically realised informational criteria’, e.g., [26], p. 259; in future situations, only when these criteria are met, the neuron will fire. As part of the neural processing, not only activation of neurons, but also the network characteristics defining these criteria change over time. Such changes of the network characteristics depend on the patterns in the past that affect them. The changes can be slow compared to propagation of activation of neurons, like in learning from a larger number of experiences, but they can also happen almost instantly, like in memory formation; the latter is called *rapid resetting* of the criteria by [25, 26]. From the informational perspective on the criteria, this form of network adaptation can be considered as a form of *emerging information formation*, and the firing of neurons as driven by this information [25, 26].

This paper presents a computational adaptive network model that makes the above more precise, and illustrates it by a simulation for an example scenario. Moreover, the paper shows how the perspective as sketched can be considered a special case of the more general perspective on the dynamics of the world based on temporal factorisation by mediating state properties [17, 18]. Mediating state properties on the one hand encode in the present world configuration, information on the past pattern in the world, and on the other hand they determine the possible future patterns for the world from there. This wider perspective generalises specific processes of emerging information formation and usage taking place in the brain, to more general emerging information formation and usage as an inherent characteristic of the world's dynamics. Viewed in this more general way, it may be argued that the brain's functioning by criterial causation is entailed by the more general principle of temporal factorisation of the world's dynamics, or at least makes clever use of that general principle. Then, viewed from an informational perspective, the temporal factorisation principle can be seen as a way in which in general the world's dynamics creates emergent information in its configuration, and the more specific principle of criterial causation describes particularly how the brain creates emergent information in its configuration. In both cases this emergent information determines the options for the future patterns.

2 Temporal Factorisation and Criterial Causation

The temporal factorisation principle [17, 18] states that any systematic temporal ‘past pattern implies future pattern’ relationship $a \Rightarrow b$ between a past pattern a and a future pattern b can be factorised in the form of two temporal relationships $a \Rightarrow p$ and $p \Rightarrow b$ for some state property p (called *mediating state property*) of the present world state; see Fig. 1, left hand side. More specifically, the principle claims that for any ‘past pattern implies future pattern’ relationship $a \Rightarrow b$ there exists a world state property p (describing some configuration of the present world state) such that temporal relationships ‘past pattern implies present state property’ $a \Rightarrow p$ and ‘present state property implies future pattern’ $p \Rightarrow b$ hold: $a \Rightarrow b \Rightarrow \exists p \ a \Rightarrow p \ \& \ p \Rightarrow b$. Note that the mediating state property p does not need to be one simple state property; it can (and often will) be a combination of multiple state properties occurring at that time point.

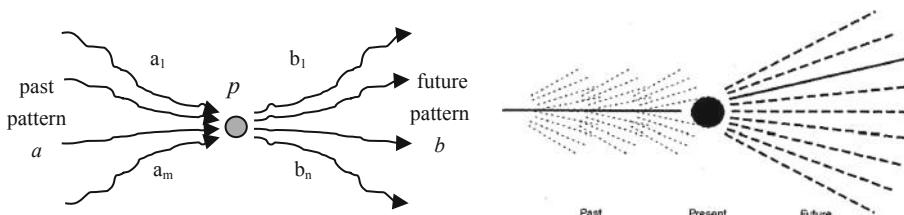


Fig. 1. Left hand picture: Mediating state property p in the present for the past pattern implies future pattern relation $a \Rightarrow b$, adopted from [17], p. 60, Fig. 1. Right hand picture: Criterial causation; adopted from [25], p. 125, Fig. 6.2.

The temporal factorisation principle claims that the present world state configuration p encodes sufficient information so that the world can forget about the temporal pattern a in the past if it makes temporal pattern b occur in the future; therefore it essentially is a claim that world state configurations are sufficiently rich to encode all (future-)relevant information on the past (which in theory could concern an almost infinite number of world states, with their temporal relations) in some condensed form in one state configuration. In [17] it is discussed in more detail how this principle relates to views on the world's dynamics from Descartes [8], Laplace [13], Ashby [2] and van Gelder and Port [27]. Descartes [8] puts forward that systematic relationships (laws of nature) exist for world states over time, in the sense that past world states imply future world states (called the clockwork universe). Laplace [13] claims: 'We may regard the present state of the universe as the effect of its past and the cause of its future'. In [27], following Ashby [2] the notion of state-determined system is taken as a basis for dynamics: '... its current state always determines a unique future behaviour ... the future behaviour cannot depend in any way on whatever states the system might have been in *before* the current state' [27], p. 6. Note that the temporal factorisation principle relates to these views but, in contrast, does not assume an overall deterministic world. It only applies to aspects of the world that happen to be deterministic (as expressed by the conditional $a \Rightarrow b$). In [17, 18], the temporal factorisation principle is modeled in a formalised manner, and by many examples it is shown how the temporal factorisation principle plays its role in the world's dynamics, taken from Physics and from Cognitive Science.

One example to illustrate the principle is as follows. Suppose in reality or in a virtual game context in the present state there is a door that was locked by someone in a past pattern a and therefore only can be opened in a future pattern b in which you bring the right key with you. From $a \Rightarrow b$, the temporal factorisation principle concludes that there is a mediating state property p that holds in the present state such that $a \Rightarrow p \ \& \ p \Rightarrow b$. Indeed this p is the state property of the door being locked; then $a \Rightarrow p$ expresses that if in the past someone locked the door, it is locked now, and $p \Rightarrow b$ expresses that (only) when you bring the right key with you in the future it can be opened. The informational perspective here is that within the world the lock represents some form of information, and only when in the future pattern b the right key (with the right key shape, according to that lock information) occurs, the door opens. This is a clear case illustrating the way in which the mediating state configuration encodes information, and it is this information what drives the world to future pattern b , in which when the key occurs, the door can be opened. In this case, humans are actors encoding the information in the world, as the lock and the matching key are human-made: humans informationalise the world, the world is becoming more informational due to human intervention. A similar example of human-made informationalisation of the world is when Little Thumb drops pebbles to find his way back. However, the temporal factorisation principle claims in general that the world (as a kind of actor) is doing a similar encoding of information concerning past patterns in present world state configurations without human intervention.

The temporal factorisation principle can also be illustrated by the behavioural notion of ‘delayed response behaviour’, that has a long tradition in the psychology literature concerning animal cognition and behaviour; e.g., [7, 9, 11, 15]. Consider c is food at location l_0 visible for an animal, d is the animal gets released, and e is that the animal gets at l_0 . An example of a past pattern a is: for at least two different time points in the past, state c (food visible at l_0) occurred. An example of a future pattern b is: if in some future state d occurs (animal is released), then at some later time point state e will occur (animal at l_0). The temporal factorisation principle says: if $a \Rightarrow b$, then there is some state property p such that $a \Rightarrow p$ and $p \Rightarrow b$. In this example, the (mediating) state property p postulated by the temporal factorisation principle would refer to an internal cognitive state, functioning as a form of memory for the animal. More specifically, in this case, after observing in animal experiments many times that the past-future relationship $a \Rightarrow b$ holds, the temporal factorisation principle postulates that some kind of (memory) state is formed after a past pattern a occurred, and that this memory state drives the organism’s future behaviour in the sense that b holds. Also in this case, such a memory state can be considered to encode information about the world, and this information drives the future behaviour. Here, the information formation is an emerging process taking place without any human intervention, and also for the animal probably it will not happen as an intentional process. So, the world itself does the information formation, in this case via the brain.

Criteria Causation as Temporal Factorisation

After Sect. 1 and the above explanation, it may already have become clear that temporal factorisation and criterial causation have a close relationship; even the two pictures shown in Fig. 1 for temporal factorisation (left hand side) and criterial causation (right hand side) have a high extent of similarity. The correspondence can essentially be formulated as follows. In the above explanation the mediating state property p in the present state for temporal factorisation corresponds to the locked door; this lock defines the criteria for criterial causation. Fulfilment of the criteria in a future pattern b correspond to the fitting of a key in the lock, after which in b the door opens. This fulfilment corresponds to the firing of a neuron and its consequences in future pattern b .

3 Criterial Causation in Temporal-Causal Networks

In the above general formulation of the temporal factorisation principle, world states and past and future world patterns are kept abstract. However, often a notion of causality is considered as a way to describe the world’s dynamics. Also in Tse [25]’s perspective based on criterial causation, causal relations play an important role. Therefore it makes sense to analyse how temporal factorisation and criterial causation work in combination with a description of world dynamics by a temporal-causal network [19, 21]. A temporal-causal network is characterised by *connectivity characteristics* (the connections from nodes X to Y and their weights $\omega_{X,Y}$), *aggregation characteristics* (for each node Y , by a combination function $c_Y(...)$ some form of aggregation is applied to the causal impacts from the incoming connections), and

timing characteristics (nodes Y have speed factors η_Y indicating how fast they change upon causal impact). The difference equations used for simulation and mathematical analysis incorporate these three types of network characteristics $\omega_{X,Y}$, $c_Y(\dots)$, η_Y : for any state Y it holds

$$Y(t + \Delta t) = Y(t) + \eta_Y [c_Y(\omega_{X_1,Y} X_1(t), \dots, \omega_{X_k,Y} X_k(t)) - Y(t)] \Delta t \quad (1)$$

where X_1, \dots, X_k are the states from which Y gets incoming connections. These concepts enable to design networks with their dynamics in a declarative manner, by mathematically defined relations; see [19, 21] for more information on Network-Oriented Modeling based on temporal-causal networks.

Criteria for Criterial Causation in Temporal-Causal Networks

Based on (1) the firing criterion of state Y can be expressed by putting that the aggregated impact on Y is higher than 0.5:

$$c_Y(\omega_{X_1,Y} X_1(t), \dots, \omega_{X_k,Y} X_k(t)) > 0.5 \quad (2)$$

So, this (2) is taken as the general criterion for criterial causation for a state Y in a temporal-causal network. Often used combination functions $c_Y(\dots)$ are the simple logistic **slogistic** $_{\sigma,\tau}(\dots)$ and advanced logistic sum function **alogistic** $_{\sigma,\tau}(\dots)$, both with steepness parameter $\sigma > 0$ and excitability threshold parameter τ :

$$\text{slogistic}_{\sigma,\tau}(V_1, \dots, V_k) = \frac{1}{1 + e^{-\sigma(V_1 + \dots + V_k - \tau)}} \quad (3)$$

$$\text{alogistic}_{\sigma,\tau}(V_1, \dots, V_k) = \left[\frac{1}{1 + e^{-\sigma(V_1 + \dots + V_k - \tau)}} - \frac{1}{1 + e^{\sigma\tau}} \right] (1 + e^{-\sigma\tau}) \quad (4)$$

Here the V_i denote the single impacts $\omega_{X_i,Y} X_i(t)$ on state Y for each of the incoming connections from states X_1, \dots, X_k . For the simple logistic sum function (3), criterion (2) is equivalent to

$$\frac{1}{1 + e^{-\sigma(V_1 + \dots + V_k - \tau)}} > 0.5$$

with the V_i denoting the single impacts $\omega_{X_i,Y} X_i(t)$ on state Y . By rewriting (see Box 1 left), this is equivalent to

$$\omega_{X_1,Y} X_1(t) + \dots + \omega_{X_k,Y} X_k(t) > \tau \quad (5)$$

So, (5) is the more specific criterion for criterial causation for the simple logistic combination function. For the advanced logistic combination the following similar but more complicated criterion for criterial causation can be derived (see Box 1 right).

$$\omega_{X_1,Y}X_1(t) + \dots + \omega_{X_k,Y}X_k(t) > \tau - \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right)/\sigma \quad (6)$$

$1 + e - \sigma(V_1 + \dots + V_k - \tau) < 2 \Leftrightarrow$ $e - \sigma(V_1 + \dots + V_k - \tau) < 1 \Leftrightarrow$ $\sigma(V_1 + \dots + V_k - \tau) > 0 \Leftrightarrow$ $V_1 + \dots + V_k > \tau \Leftrightarrow$ $\omega_{X_1,Y}X_1(t) + \dots + \omega_{X_k,Y}X_k(t) > \tau$	$\left[\frac{1}{1+e^{-\sigma(V_1+\dots+V_k-\tau)}} - \frac{1}{1+e^{\sigma\tau}}\right](1+e^{\sigma\tau}) > 0.5 \Leftrightarrow$ $\left[\frac{1}{1+e^{-\sigma(V_1+\dots+V_k-\tau)}} - \frac{1}{1+e^{\sigma\tau}}\right] > \frac{0.5}{1+e^{-\sigma\tau}} \Leftrightarrow$ $\frac{1}{1+e^{-\sigma(V_1+\dots+V_k-\tau)}} > \frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}} \Leftrightarrow$ $1 + e - \sigma(V_1 + \dots + V_k - \tau) < \frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} \Leftrightarrow$ $e - \sigma(V_1 + \dots + V_k - \tau) < \frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1 \Leftrightarrow$ $-\sigma(V_1 + \dots + V_k - \tau) < \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right) \Leftrightarrow$ $\sigma(V_1 + \dots + V_k - \tau) > -\log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right) \Leftrightarrow$ $V_1 + \dots + V_k > \tau - \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right)/\sigma \Leftrightarrow$ $\omega_{X_1,Y}X_1(t) + \dots + \omega_{X_k,Y}X_k(t) > \tau - \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right)/\sigma$
--	--

Box 1. Deriving criteria (5) (left) and (6) (right) for criterial causation for combination functions **slogistic** _{σ,τ} (...) and **alogistic** _{σ,τ} (...)

Note that for these combination functions the criteria are expressed as linear inequalities for state values X_1, \dots, X_k with as coefficients expressions in terms of network characteristics $\boldsymbol{\omega}$, σ , τ . The criteria for criterial causation for other combination functions (scaled maximum **smax** _{λ} (...) and minimum **smin** _{λ} (...), scaled sum **ssum** _{λ} (...), Euclidean **eucl** _{n,λ} (...) and scaled geometric mean **sgeomean** _{λ} (...)) found are in Table 1.

Table 1. Overview of the criteria for criterial causation for different combination functions

Combination function	Criterion for criterial causation
Name	Formula
$c_Y(V_1, \dots, V_k)$	$c_Y(V_1, \dots, V_k)$
$\text{slogistic}_{\sigma,\tau}(V_1, \dots, V_k)$	$c_Y(\omega_{X_1,Y}X_1(t), \dots, \omega_{X_k,Y}X_k(t)) > 0.5$
$\text{alogistic}_{\sigma,\tau}(V_1, \dots, V_k)$	$\omega_{X_1,Y}X_1(t) + \dots + \omega_{X_k,Y}X_k(t) > \tau$
	$\left[\frac{1}{1+e^{-\sigma(V_1+\dots+V_k-\tau)}} - \frac{1}{1+e^{\sigma\tau}}\right](1+e^{\sigma\tau}) > 0.5 \Leftrightarrow \omega_{X_1,Y}X_1(t) + \dots + \omega_{X_k,Y}X_k(t) > \tau - \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right)/\sigma$
$\text{smax}_{\lambda}(V_1, \dots, V_k)$	$\omega_{X_i,Y}X_i(t) > 0.5 \lambda \text{ for some } i$
$\text{smin}_{\lambda}(V_1, \dots, V_k)$	$\omega_{X_i,Y}X_i(t) > 0.5 \lambda \text{ for all } i$
$\text{ssum}_{\lambda}(V_1, \dots, V_k)$	$\omega_{X_1,Y}X_1(t) + \dots + \omega_{X_k,Y}X_k(t) > 0.5\lambda$
$\text{eucl}_{n,\lambda}(V_1, \dots, V_k)$	$\sqrt[n]{\frac{V_1^{\lambda} + \dots + V_k^{\lambda}}{\lambda}} > 0.5^n\lambda$
$\text{sgeomean}_{\lambda}(V_1, \dots, V_k)$	$\omega_{X_1,Y}X_1(t) * \dots * \omega_{X_k,Y}X_k(t) > 0.5^k\lambda$

Emerging Criteria for Criterial Causation in Adaptive Temporal-Causal Networks

Networks considered for real world domains are often adaptive, so that some or all of the above network characteristics can change over time as well. This is the way in which the criteria for criterial causation are set dynamically, and the information based on the criteria is not fixed but emerges. For example, in the above criteria (2), (5), (6) the connection weights ω , excitability threshold τ and steepness σ can change over time, thereby changing the criterion. Then the overall dynamics is an interaction (or co-evolution) of two types of dynamics, one of which (dynamics of the nodes) is modeled in a declarative mathematical manner from a Network-Oriented Modeling perspective, and the other one (dynamics of the characteristics and the criteria they define) is usually described in a different, nondeclarative (procedural or algorithmic) manner. This leads to a kind of hybrid model. By using the notion of *network reification*, the Network-Oriented Modeling perspective can also be used to design *adaptive* networks in a declarative manner by mathematically defined relations. This works by adding the adaptive network characteristics (in a reified form) to the (base) network as nodes at a second level, called *reification level*, while the original network forms the *base level*. In this way an extended, reified network is obtained, which is again a temporal-causal network. As for any temporal-causal network model, the dynamics of such a reified network is described in a declarative mathematical Network-Oriented manner by the nodes and their connections, including causal interlevel connections for the impact from one level to the other. This can iteratively be applied to obtain multiple reification levels to model multiple orders of adaptation of a network. For more details, see [20, 23], or the forthcoming book [24].

Using a reified temporal-causal network model to describe the world's dynamics, in a relatively easy manner causality can be modeled for criterial causation and the temporal factorisation principle. For example, then a mediating state configuration p can be described by the state values of a number of nodes, which each can be at the base level, or at any reification level. And also the past and future patterns a and b are described as patterns of state values for a number of nodes of any level over time. In particular, for the criteria expressed by linear inequalities (5) and (6), the coefficients are based on reification states at the reification level, whereas the states X_1, \dots, X_k to which the criteria are applied are at the base level.

4 An Example Reified Network for Criterial Causation

As discussed in Sect. 3, a temporal-causal network model involves three main characteristics connectivity, aggregation, and timing of the network structure, modeled by $\omega_{X,Y}$, $c_Y(\dots)$, η_Y . The difference equations used for simulation and mathematical analysis incorporate these three types of network characteristics as expressed in (1) above. For the sake of practicality, for each application from a library basic combination functions $bcf_i(\dots)$, $i = 1, \dots, m$ can be selected according to weights $\gamma_{i,Y}$, so that the combination function used for any state Y is the weighted average

$$\mathbf{c}_Y(\dots) = (\gamma_{1,Y} \text{bcf}_1(\dots) + \dots + \gamma_{m,Y} \text{bcf}_m(\dots)) / (\gamma_{1,Y} + \dots + \gamma_{m,Y})$$

Moreover, parameters of these combination functions can be considered, so that $\text{bcf}_i(\dots) = \text{bcf}_i(\mathbf{p}, \mathbf{v})$ with \mathbf{p} a list of parameters and \mathbf{v} a list of values. For reified network models additional reification states are introduced in the network that explicitly represent characteristics of the network such as *connectivity*, *aggregation*, and *timing*, and makes them adaptive; these reification states are indicated by $\mathbf{W}_{X,Y}$, \mathbf{C}_i , $\mathbf{P}_{i,j,Y}$, and \mathbf{H}_Y :

- **Adaptive connection weight $\omega_{X,Y}$:** reified connection weight representations $\mathbf{W}_{X,Y}$
- **Adaptive combination function weight $\gamma_{i,Y}$ for $\mathbf{c}_Y(\dots)$:** reified combination function weight representations $\mathbf{C}_{i,Y}$ (for the i^{th} combination function used)
- **Adaptive combination function parameter \mathbf{p}_Y for $\mathbf{c}_Y(\dots)$:** reified combination function parameter representations $\mathbf{P}_{i,j,Y}$ (the j^{th} parameter of the i^{th} combination function for Y)
- **Adaptive speed factor η_Y :** reified speed factor representations \mathbf{H}_Y

Example Scenario

The considered scenario is as follows. A person who is new in an organisation has to recognize a colleague from seeing his face, modeled by stimulus s . There are two options, colleagues a_1 and a_2 . Deciding for one of them is represented by preparation states ps_{a_i} . A belief bs_1 suggests that it should be colleague a_1 , and a belief bs_2 that it should be colleague a_2 ; however, these beliefs are indicative (for example, based on the location at which the person is seen), but not sufficient to firmly decide for one of the two. The beliefs and s are generated from independent circumstantial environmental factors; for the model they just happen. Two types of adaptive network characteristics are involved: the weights of the connections from the sensory state srs_s for s to ps_{a_1} and ps_{a_2} , and the excitability thresholds for states ps_{a_1} and ps_{a_2} . During the scenario these characteristics are adapted so that a decision results. The obtained settings define the criteria for criterial causation of the recognition. Based on them, in future situations any encounter with s (also at unexpected locations, such as a supermarket or during holidays) leads to fulfilment of the criteria and as a consequence to recognition.

In a graphical representation the reification states are depicted in a 3D format in a second plane, above the (pink) plane for the base network; see the blue plane in the example reified network model depicted in Fig. 2, also indicated as reification level; see Table 2 for an explanation of the states. Three types of causal connections are distinguished: upward causal connections, downward causal connections and leveled (horizontal) causal connections. The downward causal connections have their own fixed role and meaning in the sense that as their special effect they are causally effectuating one of the four types of adaptive values listed above. In the reified network model for criterial causation described here, for almost all states logistic function $\text{allogistic}_{\sigma,\tau}(\dots)$ is used; see (4) above. As an exception, for Hebbian learning [10] the following combination function is used for the reification states $\mathbf{W}_{X,Y}$ at the reification level:

$$\mathbf{hebb}_{\mu}(V_1, V_2, W) = V_1 V_2 (1 - W) + \mu W \quad (7)$$

where V_1, V_2 indicate the single impacts from the connected states (base states in the bottom plane in Fig. 1) and W the connection weight (represented by reification state $\mathbf{W}_{X,Y}$ in the upper plane in Fig. 1), and μ is a persistence parameter. In Fig. 1 only the following reification states are used:

- $\mathbf{W}_{X,Y}$ plays the role of connection weight for the connection from X to Y [10]
- \mathbf{T}_Y plays the role of combination function parameter value for the excitability threshold parameter τ for state Y [5]

The roles of the different base and reification states are specified by *role matrices* **mb** (base connection role), **mcw** (connection weight role), **ms** (speed factor role), **mcfw** (combination function weight role), and **mcfp** (combination function parameter role); e.g., [22–24]. In these role matrices (see Box 2) at each row for the given state it is specified which other states have impact on it (incoming arrows in Fig. 2).

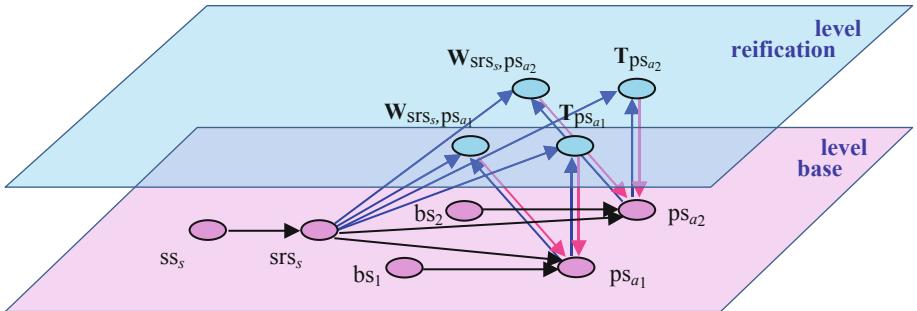


Fig. 2. Overview of the example reified network model for criterial causation, with: (1) *base level* for face recognition (lower plane, pink), (2) *reification level* (upper plane, blue) for the criteria represented by the weights ω of the base connections from srs_s to ps_{a_1} and ps_{a_2} (reified by the two \mathbf{W} states) and the excitability thresholds τ of these two base states ps_{a_1} and ps_{a_2} (reified by the two \mathbf{T} states)

This is distinguished according to their role: *base* or *non-base* connections, from which for the latter a distinction is made for the roles *connection weight*, *speed factor*, *combination function weight* and *combination function parameter reification*. The matrices all have rows according to the numbered states X_1, X_2, X_3, \dots . For a given application a limited sequence of combination functions is specified by **mcf** = [...], for the example this is **mcf** = [2 3 35], where the numbers 2, 3 refer to the numbering in the function library, the first three being **eucl_{n,λ}**(...), **alognistic_{σ,τ}**(...), **hebb_μ**(...). In Box 2 the role matrices for the reified network model for criterial causation are shown.

Table 2. Overview of the states in the example reified network model

nr	state name	explanation
X_1	ss_s	Sensor state for external stimulus s (seeing a face)
X_2	srs_s	Sensory representation state for stimulus s
X_3	bs_1	Belief state 1 (belief that it is Person 1)
X_4	bs_2	Belief state 2 (belief that it is Person 2)
X_5	ps_{a1}	Preparation state for recognition as Person 1
X_6	ps_{a2}	Preparation state for recognition as Person 2
X_7	$W_{srs_s, ps_{a1}}$	Reification state for the weight of the connection from srs_s to ps_{a1}
X_8	$W_{srs_s, ps_{a2}}$	Reification state for the weight of the connection from srs_s to ps_{a2}
X_9	$T_{ps_{a1}}$	Reification state for the excitability threshold of ps_{a1}
X_{10}	$T_{ps_{a2}}$	Reification state for the excitability threshold of ps_{a2}

The first role matrix **mb** for *base connectivity* specifies on each row for a given state from which states at the same or a lower level it has incoming connections. For example, in the fifth row it is indicated that state X_5 ($= ps_{a1}$) has two incoming base connections, from state X_2 ($= srs_s$), and from state X_3 ($= bs_1$). As another example, the 7th row indicates that state X_7 ($= W_{srs_s, ps_{a1}}$) has incoming base connections from X_2 ($= srs_s$), X_5 ($= ps_{a1}$) and from X_7 itself, and in that order, which is important as the Hebbian combination function **hebb_μ**(...) used here is not symmetric in its arguments.

In a similar way the four types of role matrices for *non-base connectivity* (i.e., connectivity from reification states at a higher level of reification: the downward arrows in Fig. 2), were defined: role matrices **mew** for connection weights and **ms** for speed factors, and role matrices **mcfw** for combination function weights and **mcfp** for combination function parameters (see Box 2).

Within each of the role matrices **mew**, **mcfw**, **mcfp** and **ms** a difference is made between cell entries indicating (in red) a reference to the name of another state that as a form of reification represents in a dynamic manner an adaptive characteristic, and entries indicating (in green) fixed values for nonadaptive characteristics. Indeed, in Box 1 it can be seen that the red cells of the non-base role matrices are filled with the (reification) states X_7 to X_{10} of the first reification level. For example, in Box 1 the name X_7 in the red cell row-column (5, 1) in role matrix **mew** indicates that the value of the connection weight from srs_s to ps_{a1} (as indicated in role matrix **mb**) can be found as value of the seventh state X_7 . In contrast, the 1 in green cell (7, 1) of **mew** indicates the static value of the connection weight from X_2 ($= srs_s$) to X_7 ($= W_{srs_s, ps_{a1}}$).

As yet another example, in role matrix **mcfp** for the combination function parameters, in cell (5, 2) it is indicated that the value of the excitability threshold of ps_{a1} is represented by reification state X_9 ($= T_{ps_{a1}}$). For more explanation about this role matrix specification format, see [22, 23], or the forthcoming book [24].

mb	base connectivity	1	2	3		mcw connection weights	1	2	3	
X_1	ss_s	X_1				X_1	ss _s	1		
X_2	srs_s		X_1			X_2	srs _s	1		
X_3	bs_1		X_3			X_3	bs ₁	1		
X_4	bs_2		X_4			X_4	bs ₂	1		
X_5	ps_{a1}		X_2	X_3		X_5	ps _{a1}	X_7	0.5	
X_6	ps_{a2}		X_2	X_4		X_6	ps _{a2}	X_8	0.5	
X_7	$W_{srs_s, ps_{a1}}$	X_2	X_5	X_7		X_7	$W_{srs_s, ps_{a1}}$	1	1	1
X_8	$W_{srs_s, ps_{a2}}$	X_2	X_6	X_8		X_8	$W_{srs_s, ps_{a2}}$	1	1	1
X_9	$T_{ps_{a1}}$	X_2	X_5	X_9		X_9	$T_{ps_{a1}}$	-0.2	-0.2	1
X_{10}	$T_{ps_{a2}}$	X_2	X_6	X_{10}		X_{10}	$T_{ps_{a2}}$	-0.2	-0.2	1
mcfw combination function weights		1	2	3		function	1	2	3	
X_1	ss_s					alo-gistic	1	2	3	
X_2	srs_s	1				hebb				
X_3	bs_1					step-mod				
X_4	bs_2									
X_5	ps_{a1}	1								
X_6	ps_{a2}	1								
X_7	$W_{srs_s, ps_{a1}}$			1						
X_8	$W_{srs_s, ps_{a2}}$		1							
X_9	$T_{ps_{a1}}$	1								
X_{10}	$T_{ps_{a2}}$	1								
ms speed		1				initial values				
X_1	ss_s		2			X_1	ss _s	0		
X_2	srs_s		0.5			X_2	srs _s	0		
X_3	bs_1		2			X_3	bs ₁	0		
X_4	bs_2		2			X_4	bs ₂	0		
X_5	ps_{a1}		0.2			X_5	ps _{a1}	0		
X_6	ps_{a2}		0.5			X_6	ps _{a2}	0		
X_7	$W_{srs_s, ps_{a1}}$		0.3			X_7	$W_{srs_s, ps_{a1}}$	0.3		
X_8	$W_{srs_s, ps_{a2}}$		0.3			X_8	$W_{srs_s, ps_{a2}}$	0.3		
X_9	$T_{ps_{a1}}$		0.07			X_9	$T_{ps_{a1}}$	0.8		
X_{10}	$T_{ps_{a2}}$		0.07			X_{10}	$T_{ps_{a2}}$	0.8		

Box 2. Specification in role matrices format for the reified example network for criterial causation

5 Example Simulation of Criterial Causation

Using a dedicated modeling environment [22] for reified network models, simulations have been performed. In Fig. 3 simulation results are shown for the example Scenario described in Sect. 4. Here the settings shown in Box 2 were used. Like in the temporal factorisation principle, for the overall process a past pattern a , a future pattern b , and a mediating (present) state property p , are distinguished, each of which will be briefly discussed.

Past Pattern a (From Time Point 0 to 100)

During the past pattern a , the stimulus s (the observed face) occurs twice: from time 25 to 50 and from time 75 to 100. In these time periods also belief state bs_2 occurs. The upper and middle graph in Fig. 3 display the past pattern a and show how the recognition of stimulus s as Person 2 is emerging: during the first encounter, ps_{a_2} (the red line) increases relatively slowly, and during the second encounter this happens faster; apparently already a more adequate informational criterion has been set for the recognition.

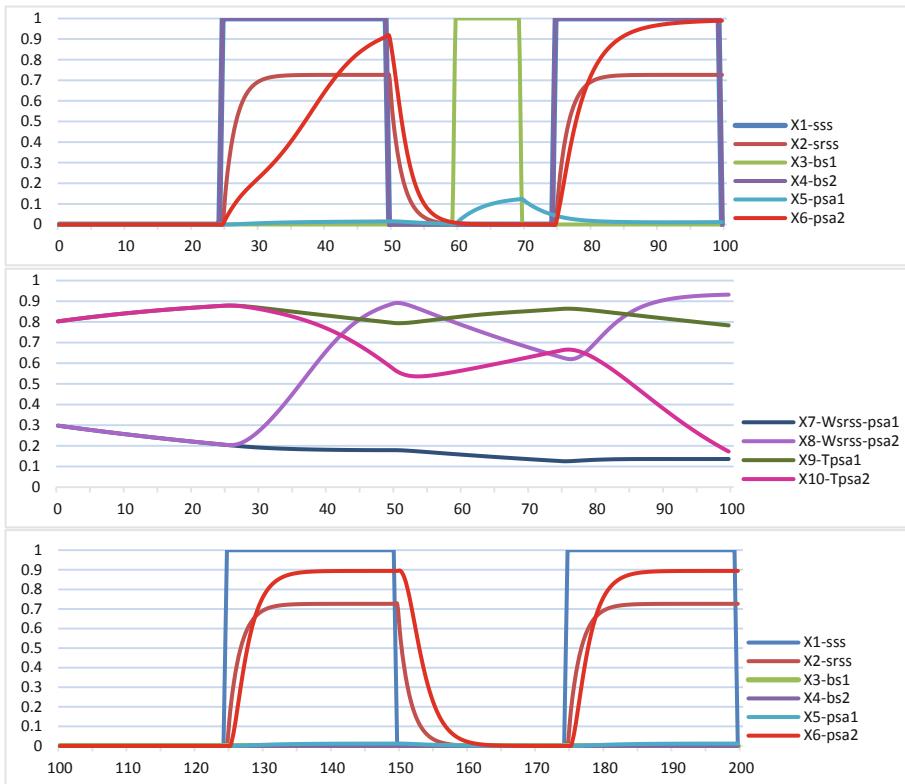


Fig. 3. A past pattern displayed in the upper graph and middle graph, and a future pattern displayed in the lower graph, where the criterion set as a mediating (present) state by the past pattern drives the future pattern.

In the middle graph, the emergence of the characteristics for this criterion are shown. In particular, it is shown that (due to belief state bs_2) the reified adaptive connection weight $W_{srs_s, ps_{a_2}}$ from srs_s to ps_{a_2} (purple line) becomes stronger, and the reified excitability threshold $T_{ps_{a_2}}$ of ps_{a_2} (pink line) becomes lower. Note that between time 60 and 70 for a short period belief state bs_1 occurs, but this disturbance has no substantial consequences.

Criterion Set in the Mediating State Property p at Time Point 100

The mediating state property p describes criterion (6) for criterial causation. In this example scenario, the mediating state property consists of the values of the following relevant characteristics: the reified weights $\mathbf{W}_{\text{srs}_s, \text{ps}_{a1}}$ and $\mathbf{W}_{\text{srs}_s, \text{ps}_{a2}}$ of the connections from srs_s to ps_{a1} and ps_{a2} , and the reified excitability thresholds $\mathbf{T}_{\text{ps}_{a1}}$ and $\mathbf{T}_{\text{ps}_{a2}}$ for ps_{a1} and ps_{a2} , so the reification states X_7 to X_{10} . Note that all of them are at the reification level. At time point 100 they have the following values: $X_7 = 0.136275$, $X_8 = 0.93172$, $X_9 = 0.78281$, $X_{10} = 0.17232$. So, this configuration described at the reification level, defines the mediating state property, and, equivalently, the coefficients of criterion (6) for criterial causation. It can be seen that the connection weight from srs_s to ps_{a1} is low (0.136275) and the excitability threshold for ps_{a1} is high (0.78281). Therefore, for the choice for Person 1 the criterion for firing cannot be met in a reasonable way. For ps_{a1} it is the opposite: the connection weight from srs_s to ps_{a2} is high (0.93171758) and the excitability threshold for ps_{a2} is low (0.17232). This means that the criterion for this choice for Person 2 is easy to fulfill by the causal impact coming from srs_s . Indeed, using (6) from Sect. 3 substituted by the values of the \mathbf{W} states and \mathbf{T} states (reifying ω and τ , respectively) at time point 100, and the value 5 of σ , and 0.5 for the ω from the belief state, the criterion for firing becomes

$$0.5 \text{bs}_2(t) + 0.93172 \text{srs}_s(t) > 0.0857$$

Not assuming any positive value of $\text{bs}_2(t)$, this is already fulfilled if

$$\text{srs}_s(t) > 0.0857 / 0.93172 = 0.092$$

So already a very weak sensory representation signal of the observed face as low as 0.1 would be enough to recognize the face.

Future Pattern b (From Time Point 100 to 200)

In the lower graph in Fig. 3 the future pattern is displayed; it is shown how based on the emerged criterion (represented by the present mediating state at time 100), indeed instant recognition takes place in the absense of any of the belief states. Here the criterion is based on the (constant) values for the reified connection weights and excitability thresholds defining the mediating state property: the values $X_7 = 0.136275$, $X_8 = 0.93172$, $X_9 = 0.78281$, $X_{10} = 0.17232$. In the future pattern at times 125 and 175 the person is seen again (ss_s whereby the belief state bs_2 is kept 0), and the criterion becomes fulfilled so that indeed firm recognition ps_{a2} as Person 2 takes place.

6 Discussion

Criterial causation as introduced by Tse [25] describes how as a form of plasticity, in the brain, configurations emerge that provide informational criteria for future processing and behaviour. In the current paper, first it was shown how this notion relates to the more general notion of temporal factorisation based on mediating state properties to describe the world's dynamics as introduced in [17]. The core of both of these two

notions is that by some adaptive process, over time a (past) brain or world pattern leads to the formation of a present brain or world configuration that in turn drives the (future) brain or world pattern. From an informational perspective, this configuration in the present encodes emergent information based on the past that is relevant for the future. Choices made in the future are based on this information, by a person, or by the world. In this paper it has been shown how these processes can be modeled by an adaptive temporal-causal network.

For future work, it will be explored how the notion of Extended Mind [3, 6, 16] can be addressed in a similar manner and how a notion of representational content [4, 12] known from Philosophy of Mind can be used to describe the information in the emerging brain or world configurations. Moreover, it will be explored how also metaplasticity [1, 23] can be incorporated in the adaptive processes for criterial causation.

On purpose, in the current paper any link to notions such as the free will problem or the mental causation problem from Philosophy of Mind (as discussed extensively by Tse) has been left aside. The criterial causation perspective of Tse [25] has much value independent of such links (as also Levy [14] emphasizes), and that value has been the focus here. However, in future work, such philosophical links might be considered as well.

References

1. Abraham, W.C., Bear, M.F.: Metaplasticity: the plasticity of synaptic plasticity. *Trends Neurosci.* **19**(4), 126–130 (1996)
2. Ashby, W.R.: *Design for a Brain*. Chapman & Hall, London (1952). Revised edition 1960
3. Bosse, T., Jonker, C.M., Schut, M.C., Treur, J.: Simulation and analysis of shared extended mind. *Simul. J. (Soc. Model. Simul.)* **81**, 719–732 (2005)
4. Bosse, T., Jonker, C.M., Schut, M.C., Treur, J.: Collective representational content for shared extended mind. *Cogn. Syst. Res.* **7**, 151–174 (2006)
5. Chandra, N., Barkai, E.: A non-synaptic mechanism of complex learning: modulation of intrinsic neuronal excitability. *Neurobiol. Learn. Mem.* **154**(2018), 30–36 (2018)
6. Clark, A., Chalmers, D.: The extended mind. *Analysis* **58**, 7–19 (1998)
7. Cromwell, H.C., Tremblay, L., Schultz, W.: Neural encoding of choice during a delayed response task in primate striatum and orbitofrontal cortex. *Exp. Brain Res.* **236**(6), 1679–1688 (2018)
8. Descartes, R.: *The World*, Ch 6: Description of a New World, and on the Qualities of the Matter of Which it is Composed (1634)
9. Foster, J.M.: Unit activity in the prefrontal cortex during delayed response performance: neuronal correlates of short-term memory. *J. Neurophysiol.* **36**, 61–78 (1973)
10. Hebb, D.O.: *The Organization of Behavior: A Neuropsychological Theory*. Wiley, Hoboken (1949)
11. Hunter, W.S.: The delayed reaction in animals. *Behav. Monogr.* **2**, 1–85 (1912)
12. Kim, J.: *Philosophy of Mind*. Westview Press (1996)
13. Laplace, P.S.: *Philosophical Essays on Probabilities*. Springer, New York (1995). Translated by A.I. Dale from the 5th French edition of 1825 (1825)
14. Levy, N.: Review of P.U. Tse – the neural basis of free will: criterial causation. *Philos. Rev.* **33**(4), 331–333 (2013)

15. Tinklepaugh, O.L.: Multiple delayed reaction with chimpanzees and monkeys. *J. Comput. Psychol.* **13**, 207–243 (1932)
16. Tollesen, D.P.: From extended mind to collective mind. *Cogn. Syst. Res.* **7**, 140–150 (2006)
17. Treur, J.: Temporal factorisation: a unifying principle for dynamics of the world and of mental states. *Cogn. Syst. Res.* **8**(2), 57–74 (2007)
18. Treur, J.: Temporal Factorisation: Realisation of mediating state properties for dynamics. *Cogn. Syst. Res.* **8**(2), 75–88 (2007)
19. Treur, J.: Network-Oriented Modeling: Addressing Complexity of Cognitive, Affective and Social Interactions. Springer, Cham (2016). <https://doi.org/10.1007/978-3-319-45213-5>
20. Treur, J.: Multilevel network reification: representing higher order adaptivity in a network. In: Aiello, L., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., Rocha, L. (eds.) *Complex Networks and Their Applications VII, Proceedings of the Complex Networks 2018*, vol. 1. *Studies in Computational Intelligence*, vol. 812, pp. 635–651. Springer, Heidelberg (2018)
21. Treur, J.: The ins and outs of network-oriented modeling: from biological networks and mental networks to social networks and beyond. *Trans. Comput. Coll. Intell.* **32**, 120–139 (2019)
22. Treur, J.: Design of a Software Architecture for Multilevel Reified Temporal-Causal Networks (2019). <https://www.researchgate.net/publication/333662169>
23. Treur, J.: Modeling higher-order adaptivity of a network by multilevel network reification. *Netw. Sci.* (2019, in press)
24. Treur, J.: Network-Oriented Modeling for Adaptive Networks: Designing Higher-Order Adaptive Biological, Mental and Social Network Models. Springer, Heidelberg (2020, to appear)
25. Tse, P.U.: *The Neural Basis of Free Will: Criterial Causation*. MIT Press, Cambridge (2013)
26. Tse, P.U.: Two types of libertarian free will are realized in the human brain. In: Caruso, G. D., Flanagan, O.J. (eds.) *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience*, pp. 248–290. Oxford University Press (2018)
27. van Gelder, T.J., Port, R.F.: It's about time: an overview of the dynamical approach to cognition. In: Port, R.F., van Gelder, T. (eds.) *Mind as Motion: Explorations in the Dynamics of Cognition*, pp. 1–43. MIT Press, Cambridge (1995)



Network Influence Based Classification and Comparison of Neurological Conditions

Ruaridh Clark^{1(✉)}, Niia Nikolova², Malcolm Macdonald¹,
and William McGeown²

¹ Mechanical & Aerospace Engineering,
University of Strathclyde, Glasgow, UK
ruaridh.clark@strath.ac.uk

² School of Psychological Sciences and Health,
University of Strathclyde, Glasgow, UK

Abstract. Variations in the influence of brain regions are used to classify neurological conditions by identifying eigenvector-based communities in connectivity matrices, generated from resting state functional magnetic resonance imaging scans. These communities capture the network influence of each brain region, revealing that the subjects with Alzheimer's disease (AD) have a significantly lower degree of variation in their most influential brain regions when compared with healthy control (HC) and amnestic mild cognitive impairment (aMCI) subjects. Classification of subjects based on their pattern of influential regions is demonstrated with neural networks identifying HC, aMCI and AD subjects. The difference between these conditions are investigated by altering brain region influence so that a neural network changes a subject's classification. This conversion is performed on healthy subjects changing to aMCI or AD, and for aMCI subjects changing to AD. The results highlight potential compensatory mechanisms that increase or maintain functional connectivity in certain regions for those with aMCI, such as in the right parahippocampal gyrus and regions in the default mode network, but these same regions experience significant decline in those that convert from aMCI to AD.

Keywords: Functional connectivity · Community detection · Dementia

1 Introduction

Alzheimer's disease (AD) is the most common type of dementia. It is typically characterised by marked decline in episodic memory, with deficits occurring in other cognitive domains such as in language, visuospatial and executive functioning. Individuals with amnestic mild cognitive impairment (aMCI) present with impairments in memory, with other cognitive domains remaining relatively

intact. Although not all individuals with aMCI are in the early stages of AD, people with this pattern of symptomology are at high risk of conversion to the disorder [1].

There has been a large body of work documenting the changes in regional brain volume as the disease progresses, for example atrophy of the hippocampus is an indicator that subjects with aMCI will develop AD [2]. Whereas volumetric analysis does not inform on how brain dynamics are modified by AD, *functional magnetic resonance imaging* (fMRI) data may offer information on the regulation of brain networks and provide additional markers of disease. This is an important distinction, as the volume of the parahippocampal gyri (PHG) has been found to remain constant for those with aMCI regardless of whether they declined further into AD or not, but a greater extent of activation within the PHG was a reliable marker of future cognitive decline [3].

This paper will focus on functional connectivity analyses to differentiate between people with AD and aMCI. The approach taken here is to identify and compare the relative influence of brain regions. This influence is captured from resting state fMRI data that is converted into a network of brain regions, with connections weighted by the strength of their signal correlations [4]. A region's influence is determined using an eigenvector-based community detection, communities of dynamical influence (CDI), as introduced by Clark, Punzo and Macdonald [5]. For a directed graph, this influence represents the nodes that can rapidly lead the network to a new state of consensus. For an undirected graph, used here to represent brain region connectivity, information is lost on which nodes are leading or following. Therefore, influential nodes are either important sources or sinks for information in the network. The CDI method relies on the relationship between eigenvectors to determine the communities, where the most dominant eigenvectors form a coordinate system with communities displaying as an alignment of nodes from that system's origin. Eigenvectors have been used previously with the neuronal network of the *C. elegans* to identify brain circuitry [6]. CDI is a progression from normalised cuts [7], and other spectral bisection methods, as it considers a combination of eigenvectors before determining community assignment.

By associating each brain region with an influence ranking, based on which community it belongs to, a subject can be characterised based on the relative influence of their brain regions. Pattern recognition enables the detection and association of these rankings with different neurological conditions. Machine learning (ML) is a highly capable and popular method of recognising statistical patterns, where most of the research on classifying aMCI and AD with ML has focused on low-level features such as cortical thickness and/or grey matter tissue volumes from MRI, mean signal intensities from positron emission tomography (PET) and other common biomarkers [8]. There are also examples of graph theoretical metrics being used in combination with machine learning, such as [9] & [10] that employed a support vector machine (SVM) as well as [11] that used a naïve Bayes classifier to perform the classification. Of significant clinical utility, functional brain connectivity can be combined with machine learning into models capable of identifying neurological conditions such as AD [12], autism or depression.

In this paper, we describe a novel method of determining the influence of brain regions in neurological conditions, by identifying communities of dynamical influence in healthy controls, individuals with aMCI and patients with AD. We also determine how influence may change during conversion between these conditions, by developing a new technique to classify individuals based on the influence of their brain regions.

2 Methods

In order to classify and compare influential communities between the AD, aMCI and control groups, a connectivity matrix is first generated for a region-of-interest (ROI) set for each subject from their resting state fMRI data. This all-to-all connectivity matrix is reduced to only include significant connections by applying a threshold on the minimum weight for edges included in the network. This new topology undergoes eigenvector-based community designation, where the communities are ranked by their influence over the network. This ranking of nodes based on their community's influence is used to train neural networks to recognise whether a subject is healthy or has amnestic mild cognitive impairment (aMCI) or Alzheimer's disease (AD). To understand the changes required to convert a subject from one condition to another, the influence of their nodes are altered so that the neural network would change their classification to an alternative state. This alteration, to achieve this change in classification, captures the difference in the influence of ROIs when comparing between different neurological conditions.

2.1 Dataset

The data used here was from the 'Resting-state fMRI in Dementia Patients' dataset [13], obtained from the Harvard Dataverse database. The MRI data was obtained using a Siemens 3T MRI system (Magnetom Allegra, Siemens, Erlangen, Germany) for ten patients with a probable AD diagnosis (by NINCDS-ADRDA consensus criteria [14]), 10 aMCI patients [15] and 10 healthy elderly subjects (HC).

The subjects underwent a resting state EPI fMRI scan ($TR = 2080\text{ ms}$, $TE = 30\text{ ms}$, 32 axial slices parallel to AC-PC plane, matrix 64×64 , in plane resolution $3 \times 3\text{ mm}^2$, slice thickness = 2.5 mm , 50% skip, flip angle 70°). The duration of the scan was 7 min and 20 s, yielding 220 volumes. Subjects were instructed to keep their eyes closed throughout, refrain from thinking of anything in particular and to avoid falling asleep. An anatomical T1-weighted three dimensional MDEFT (modified driven equilibrium Fourier transform) scan was also acquired for each subject ($TR = 1338\text{ ms}$, $TE = 2.4\text{ ms}$, $TI = 910\text{ ms}$, flip angle = 15° , matrix = $256 \times 224 \times 176$, $FOV = 256 \times 224\text{ mm}^2$, slice thickness = 1 mm , total scan time = 12 min).

2.2 Preprocessing

The functional data was preprocessed using the CONN toolbox (CONN: functional connectivity toolbox, [16]) for SPM12 (www.fil.ion.ucl.ac.uk/spm) and MATLAB version 2018a.

Spatial Preprocessing. The first four volumes of the functional scans were removed in order to eliminate saturation effects and to allow the signal to stabilise. Functional data were slice-time adjusted and corrected for motion. The high resolution T1 weighted anatomical images were coregistered with the mean EPI image. They were segmented into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) masks and were spatially normalised to the Montreal Neurological Institute (MNI) space [17]. The obtained transformation parameters were applied to the motion corrected functional data, and an 8 mm FWHM Gaussian kernel was applied for spatial smoothing. It should be noted that the use of spatial smoothing on fMRI data can affect the properties of functional brain networks, including a possible over-emphasis of strong, short-range links, changes in the identities of hubs of the network, and decreased inter-subject variation [18].

Temporal Filtering. In order to mitigate physiological and movement-related noise, the aCompCor technique was used. aCompCor identifies and removes the first five principal components of the signal from the CSF and WM masks (eigenvectors of the PCA decomposition of the EPI timecourse averaged over the CSF and WM), as well as the motion parameters, their first-order temporal derivatives and a linear detrending term [19]. One subject's scan was excluded from the analysis due to excessive motion. Scrubbing and motion regression were also performed. The preprocessed functional data were then bandpass filtered ($0.008 \text{ Hz} < f < 0.09 \text{ Hz}$) using a fast Fourier transform (FFT).

2.3 Connectivity Matrix Generation

One hundred and thirty-two (132) ROIs were defined by the default CONN atlas which combines the FSL Harvard-Oxford cortical and subcortical areas and the AAL atlas cerebellar areas. Connectivity between the 132 ROIs was assessed for the 7-min resting state scan for each subject. We constructed 132×132 ROI-to-ROI correlation (RRC) matrices of Fisher z-transformed bivariate correlation coefficients (Pearson's r) using the ROIs described above. For each subject, a graph adjacency matrix $A(i, j)$ was computed by thresholding the RRC matrix $r(i,j)$ using a Cluster-Span Threshold (CST [20]).

2.4 Cluster-Span Threshold

An unbiased Cluster-Span threshold (CST) [20] was used in generating the adjacency matrix. CST is especially suitable as it performs well in distinguishing

functional connectivity between HC and AD subjects [21]. The threshold generates a topology that excludes edges with weights smaller than the chosen value. CST is selected so that the topology generated contains the same number of clustered triples and spanning triples.

2.5 Communities of Dynamical Influence

The network is assigned into Communities of Dynamical Influence (CDI) based on the connections and influence of nodes in the network. CDI are defined in [5] where community designation is achieved by using multiple (often three) eigenvectors to define a coordinate system. The nodes, which are further from the origin of this system than any of their connections, are defined as leaders of separate communities. Each of these communities is populated with other nodes that lie on a path that connects to the leader node of that community. Each node is assigned to only one community, the community is chosen by assessing which leader is most closely aligned to that node. This alignment is assessed by comparing the position vector, from the origin of the coordinate system, for the leader nodes and the node to be assigned. The dot product of these position vectors determines the leader that is best aligned to the node.

Once community designation is complete, the order of influence is determined by evaluating the largest entry of the most dominant eigenvector for each community. The dominant eigenvector of the connectivity matrix is known to be a nonnegative vector. The community that contains the node with the largest v_1 value is ranked as the most influential community, with the other communities ranked in descending order according to their largest v_1 . For each subject, a vector is produced that denotes the ranking of the community each node is in. This vector shall be referred to as the *Influence Vector* with values assigned to ROIs between 0 and 1, where these extremes mark the least and most influential community respectively.

In this paper, CDI is determined from the three most dominant eigenvectors of the undirected connectivity matrix after applying the CST. These are the eigenvectors associated with the largest eigenvalues in magnitude and are shown in [5] to identify the nodes that are most effective at driving the network to consensus.

2.6 Pattern Recognition

A neural network (NN) is employed to recognise patterns in the *influence vector* and associate these patterns with the subject status of healthy control (HC) or aMCI or AD. A two-layer, feed-forward, neural network was used since this simple architecture was shown to be capable of capturing sufficient information from the influence vector to produce accurate classifications. The NN employed sigmoid output neurons, scaled conjugate gradient backpropagation and 10 hidden neurons [22]. Each input vector, \mathbf{x} , is scaled to fit in the range $[-1, 1]$ and the performance is evaluated using cross-entropy [23], with the cost function, c , a mean of the individual values,

$$c = -\frac{1}{n} \sum_{i=1}^x y \ln(a) \quad (1)$$

where n is the total number of items of data in a set of inputs x , with the network output a and the desired output y .

To identify and reduce variance when using a small data set, five separate neural networks were trained with different compositions of training and validation sets. No test set was used due to the small size of the data set and the fact that the risk of over-fitting is not a major concern in the intended application. For a neural network to be trained successfully the cross-entropy cost function had to be below $c = 0.1$ with all networks reporting 100% accuracy from their confusion matrices. The five neural networks were trained on 29 subjects with the training sets varied between 21 and 25 subjects, where the validation set contained the remaining subjects and was subjects from all three of the classifications. Below 21 there was insufficient training data to effectively train the network. The mean of the five NN outputs was assessed and in Fig. 1 the variance between different NNs is also reported.

2.7 Optimisation of the Influence Vector

The neural network outputs a three element vector, a , with non-negative entries that represents the three possible subject conditions, where $\sum_{j=1}^3 a_j = 1$ and the largest element indicates the condition selected by the network. An optimiser, using a sequential quadratic programming method [24], is employed to alter the input vector so that the neural network identifies it as having a different condition. The optimiser aims to maximise a_j where j represents the target condition. The alteration to the input vector is recorded to identify the ROIs that were altered to change a subject from their current condition to another state.

3 Results

A large variability between subjects is observed, even when they are represented by their ROI community influence. For example, Table 1 shows that for each classification most of the ROIs are, for at least one of the subjects, placed in both the most and the least influential communities (top and bottom CDI respectively). Despite this variability, trends emerge when comparing the different classification in Table 1 with the percentage of ROIs displaying either an upward or downward trend from HC, through aMCI to AD in each row. One of the most notable results is that, for AD, 43% of the ROIs present in the top CDI, for at least one subject, were not in the bottom CDI for any other subject. Whilst only 53% of the 132 ROIs were present in the top CDI for at least a single AD subject. This means that around 81% of the ROIs, in the most influential community for at least one AD subject, are not in the least influential community for any other AD subject. This is a far greater consistency than seen from the aMCI ($\sim 29\%$)

and HC ($\sim 15\%$) subjects. To a less significant extent the HC subjects present the inverse pattern, whereby they have the highest number of ROIs that are present in the bottom CDI and not included in the top CDI for any other HC subject (38%), compared to aMCI (24%) and AD (14%).

Table 1. The number of ROIs that are included in the most (top) and least (bottom) influential communities of dynamical influence (CDI) are shown alongside the number of ROIs that are in the top community for at least one subject but not the bottom for any others and vice versa.

	% of ROIs		
	HC	aMCI	AD
Present in bottom CDI	71	79	84
Present in bottom CDI & not in top CDI	27	19	12
Present in top CDI	85	77	53
Present in top CDI & not in bottom CDI	13	22	43

3.1 Altering Influence in Healthy Control Subjects

A mean alteration was produced from alterations generated using five separate neural networks, as described in the Methods section. Three different conversions are detailed here; Ten HC subjects were converted to the classification of aMCI, ten HC subjects were also converted to AD, and finally ten aMCI subjects were converted to the classification of AD. The results of these conversions highlight the ROIs that undergo the greatest changes in connectivity and influence when a person is affected by MCI and/or AD, see Fig. 1(a), (b) & (c). These three conversions were also investigated in the other direction, eg. converting from AD to HC, with the results producing similar findings to those detailed herein and so not reported separately.

Parahippocampal Gyri. The hippocampus and parahippocampal regions are well known to have a significant role in memory formation [3], which is an area that notably declines in those with aMCI and AD. In [25] the parahippocampal gyri were noted to decrease in functional connectivity in AD subjects when compared with aMCI subjects. Our results support these findings for the right PHG, where it loses significant influence in the conversion from HC to AD and from aMCI to AD, in Fig. 1(d). The anterior and posterior divisions of the right PHG reporting some of the largest alterations in these conversion (see Fig. 1(a) and (c) respectively). Interestingly in [3], subjects with greater clinical impairment were found to rely more on their right parahippocampal gyrus (PHG). Those with a reliance on the right PHG were also those whom declined the most over 2.5 years of clinical follow-up. It was hypothesised that this extra reliance on the right PHG could be a marker for impending clinical decline. This compensatory mechanism appears to be visible in the results of Fig. 1(d) where the anterior

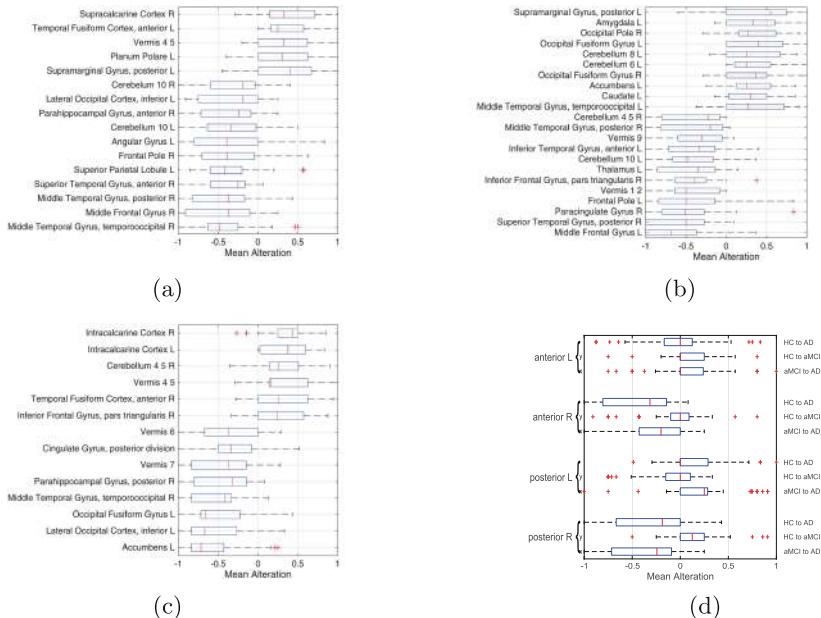


Fig. 1. The largest alterations of ROI influence (± 1.5 z-score) required to change the classification of (a) 10 HC subjects to AD, (b) 10 HC subjects to aMCI, and (c) 10 aMCI subjects to AD. In (d) the alterations to the parahippocampal regions are reported for each conversion. A positive alteration indicates that a ROI's influence has increased. The mean alteration is assessed from five conversions using different neural networks. The median, 25th and 75th percentile are detailed with the whiskers extending to the most extreme data points. Outliers lie more than three scaled median absolute deviations away from the median and are excluded.

division of the right PHG gains influence when converting HC subjects to aMCI with a mean alteration z-score of 0.83.

The behaviour of the hippocampus mirrors that of the parahippocampal regions in the HC to AD conversions. The right hippocampus loses a significant amount of influence (z-score of -1.18) while the left side gains some influence (z-score of 0.41). This finding is supported by [26] where disrupted connectivity between the right hippocampus and several brain regions was seen for AD subjects, whilst connectivity between the left hippocampus and the prefrontal cortex was relatively increased. In [27] a more extensive connectivity disruption was found in both sides of the hippocampus, where the contrast in results to [26] was attributed to greater severity of AD in subjects studied in [27] with hippocampal connectivity thought to decline progressively during the disease.

Default Mode Network. Two prominent regions in the default mode network (DMN), the posterior cingulate cortex and the precuneus cortex, have frequently been identified as early markers in AD [28]. Compared to controls, AD patients

have been shown to exhibit lower connectivity in the precuneus and posterior cingulate cortex within the DMN [29].

Our findings support [29] where the posterior cingulate cortex loses influence in the HC to AD case with a z-score of -0.96 , while in the HC to aMCI case it gains some influence with a z-score of 0.49 . It is therefore not surprising that this region loses significant influence in the aMCI to AD conversion in Fig. 1(c) with a z-score of -1.53 . In the conversions from HC to AD and from HC to aMCI, the precuneus cortex becomes notably less influential, indicated by a negative mean alteration with a z score of -1.14 and -1.00 , respectively. These results indicate that both ROIs are clear indicators of AD, but this analysis suggests their decline appears to occur at different stages of the disease with results suggesting that the precuneus cortex declines earlier as depicted in Fig. 2.

Furthermore, a network comparison in [30] between controls and AD highlighted a decrease in functional connectivity between the DMN and posterior cingulate gyrus, precuneal cortex, and lateral occipital cortex (LOC). Our results support this finding where, in Fig. 1(a), the left inferior division of the LOC loses significant influence in the HC to AD conversion (z-score of -1.53) while the right side only changes slightly (z-score of -0.16). An even more significant swing in influence is seen, in Fig. 1(b), for the LOC's left inferior division in the aMCI to AD conversion (z-score of -2.28). This ROI presents a similar pattern to the posterior cingulate cortex by gaining influence in the HC to aMCI conversion (z-score of 0.99), but losing influence in both conversions to AD. It is also worth noting that the superior divisions of the LOC presents the opposite changes in influence but to a less significant degree than the left inferior division.

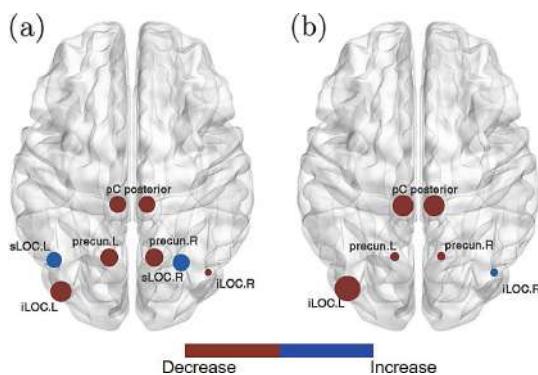


Fig. 2. Superior view of mean alterations to influence of default mode network and lateral occipital cortex (LOC). Circle size proportional to z-score for (a) HC to AD and (b) aMCI to AD conversion. ROIs shown: posterior cingulate (pC), Precuneus (Precun), inferior and superior LOC (iLOC & sLOC) with L and R indicating left and right respectively.

Other key DMN ROIs are seen to lose influence in the conversions, with the left Angular gyrus and middle frontal gyri displaying prominently in Fig. 1(a) and (b).

3.2 Calcarine

The calcarine, sensorimotor and anterior cingulate regions have been shown to be spared significant damage until the late stages of AD [31]. Our results indicate that by avoiding significant damage the intracalcarine and supracalcarine regions gain greater influence. The right supracalcarine cortex tops the positive alterations in Fig. 1(a) and both sides of the intracalcarine cortex top the positive alterations in Fig. 1(c). However, the bi-lateral intracalcarine cortex and the left supracalcarine cortex lose influence when converting from HC to MCI, so avoiding damage might not provide the whole picture for why these calcarine regions become more influential in the HC to AD conversion.

4 Conclusions

This paper presents a new approach to understanding the impact of changes in brain region connectivity, which are brought about by neurological conditions. Brain region assignment to influence ranked communities reveals that those with AD have a significantly higher degree of commonality in their most influentially connected regions, when compared with HC and aMCI subjects. The communities identified as influential therefore appear to be shared to a greater extent in AD subjects, compared to the aMCI and HC.

The detection of brain region influence enables the classification of conditions and captures the patterns of functional changes that lead to aMCI and AD. The right parahippocampal gyrus (PHG) is confirmed as playing a key role in the decline to AD. In particular, the results supported findings of compensatory activity in this region, where it was seen to maintain, if not gain, some influence in the conversion from HC to aMCI. The right PHG then experienced a significant decline for the conversion to an AD classification. The results supported previous findings on the importance of the default mode network (DMN) in the development of aMCI and AD. Additionally, the posterior cingulate gyrus and the lateral occipital cortex are noted as promising indicators of future conversions from aMCI to AD. Finally, the calcarine is confirmed as a region that defies decline in those with AD. Interestingly, the intra- and supracalcarine regions are seen to decline for those with aMCI but their influence then increases for subjects with AD.

References

1. Dawe, B., Procter, A., Philpot, M.: Concepts of mild memory impairment in the elderly and their relationship to dementia—a review. *Int. J. Geriatr. Psychiatry* **7**(7), 473–479 (1992)

2. Apostolova, L.G., Dutton, R.A., Dinov, I.D., Hayashi, K.M., Toga, A.W., Cummings, J.L., Thompson, P.M.: Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. *Arch. Neurol.* **63**(5), 693–699 (2006)
3. Dickerson, B.C., Salat, D.H., Bates, J.F., Atiya, M., Killiany, R.J., Greve, D.N., Dale, A.M., Stern, C.E., Blacker, D., Albert, M.S., Sperling, R.A.: Medial temporal lobe function and structure in mild cognitive impairment. *Ann. Neurol.* **56**(1), 27–35 (2004). <https://doi.org/10.1002/ana.20163>
4. Friston, K.J.: Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* **2**, 56–78 (1994). <https://doi.org/10.1002/hbm.460020107>
5. Clark, R., Punzo, G., Macdonald, M.: Network communities of dynamical influence. *arXiv* (2019). [arXiv:1908.10129](https://arxiv.org/abs/1908.10129)
6. Varshney, L.R., Chen, B.L., Paniagua, E., Hall, D.H., Chklovskii, D.B.: Structural properties of the *Caenorhabditis elegans* neuronal network. *PLoS Comput. Biol.* **7**(2), e1001066 (2011)
7. Shi, J., Malik, J.: Normalized cuts and image segmentation. *Departmental Papers (CIS)*, p. 107 (2000)
8. Suk, H.I., Lee, S.W., Shen, D., A.D.N. Initiative: Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct. Funct.* **220**(2), 841–859 (2015)
9. Hojjati, S.H., Ebrahimzadeh, A., Khazaee, A., Babajani-Feremi, A., A.D.N. Initiative: Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. *J. Neurosci. Methods* **282**, 69–80 (2017)
10. Khazaee, A., Ebrahimzadeh, A., Babajani-Feremi, A.: Application of advanced machine learning methods on resting-state fMRI network for identification of mild cognitive impairment and Alzheimer's disease. *Brain Imaging Behav.* **10**(3), 799–817 (2016)
11. Khazaee, A., Ebrahimzadeh, A., Babajani-Feremi, A., A.D.N. Initiative: Classification of patients with MCI and AD from healthy controls using directed graph measures of resting-state fMRI. *Behav. Brain Res.* **322**, 339–350 (2017)
12. Forouzannezhad, P., Abbaspour, A., Fang, C., Cabrerizo, M., Loewenstein, D., Duara, R., Adjouadi, A.: A survey on applications and analysis methods of functional magnetic resonance imaging for Alzheimer's disease. *J. Neurosci. Methods* **317**, 121–140 (2019). <https://doi.org/10.1016/j.jneumeth.2018.12.012>
13. Mascali, D., DiNuzzo, M., Gili, T., Moraschi, M., Fratini, M., Maraviglia, B., Serra, L., Bozzali, M., Giove, F.: Resting-state fMRI in dementia patients (2015). Harvard Dataverse. <https://doi.org/10.7910/DVN/29352>
14. McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E.M.: Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology* **34**, 939–944 (1984). <https://doi.org/10.1212/wnl.34.7.939>
15. Petersen, R.C., Doody, R., Kurz, A., Mohs, R.C., Morris, J.C., Rabins, P.V., Ritchie, K., Rossor, M., Thal, L., Winblad, B.: Current concepts in mild cognitive impairment. *Arch. Neurol.* **58**, 1985–1992 (2001). <https://doi.org/10.1001/archneur.58.12.1985>
16. Whitfield-Gabrieli, S., Nieto-Castanon, A.: Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* **2**, 125–141 (2012). <https://doi.org/10.1089/brain.2012.0073>
17. Ashburner, J., Friston, K.J.: Unified segmentation. *NeuroImage* **26**, 839–851 (2005). <https://doi.org/10.1016/j.neuroimage.2005.02.018>

18. Alakörkkö, T., Saarimäki, H., Glerean, E., et al.: Effects of spatial smoothing on functional brain networks. *Eur. J. Neurosci.* **46**(9), 2471–2480 (2017)
19. Behzadi, Y., Restom, K., Liau, J., Liu, T.T.: A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007). <https://doi.org/10.1016/j.neuroimage.2007.04.042>
20. Smith, K., Azami, H., Parra, M.A., Starr, J.M., Escudero, J.: Cluster-span threshold: an unbiased threshold for binarising weighted complete networks in functional connectivity analysis. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), vol. 147, pp. 2840–2843. IEEE (2015). <https://doi.org/10.1109/EMBC.2015.7318983>
21. Smith, K., Abasolo, D., Escudero, J.: A comparison of the cluster-span threshold and the union of shortest paths as objective thresholds of EEG functional connectivity networks from Beta activity in Alzheimer's disease. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2826–2829. IEEE, August 2016
22. Mathworks: nprtool: Neural Net Pattern Recognition tool (r2019a) (2019). <https://uk.mathworks.com/help/deeplearning/ref/nprtool.html>. Accessed 16 Aug 2019
23. Mathworks: crossentropy: Neural Network performance (r2019a) (2019). <https://uk.mathworks.com/help/deeplearning/ref/crossentropy.html>. Accessed 16 Aug 2019
24. Mathworks: fminunc Unconstrained Minimization (r2019a) (2019). <http://uk.mathworks.com/help/optim/ug/fminunc-unconstrained-minimization.html>. Accessed 16 Aug 2019
25. Gili, T., Cercignani, M., Serra, L., Perri, R., Giove, F., Maraviglia, B., Caltagirone, C., Bozzali, M.: Regional brain atrophy and functional disconnection across Alzheimer's disease evolution. *J. Neurol. Neurosurg. Psychiatry* **82**(1), 58–66 (2011). <https://doi.org/10.1136/jnnp.2009.199935>
26. Wang, L., Zang, Y., He, Y., Liang, M., Zhang, X., Tian, L., Wu, T., Jiang, T., Li, K.: Changes in hippocampal connectivity in the early stages of Alzheimer's disease: evidence from resting state fMRI. *NeuroImage* **31**(2), 496–504 (2006)
27. Allen, G., Barnard, H., McColl, R., et al.: Reduced hippocampal functional connectivity in Alzheimer disease. *Arch. Neurol.* **64**(10), 1482–1487 (2007). <https://doi.org/10.1001/archneur.64.10.1482>
28. Rombouts, S.A.R.B., Barkhof, F., Goekoop, R., Stam, C.J., Scheltens, P.: Altered resting state networks in mild cognitive impairment and mild Alzheimer's disease: an fMRI study. *Hum. Brain Mapp.* **26**(4), 231–239 (2005). <https://doi.org/10.1002/hbm.20160>
29. Binnewijzend, M.A., Schoonheim, M.M., Sanz-Arigita, E., Wink, A.M., van der Flier, W.M., Tolboom, N., Adriaanse, S.M., Damoiseaux, J.S., Scheltens, P., van Berckel, B.N., Barkhof, F.: Resting-state fMRI changes in Alzheimer's disease and mild cognitive impairment. *Neurobiol. Aging* **33**(9), 2018–2028 (2012). <https://doi.org/10.1016/j.neurobiolaging.2011.07.003>
30. Hafkemeijer, A., Möller, C., Doppen, E.G., Jiskoot, L.C., Schouten, T.M., van Swieten, J.C., van der Flier, W.M., Vrenken, H., Pijnenburg, Y.A., Barkhof, F., Scheltens, P.: Resting state functional connectivity differences between behavioral variant frontotemporal dementia and Alzheimer's disease. *Front. Human Neurosci.* **9**, 474 (2015). <https://doi.org/10.3389/fnhum.2015.00474>
31. Brun, A., Englund, E.: Regional pattern of degeneration in Alzheimer's disease: neuronal loss and histopathological grading. *Histopathology* **5**(5), 549–564 (1981)



Characterization of Functional Brain Networks and Emotional Centers Using the Complex Networks Techniques

Richa Tripathi¹(✉), Dyutiman Mukhopadhyay², Chakresh Kumar Singh¹, Krishna Prasad Miyapuram¹, and Shivakumar Jolad³

¹ Indian Institute of Technology Gandhinagar, Gandhinagar 382355, Gujarat, India
richa.tripathi@iitgn.ac.in

² University College London, London WC1E 6BT, UK

³ Flame University, Pune 412115, India

Abstract. In this work, we construct functional networks of the human brain using the coherence measure on the EEG time-series data, in response to external audio-visual stimuli. These stimuli were nine different movie clips selected to evoke different emotional states. The constructed networks for each emotion were characterized using network measures such as clustering coefficient, small worldness, the efficiency of information propagation, etc. in different frequency bands corresponding to brain waves. We used a community detection algorithm to infer the segregation of functional correlations in the brain into modules. Further, using the variation of information measure, we compare and contrast the modular organizations of different brain networks. We observe that the different brain networks are closest in their organization into modules in alpha frequency band while they farther apart in other bands. We identified crucial network nodes or hubs using centrality measure, and find that most of the hubs were common for all networks and belong to a specific location on the brain map. In summary, our work demonstrates the utilization of the network theoretical and statistical tools for understanding and differentiating different brain networks corresponding to the perception of varieties of emotional stimuli.

Keywords: Brain networks · Functional connectivity · Modular organization · Hubs

1 Introduction

In the recent years, analysis of brain networks constructed from EEG and advanced brain imaging techniques has contributed to the understanding of the complex structure and functioning of the human brain [8,39] and has given clues

S. Jolad—Part of the work carried at: Indian Institute of Technology Gandhinagar, Gandhinagar-382355, India.

to understanding its higher cognitive exhibits such as emotions and reasoning abilities [26, 28, 35]. The functional brain network studies have given credible evidence against the locationist approach of functional segregation of brain regions, in favor of the constructionist approach where different emotions involve brain circuits comprising of brain areas not specific to a particular emotion [27, 35, 36]. Emotional stimuli affect large scale functional brain networks which can be evaluated in terms of parameters such as node betweenness and network efficiency [37]. The human brain can be decomposed into multiple, distinct, and interacting networks such as salience network, executive control network and task-negative network, and emotional stimuli can differentially affect these sub-networks [28, 37, 45].

Studies on emotion, based on the analysis of single-electrode level EEG in the frequency domain have demonstrated the association of emotion with asymmetric activity in the frontal brain in the alpha band. It has also been demonstrated that different patterns of functional connectivity are associated with different emotional states in either single or combined frequency bands ascertaining distinct response patterns of the central nervous system to different emotional stimuli [26]. In previous studies on classifying emotions, researchers have largely focused only on a few contrasting dimensions such as threat-safe, sadness-happiness, positive-negative-neutral and with a small number of recording sites of EEG activity [12, 26, 30]. Also, the film-based studies on emotions and functional responses of the brain have been previously carried out primarily based on the western classification of emotions which are happiness, sadness, fear, anger, surprise, disgust [13, 14]. However, audio-visual stimuli such as film viewing are capable of evoking sentiments which might be due to the interplay of the basic emotions (dominant states), and transitory and temperamental states of emotions. In a recent study [9], authors reported the existence of 27 different emotional experiences based on reports of emotional states elicited by a large number of emotionally evocative videos. Many of these emotional experiences were not discrete but were linked to each other through a smooth gradient.

In this paper, we aim to use tools from network theory to explore any consistent patterns or differences in brain network structures when participants are evoked with various emotional stimuli. The nine dimensions of emotions studied here correspond to the sentiments or *Rasa*'s described in the *Natyashastra* evoked through audio-visual stimuli such as performing arts (here movie clips) [16]. The *Rasas* can be understood as a superposition of basic emotional states (see methods for details) such as anger, disgust, fear, happiness, sadness, and surprise as described by Ekman [14]. We have analyzed the patterns of brain activity and examined whether these activities show specific neural signatures in viewers, in different frequency bands corresponding to the brainwaves [42]. Different network measures were used to characterize the structural and functional differences among these emotional states. Modular organization of functional connectivity of the brain for different *Rasas* networks were extracted using community detection algorithms. These networks were then compared using information flow measure to quantify the difference between their modular organizations, across all frequency bands. Central nodes for information propagation were identified

using the leverage centrality measure for each of the networks, enabling us to find the most significant nodes for emotion perception and segregate designated areas for processing of a particular emotion. The current work, though relates to *Rasa* theory and emotional states, it offers a generic approach to characterizing brain networks (corresponding to different tasks), and in different frequency wavebands.

2 Data and Methods

Prior to conducting EEG experiments, ethical clearance was taken from the Institute Ethical Committee (IEC) of the Indian Institute of Technology, Gandhinagar. Informed consent was obtained from all the participants before conducting experiments.

2.1 Subjects

Participants were 20 healthy, right-handed students from the Indian Institute of Technology Gandhinagar (mean age: 26 years, 16 males; 4 females). All of the participants were proficient in Hindi and English languages. They were all informed about the task and were asked to remain attentive while watching the film clip. An independent ranking of many movie clips corresponding to each category of emotion was done by a small number of subjects. Only those clips were selected which were ranked best in evoking a particular response for all the categories.

2.2 *Rasa* and *Natyasastra*: A Background

A major source of the Indian system of classification of emotional states comes from the ‘*Natyasastra*’ [44], the ancient Indian treatise on the performing arts, which dates back to 2nd Century AD (pg. LXXXVI: [16]). The ‘*Natyasastra*’ speaks about ‘sentiments’ or ‘*Rasas*’ (pg. 102: [16]) which are produced when certain ‘dominant states’ (*sthayi bhava*), ‘transitory states’ (*vyabhicari bhava*) and ‘temperamental states’ (*sattvika bhava*) of emotions come together (pgs. 102, 105: [16]). This *Rasa* theory, which is still widely followed in classical Indian performing arts, classifies eight *Rasas* or sentiments which are: *Sringara* (erotic), *Hasya* (comic), *Karuna* (pathetic), *Raudra* (furious), *Veera* (heroic), *Bhayanaika* (terrible), *Bibhatsa* (odious) and *Adbhuta* (marvelous). There was a later addition of the ninth sentiment or *Rasa* called *Santa* (peace) in later Sanskrit poetics (pg. 102: [16]). We drew inspiration from this classification system and selected movie clips corresponding to each of the nine *Rasas*. Please note that, as there are no standard movie clips for this classification system, our selection of movies is one possible set and is not stringent in any manner.

2.3 Western Emotional Classification and *Rasa* Theory

According to Western version of emotion classification by Ekman, there are basic or universal emotions [13–15] which are happiness, sadness, fear, anger, surprise, disgust. However, there are also background emotion sets which are: wellbeing-malaise; calm-tense; pain-pleasure; [10] as well as self-referential social emotions which are: embarrassment, guilt, shame, jealousy, envy, empathy, pride, admiration. [5, 11, 18, 25, 31, 41]. Also, there are pioneering works of scientists like Lisa Feldman Barrett [2] who question Ekman’s concepts of discreteness of emotions. One can find startling similarities between the *Rasa* theory (its concepts of the generation of *Rasas* from the *Bhavas*) with the works of Panksepp and Kagan [23, 32–34]. However, there has been very little previous work done on the perception or brain science of emotional states based on *Rasa* theory mostly due to the lack of awareness regarding the science of the *Rasa* theory among the scientific community. One behavioural study was conducted by Hejmadi [20], which investigated the identification of these emotions across cultures. An image processing study was conducted in Ref. [40] for investigating the variations in facial features based on nine *Rasas*.

2.4 Movie Clips

Complex, naturalistic stimuli like film-viewing evoke highly reliable brain activity across viewers as per current research works [19]. In this work, we used nine film clips from Bollywood films (popular Indian Hindi language cinema) made between 1970’s to the current time (see Table 1). Independent rating of the movies was done on a small number of subjects. The length of the movie clips varied from 42 s to 2 min 37 s. Length of the film segments were not kept constant since the clips included a flow of narrative which was necessary to be shown till a certain length of time (different for each segment) to evoke a specific Rasa.

Table 1. Table for the movie clips corresponding to each emotion

<i>Rasa</i> genre	Film name	Year	Duration	Start time	End time
<i>Adbhuta</i>	Mr. India	1987	1m 48 s	1 h 1 m 40 s	1 h 3 m 28 s
<i>Bhayanka</i>	Bhoot	2003	1 m 34 s	1 h 2 m 57 s	1 h 4 m 31 s
<i>Bibhatsa</i>	Rakhta Charitra	2010	1 m 12 s	43 m 55 s	45 m 7 s
<i>Hasya</i>	3 Idiots	2009	2 m 33 s	59 m 55 s	1 h 2 m 28 s
<i>Karuna</i>	Kal Ho Naa Ho	2003	2 m 37 s	2 h 47 m 41 s	2 h 50 m 18 s
<i>Raudra</i>	Ghajini	2008	2 m 9 s	2 h 38 m 43 s	2 h 40 m 52 s
<i>Santa</i>	Zindagi Na Milegi Dobara	2011	2 m 22 s	48 m 22 s	50 m 44 s
<i>Sringara</i>	Umrao Jaan	1981	42 s	43 m 08 s	43 m 50 s
<i>Veera</i>	Lagaan	2001	2 m 3 s	2 h 10 m 57 s	2 h 13 m

2.5 EEG Experimental Procedure

We conducted the EEG experiment on the participants during which they were asked to watch chosen film clips representative of nine different *Rasa* Genres. The electrical activity of the brain was recorded using 128 channel high-density Geodesic EEG SystemsTM with a sampling frequency of 250 Hz. A representative diagram of the node placement (along with node numbers from 1 to 128) used in the present work for brain network visualization, is shown in Fig. 1(a). The figure also depicts the electrode allocations to anatomical regions of the brain. The following abbreviations are used: F, P, O and T stand for Frontal, Parietal, Occipital and Temporal lobes; L and R represent Left and Right regions.

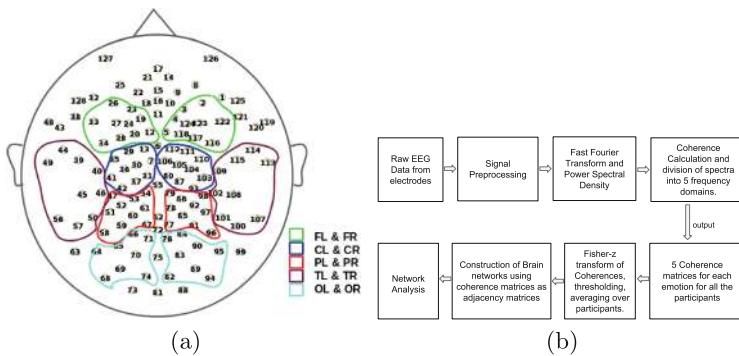


Fig. 1. (a) A brain map of the electrode positions of the EEG cap. All the Brain networks and extracted communities in the present work follow this Node placement. (b) Workflow structure governing EEG data recording, pre-processing, network construction and analysis.

The participants were shown all the film clips in random order. A white fixation cross with a black screen for ten seconds was shown before each clip. Initial few seconds of the time series recordings were neglected before analysis to avoid major fluctuations due to adjustment. Signals beyond 60 Hz frequencies were filtered out to avoid noise effects. The design of the experiment was done in E-primeTM and synced with Net-stationTM acquisition software. The entire study on the collected data from the EEG experiment is summarised in Fig. 1(b).

3 Construction of Brain Network

Standard time series analysis measures such as Coherence [7], Phase Synchronization [46], Mutual Information [38], and Granger Causality [21] offer a huge depth into understanding the synchrony and information propagation in the functional brain. We use Electroencephalographic Coherence that is used as a metric for deriving functional brain connectivity i.e. the degree of association

between any two brain regions (whose electrical activity is recorded by electrodes in the signal space). The Coherence measure is sensitive to both amplitude and phase changes and holding values between 0 (nil coherence) and 1 (complete coherence), it compares similarity of the power spectra (measured in microvolts squared, μV^2) of the time-series recorded by these electrodes. The regions having highest coherence are assumed to be the most synchronized functionally and vice-versa. The coherence measure (Eq. 1) between two time series X and Y is defined as,

$$C_{XY}(f) = \frac{|G_{XY}(f)|^2}{|G_{XX}(f)G_{YY}(f)|}, \quad (1)$$

where $G_{XY}(f)$ is the cross-power spectral density and $G_{XX}(f)$ and $G_{YY}(f)$ are the respective auto-power spectral densities [43].

The time series data (recorded by the electrodes, in microvolts, μV) was transformed to frequency domain via Fast Fourier Transform (FFT) and the power in different frequency bands [42]: δ (1–4 Hz), θ (4 - 8 Hz), α (8–12 Hz), β (12–40 Hz) and γ (40–60 Hz), corresponding to different brain waves was calculated. The coherence spectra, thus decomposed into five frequency bands resulted in five (128×128) coherence matrices and these were obtained for each of the nine stimuli. The frequency decomposition of the spectrum was crucial to estimate the power in each of the bands, capturing the brain state. Apart from power in individual bands, we also used average coherence across the whole spectrum in the analysis. The coherence matrix was mapped to a weighted adjacency matrix, and the corresponding brain network was constructed with $N = 128$ nodes. In this network, the nodes are the electrode locations on the scalp and edges are connections between them measured by the coherence values. For the overall networks construction in each frequency band, corresponding to each *Rasa*, the edges weights (coherences) representing functional correlations are Fischer Z-transformed and then averaged over all the participants. Thus, a total of 45 (9×5) weights networks were constructed, where weighted edges represent the strength of coherences between the nodes.

4 Brain Network Characterization

Functional properties of complex networks are largely determined by the statistical properties of its structure. We have calculated six well-known network measures on the networks: Clustering Coefficient (CC), Characteristic Path Length (L), Network Density (d), Local and Global efficiency (LE, GE), and Small-worldness index (SW) using Area Under Curve [3] method. The Area Under the Curve (AUC) method ensures that networks are automatically thresholded and significant edge weight ranges are retained. In the AUC procedure, two cutoffs (one upper and one lower) on the edge strengths, are consistently determined for each network, such that a range of important connections are retained, and only important network structure is dealt with. The upper bound in this threshold range, κ_+ , is an edge weight at which the network is fully connected and the lower bound, κ_- is the threshold at which the network just gets disconnected.

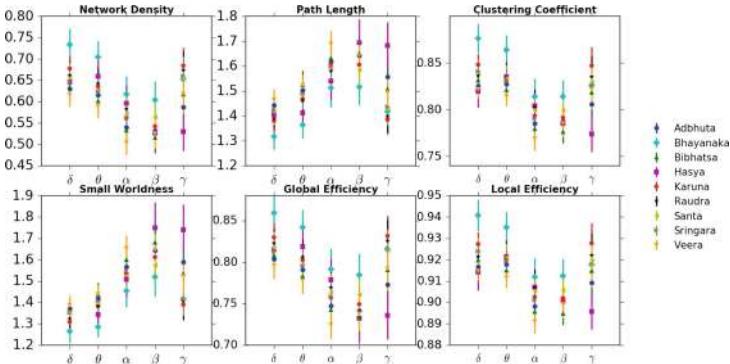


Fig. 2. Network measures Network Density (d), Path Length (L), Clustering Coefficient (CC), Small Worldness, Global Efficiency (GE), and Local Efficiency (LE) (see text for details) of brain networks corresponding to different *Rasas* across the five frequency bands ($\delta, \theta, \alpha, \beta$ and γ) along with their error bars.

The network is binarised at each threshold within this range, retaining only edges with weights larger than the threshold value and eliminating the remaining ones and network metric (M) is evaluated at each connection density. This results in a curve showing the variation of M with edge weights lying in the threshold range. The curve was then integrated over all these thresholds to yield the AUC value of the metric M .

In Fig. 2, we plot the six aforementioned measures for all the *Rasas* networks across all frequency bands. Network density (d) for all the networks has values $d > 0.45$, indicating that they are dense. Average shortest path length lies between 1.2 to 1.8, indicating the ease of information propagation. Network density decreases monotonically with increasing frequency in different bands, except in the gamma band - where it switches to higher values than its previous frequency band (beta). Conversely, we see a rise in path length values with increasing frequency, until the beta band, after which it drops in gamma band. Clustering coefficient for all networks shows small variations from 0.75 to 0.9. Gamma band has greater spread (over different networks) in the network metrics than all other bands. *Veera* and *Bhayanaka* networks show the maximum difference across δ, θ, α , and β bands. SW is greater than one for all *Rasas*, indicating deviation from randomness as well as absolute regularity of connections as in the brain. Global efficiency of the networks is high - ranging from 0.7 to 0.87. Local efficiency is even higher - 0.88 to 0.95, indicating high robustness to the failure of nodes. The trend (increase or decrease of network measures with frequency) is flipped at high frequency i.e in the γ band. This is seen for all the *Rasas* and for each network metric, suggesting a dynamic reorganisation of structure at high frequency.

5 Results of Network Analysis

Network analysis of the functional networks performed in the present study consists of five elements as explained in the following sections.

5.1 Community Structure

Quantitative analysis of brain using complex networks measures has revealed the presence of highly connected hubs and significant modular architecture [1], apart from showing small-worldness. Modules are functionally specialised and spatially localised groups of nodes that function together in unison to integrate the information globally that they process locally.

In Fig. 3, we show community structure of brain networks corresponding to all the *Rasas*. Note that the networks show only top 10% of the edges. This was done for better visualization, as the unthresholded network is fully connected and hence very dense. The community structure was extracted using an algorithm based on the modularity optimization, in Gephi [4]. All the networks had modularity value, $Q \geq 0.58$, and had five to six communities in each case. Broadly, the community structure appears to be similar across all networks, with one major community in the frontal lobes (C1), two in the left and right hemisphere's central/parietal brain regions (C2, C3), one that encompasses a large area in the visual cortex in the occipital areas (C4) and two other smaller communities along the left and right temporal regions (C5, C6). The *Bibhatsa Rasa* shows split of the community in the parietal lobe. The communities C4 and C5 are merged into one for most of the graphs.

For visualisation of dominant communities in the brain networks of all the *Rasas*, we used Gephi software which utilises a modularity optimisation algorithm [6] for detection of communities from the brain networks and Geolayout scheme for node positioning.

5.2 Distance Between Networks

To compare the similarities and differences in the network's modular organization (community structure), we used the Variation of Information (VI) measure [24]. The VI measure signifies the difference in the modular organisation of the networks, summarising the differences in segregation of coherences between nodes. For calculation of VI measure, the community structure on the overall networks was evaluated using best partition routine of python library community [17].

In Fig. 4, we show the VI measure based distance matrices for all *Rasas* networks' modular organization in different frequency bands. Networks that are farthest from each other are shown in brighter cell colour ($VI \simeq 0.5$), and those which are closest are shaded dark ($VI \simeq 0.0$). Six VI measure matrices were obtained - one corresponding to overall spectrum and five other corresponding to each of frequency bands (see Fig. 4). It can be seen that alpha frequency band shows maximum similarity in modular organization of all the brain networks.

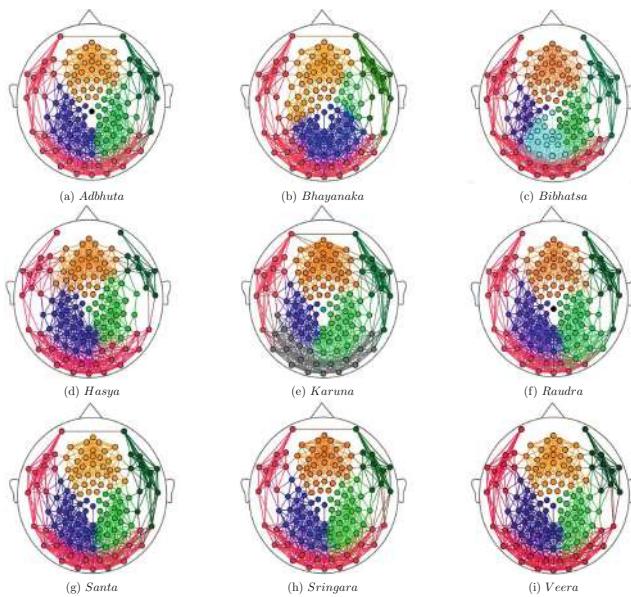


Fig. 3. Brain Networks for various *Rasas* with 10% highest edge weight edges, organised into clear community structure. For all the *Rasas* the network is dissociated into 5 to 6 communities.

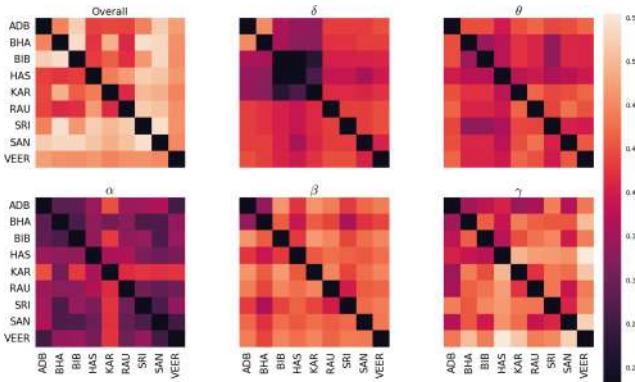


Fig. 4. Variation of Information matrices for various *Rasas* in different frequency bands arranged in order **overall**, δ , θ , α , β and γ , from left to right and top to bottom.

For all the pairs of *Rasas* networks, the VI falls in the range 0.22 to 0.56. The colour bar (common for all matrices) range is chosen to represent the maximum variation in VI. We observe that, for the full network, the dynamic range of VI variation is much small, hence we cannot infer much about the similarity or dissimilarity of different network organizations. In the lowest frequency δ band - *Bibhatsa*, *Hasya* and *Karuna* networks are closest to each other. Maximum

distance is between *Adbhuta* and *Bhayanka* networks. For θ band - *Bhayanka*, *Bibhatsa*, *Sringara* networks are relatively closest. Others are far from each other. The α band is the most distinctive frequency band where all the *Rasas* are much closer to each other (except the *Karuna* network). The α band activity for all the networks, except *Karuna* seems indistinguishable from each other. For β band, the pairs *Adbhuta* and *Bhayanka* and *Raudra*, and *Sringara* are the closest to each other, and *Hasya* and *Adbhuta* are the farthest. In the γ band, *Veera* network is farthest from *Santa*, *Hasya* and *Bhayanka* network. *Adbhuta* is much closer to *Bhayanka* and *Karuna* networks.

5.3 Hub Identification

For brain networks, work by Joyce *et al.* suggests that Leverage centrality [22] is computationally cheaper and more accurate for identifying hubs than other centrality measures (as revealed by their Receiver Operating Characteristic curves). Leverage centrality also incorporates information about local node neighbourhood, such that a node with high positive leverage centrality is more impactful to its neighbours so that its neighbours draw more information from it than any other nodes in their neighbourhood. In contrast, a negative leverage centrality node is influenced more by its neighbours than being an influencer node. In the present study, we designate central nodes as nodes with the largest positive and negative leverage centralities.

Leverage centrality (defined in Eq. 2),

$$l_v = \frac{1}{k_v} \sum_{N_v} \frac{k_v - k_w}{k_v + k_w} \quad (2)$$

measures the relationship between the degree of a vertex k_v and the degree of each of its neighbors k_w , averaging over all the neighbors N_v . For the weighted network, as in our case, degrees are weighted degrees.

In Fig. 5, are shown a brain map with coloured nodes having the highest (top 12) positive and highest negative (top 12) leverage centralities. Colour coding is based on the *Rasas* as shown in the figure. Nodes shown in pink are consistently central across all the *Rasas*. For the positive centrality case, they all lie in the periphery of the brain map, mostly in the parietal brain regions indicating that most important information relay centres for the perception of emotions are located at the parietal regions of the brain. The central nodes with most negative leverage centralities are in the frontal and central brain regions demonstrating that nodes drawing maximum information from their neighbours are in these regions.

5.4 Edge Weight Distribution

Edge weights denote the strength of connections between the nodes. The cumulative edge weight distribution of all *Rasa* networks is shown in Fig. 6. We find that edge weights are considerably high for δ , θ and α bands for all networks

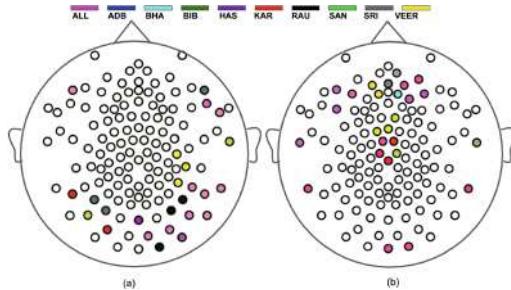


Fig. 5. Brain map showing nodes with (a) highest positive leverage centralities (b) highest negative leverage centralities for all the *Rasas*.

and lower for α and β bands. Through this we infer that functional connectivity is stronger in δ , θ , and γ bands whereas the pathways have lesser average information flow in α and β frequency regimes. We also observe that *Raudra*, *Bhayanka* and *Hasya* networks have higher correlations specially in δ , θ and α bands as compared to other *Rasas* networks.

6 Discussion

In this work, we have used concepts and tools from complex networks to understand the functional connectivity and structural organization of the brain subjected to audio-visual stimuli. To the best of our knowledge, this is the first attempt to design a network analysis probe to identify neural signatures for different *Rasas*. Our main findings, across all Rasa genres are (1) community structure in the alpha band is most similar and (2) lower synchrony of nodes in higher frequency bands and vice versa. Previous studies have reported higher correlations for negative or stress full visual stimuli [29] for lower frequency bands. In line with this, our results indicate higher coherence in these bands for *Raudra* and *Bhayanka* *Rasas* emotions (considered as negative emotions). We have identified the community structure of the brain for different *Rasas* at the level of strongest links. We find the existence of four dominant communities consistently across all *Rasas* and localisation of hubs on brain map.

There are a few limitations to our current work which we would like to highlight here. Our observations are based on the analysis of the signal space of electrodes placed on the scalp which does not have a clear source mapping inside the brain. Hence, we refrain from making inferences on the source space and only present the analysis in the signal space. However, EEG based correlations have been successfully used to extract functional connectivities as well to distinguish different emotional states [26], even at the signal space. Also, we used only one movie clip corresponding to each *Rasa* that was ranked best to elicit a particular emotion. There can be other sets of movie clips for each category. Hence, there is a need for better standardization audio-visual stimuli for *Rasas* used. We also did not distinguish between the emotional perception ability between classes

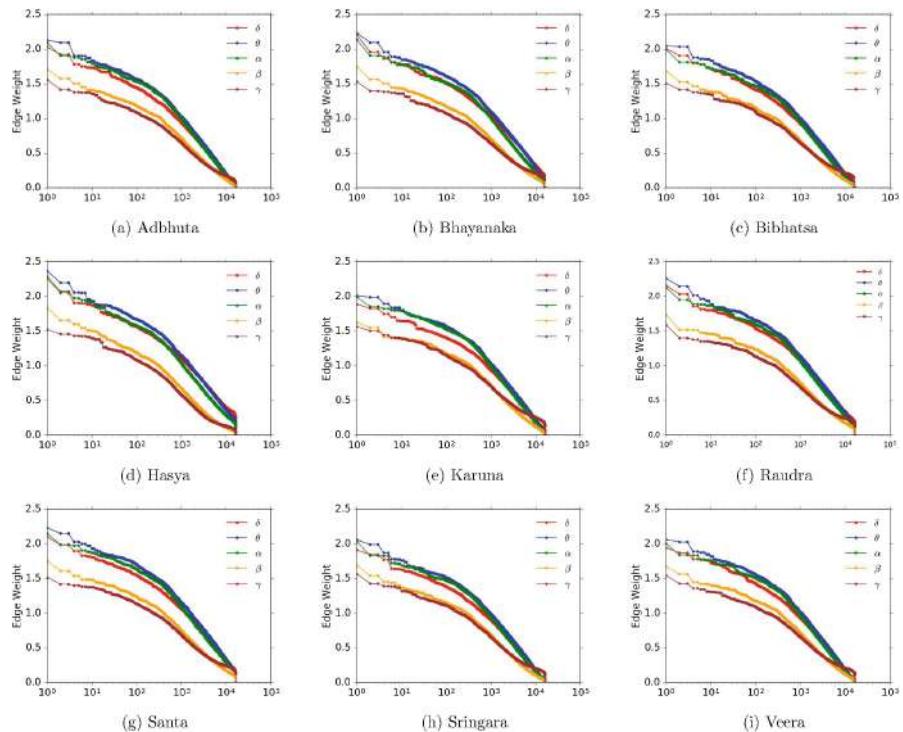


Fig. 6. Edge weight distribution matrix for various Rasas in different frequency bands.

such as gender and age group. We hope that the current work sets a precedent for using network tools for a more detailed analysis of the neural signatures of emotions based on audio-visual stimuli.

References

1. Achard, S., Salvador, R., Whitcher, B., Suckling, J., Bullmore, E.: A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci.* **26**(1), 63–72 (2006)
2. Barrett, L.F.: Are emotions natural kinds? *Perspect. Psychol. Sci.* **1**(1), 28–58 (2006)
3. Bassett, D.S., Meyer-Lindenberg, A., Achard, S., Duke, T., Bullmore, E.: Adaptive reconfiguration of fractal small-world human brain functional networks. *Proc. Nat. Acad. Sci.* **103**(51), 19518–19523 (2006)
4. Bastian, M., Heymann, S., Jacomy, M., et al.: Gephi: an open source software for exploring and manipulating networks. *ICWSM* **8**(2009), 361–362 (2009)
5. Bennett, M., Gillingham, K.: The role of self-focused attention in children's attributions of social emotions to the self. *J. Genet. Psychol.* **152**(3), 303–309 (1991)
6. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. *J. Stat. Mech.: Theory Exp.* **2008**(10), P10008 (2008)

7. Bowyer, S.M.: Coherence a measure of the brain networks: past and present. *Neuropsychiatr. Electrophysiol.* **2**(1), 1 (2016)
8. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**(3), 186 (2009)
9. Cowen, A.S., Keltner, D.: Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Nat. Acad. Sci.* **114**(38), E7900–E7909 (2017)
10. Damasio, A.R.: The feeling of what happens: body and emotion in the making of consciousness. *N. Y. Times Book Rev.* **104**, 8–8 (1999)
11. Diener, E., Suh, E.M., Lucas, R.E., Smith, H.L.: Subjective well-being: three decades of progress. *Psychol. Bull.* **125**(2), 276 (1999)
12. Dmochowski, J.P., Sajda, P., Dias, J., Parra, L.C.: Correlated components of ongoing EEG point to emotionally laden attention - a possible marker of engagement? *Front. Hum. Neurosci.* **6**, 112 (2012)
13. Ekman, P.: Expression and the nature of emotion. *Approaches Emot.* **3**, 19–344 (1984)
14. Ekman, P.: An argument for basic emotions. *Cogn. Emot.* **6**(3–4), 169–200 (1992)
15. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. *J. Pers. Soc. Psychol.* **17**(2), 124 (1971)
16. Ghosh, M.: The Natyasastra Ascribed to Bharata Muni, vol. 1. Asiatic Society of Bengal, Calcutta (1951)
17. Hagberg, A., Swart, P., Chult, D.S.: Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Laboratory (LANL), Los Alamos (2008)
18. Hareli, S., Eisikovits, Z.: The role of communicating social emotions accompanying apologies in forgiveness. *Motiv. Emot.* **30**(3), 189–197 (2006)
19. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R.: Intersubject synchronization of cortical activity during natural vision. *Science* **303**(5664), 1634–1640 (2004)
20. Hejmadi, A., Davidson, R.J., Rozin, P.: Exploring hindu indian emotion expressions: evidence for accurate recognition by Americans and Indians. *Psychol. Sci.* **11**(3), 183–187 (2000)
21. Hesse, W., Möller, E., Arnold, M., Schack, B.: The use of time-variant eeg granger causality for inspecting directed interdependencies of neural assemblies. *J. Neurosci. Methods* **124**(1), 27–44 (2003)
22. Joyce, K.E., Laurienti, P.J., Burdette, J.H., Hayasaka, S.: A new measure of centrality for brain networks. *PLoS ONE* **5**(8), e12200 (2010)
23. Kagan, J.: Temperament and the reactions to unfamiliarity. *Child Dev.* **68**(1), 139–143 (1997)
24. Karrer, B., Levina, E., Newman, M.E.: Robustness of community structure in networks. *Phys. Rev. E* **77**(4), 046119 (2008)
25. Leary, M.R., Baumeister, R.F.: The nature and function of self-esteem: sociometer theory. In: *Advances in Experimental Social Psychology*, vol. 32, pp. 1–62. Elsevier (2000)
26. Lee, Y.Y., Hsieh, S.: Classifying different emotional states by means of EEG-based functional connectivity patterns. *PLoS ONE* **9**(4), e95415 (2014)
27. Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., Barrett, L.F.: The brain basis of emotion: a meta-analytic review. *Behav. Brain Sci.* **35**(3), 121–143 (2012)

28. McMenamin, B.W., Langeslag, S.J., Sirbu, M., Padmala, S., Pessoa, L.: Network organization unfolds over time during periods of anxious anticipation. *J. Neurosci.* **34**(34), 11261–11273 (2014)
29. Miskovic, V., Schmidt, L.A.: Cross-regional cortical synchronization during affective image viewing. *Brain Res.* **1362**, 102–111 (2010)
30. Nie, D., Wang, X.W., Shi, L.C., Lu, B.L.: EEG-based emotion recognition during watching movies. In: 2011 5th International IEEE/EMBS Conference on Neural Engineering (NER), pp. 667–670. IEEE (2011)
31. Oatley, K., Johnson-Laird, P.N.: Towards a cognitive theory of emotions. *Cogn. Emot.* **1**(1), 29–50 (1987)
32. Panksepp, J.: Affective neuroscience of the emotional brainmind: evolutionary perspectives and implications for understanding depression. *Dialogues Clin. Neurosci.* **12**(4), 533 (2010)
33. Panksepp, J., Normansell, L., Cox, J.F., Siviy, S.M.: Effects of neonatal decortication on the social play of juvenile rats. *Physiol. Behav.* **56**(3), 429–443 (1994)
34. Parrott, W.G.: *Emotions in Social Psychology: Essential Readings*. Psychology Press, Philadelphia (2001)
35. Pessoa, L.: On the relationship between emotion and cognition. *Nat. Rev. Neurosci.* **9**(2), 148 (2008)
36. Pessoa, L., Adolphs, R.: Emotion processing and the amygdala: from a ‘low road’ to ‘many roads’ of evaluating biological significance. *Nat. Rev. Neurosci.* **11**(11), 773 (2010)
37. Pessoa, L., McMenamin, B.: Dynamic networks in the emotional brain. *Neurosci.* **23**(4), 383–396 (2017)
38. Salvador, R., Martinez, A., Pomarol-Clotet, E., Sarro, S., Suckling, J., Bullmore, E.: Frequency based mutual information measures between clusters of brain regions in functional magnetic resonance imaging. *NeuroImage* **35**(1), 83–88 (2007)
39. Sporns, O., Chialvo, D.R., Kaiser, M., Hilgetag, C.C.: Organization, development and function of complex brain networks. *Trends Cogn. Sci.* **8**(9), 418–425 (2004)
40. Srimani, P., Hegde, R.: Analysis of facial expressions with respect to Navarasas in Bharathanatyam styles using image processing. *Int. J. Knowl. Eng.* **3**(02), 193–196 (2012)
41. Tangney, J.P.E., Fischer, K.W.: *Self-conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*. Guilford Press, New York (1995)
42. Tatum, W.O.: Ellen R. grass lecture: extraordinary EEG. *Neurodiagn. J.* **54**(1), 3–21 (2014)
43. Thatcher, R., Krause, P., Hrybyk, M.: Cortico-cortical associations and eeg coherence: a two-compartmental model. *Electroencephalogr. Clin. Neurophysiol.* **64**(2), 123–143 (1986)
44. Thirumalai, M.: An introduction to Natya Shastra-gesture in aesthetic arts. *Lang. India* **1**(6), 27–33 (2001)
45. Menon, V., Toga, A.W.E.: Salience network. In: *Brain Mapping: An Encyclopedic Reference*. Academic Press (2015)
46. Varela, F., Lachaux, J.P., Rodriguez, E., Martinerie, J.: The brainweb: phase synchronization and large-scale integration. *Nat. Rev. Neurosci.* **2**(4), 229 (2001)



Topological Properties of Brain Networks Underlying Deception: fMRI Study of Psychophysiological Interactions

Irina Knyazeva^{1,3,4(✉)}, Maxim Kireev^{1,2}, Ruslan Masharipov²,
Maya Zheltyakova^{1,2}, Alexander Korotkov², Makarenko Nikolay^{3,4},
and Medvedev Svyatoslav²

¹ Saint-Petersburg State University, St. Petersburg, Russia
iknyazeva@gmail.com

² N.P. Bechtereva Institute of the Human Brain, Russian Academy of Sciences,
St. Petersburg, Russia

³ Institute of Information and Computational Technologies, Almaty, Kazakhstan

⁴ Central Astronomical Observatory at RAS, St. Petersburg, Russia

Abstract. In the current study, we used topological data analysis of fMRI data for exploring neurophysiological mechanisms underlying the execution of deceptive actions. We used the results of the analysis of psychophysiological interactions (PPI) of fMRI data, obtained during an earlier experiment where subjects were required to mislead an opponent through sequential execution of deceptive and honest claims. A connectivity matrix based on PPI analysis was processed with the methods of algebraic topology. With this approach, we confirmed our previous findings that the increase in local activity and psychophysiological interactions of the left caudate nucleus is associated with the execution of deceptive actions. It is also in line with our hypothesis that involvement of the left caudate nucleus in brain processing of deception reflects the process of activation of error detection mechanism. In contrast to this finding, the right caudate nucleus was most frequently observed in the selected cliques associated with honest actions in comparison with deceptive ones. This observation points to possible differential role of left and right caudate nuclei in processing deceptive and honest actions, so it can be further investigated in future research. Topological analysis of higher-order organization of functional interactions revealed three cycles encompassing different sets of brain regions. Those regions are associated with executive control, error detection and sociocognitive processes, involvement of which in deception execution was hypothesized in previous studies. The fact of observation of such loops of functionally integrated brain regions demonstrates the possibility of parallel functioning of above-mentioned mechanisms and substantially extends the current view on neurobiological basics of deceptive behavior.

Keywords: Deception · Topological data analysis · Network neuroscience · Psychophysiological interactions · Brain networks

1 Introduction

Nowadays, mainly basing on the data concerning changes in the levels of local activity or functional coupling between particular brain regions, we know a lot regarding properties of their involvement in ongoing activity. But still the understanding of how those brain regions work together to produce investigated activity is quite poor. Although raising of investigations of functional interactions between brain regions is promising, the majority of studies aimed at revealing organization of functional brain systems utilizes analysis of pair-wise interactions. Therefore it hinders revealing of higher-order properties of functional organization of brain networks, like circuitry structure of their interactions, and hampers uncovering an issue of how brain systems work. One fruitful and actively developing approach assumes application of algebraic topology which allows to reveal unique organizational properties of network activities at different levels of structural and functional brain organization, i.e. from brain regions to neuronal populations [27]. Although this approach presents new analytical possibilities its full potential is still to be uncovered. Especially revealing the cycles of functional interactions might be useful for those types of complex forms of behavior which are associated with a number of parallel, simultaneously involved processes or mechanisms. One of the examples of such multifaceted behavior is deception which is innate to human beings. The brain basis of deception is hard to investigate because execution of deceptive action can be associated with a number of hypothetical processes like inhibition of predominant honest action [7], action selection [14], conflict monitoring and error detection [15, 16], as well as sociocognitive processes needed for inferring mental state of the opponent [18, 30]. Although those processes can be involved simultaneously, there are only a few studies aimed at revealing interregional interactions between brain regions underlying deceptive behavior with the usage of functional MRI [1, 14, 19, 32] or electrophysiological data [10, 31]. And even less studies addressed an issue about topological properties of neuronal network activity by applying graph theoretical approach [9, 33]. The possibility of simultaneous involvement of above mentioned cases in brain processing of deception can be assessed via application of algebraic topology analysis allowing to reveal so called cycles - mathematical representation of neuronal circuits of functionally integrated brain regions [27]. Taking into account the lack of such studies, in the current research we aimed to fill this gap by applying methods of algebraic topology to fMRI data revealed in the settings of deception execution [15]. To analyse topological structure of functional interregional interactions we utilized the data about psychophysiological interactions between predefined ROI set covering whole brain [26]. First of all it was expected that deception should be characterised by greater number of interacted nodes (both at the level of cliques and cycles). Based on previous research, we focused on a number of brain regions located within prefrontal cortex and caudate nuclei, which previously demonstrated dominant involvement in functional interactions underlying deception execution [14, 16, 32]. It was also expected that in accordance with widely reported greater involvement of brain regions comprising fronto-parietal executive control network in deception, these regions will

be integrated in higher-order topological structures of interacting brain regions. And, finally, if deception assumes involvement of sociocognitive processes [18, 30] needed for inferring mind states of others [25] while deceiving in the settings of social interaction, brain regions pertaining to so called theory of mind (ThOM) brain system [20] should be more involved in the higher-order interactions.

2 Experiment Design

Participants. Twenty-four healthy volunteers (14 women and 10 men) participated in the study. All participants were native Russian speakers, 19–44 years of age, with no history of neurological or psychological disorders. All subjects were also right-handed, as assessed by the Edinburgh Handedness Inventory [23]. The participants were given no information about the specific purpose of the study. All subjects provided written informed consent prior to the study, and were paid for their participation. All procedures were conducted in accordance with the Declaration of Helsinki and were approved by the Ethics Committee of the N.P. Bechtereva Institute of the Human Brain, Russian Academy of Sciences.

Study Design. In the MRI scanner participants played an interactive “I doubt it” game. They were given an interface with two arrows oriented up or down and they were instructed to send the information about arrow orientation to an opponent (computer program). Subjects were free in deciding to send honest or dishonest information to an opponent by pressing the corresponding buttons. Afterwards the opponent decided to accept or refuse the offer regarding arrow orientation sent by the participants. At the end of a probe the participants were presented with the opponent decision together with the amount of payoff or penalty. Monetary reward was applied for accepted-deceptive or refused-truthful claims. Likewise monetary penalties were incurred for accepted-honest and refused-deceptive claims. Therefore the primary goal of the experimental task was to mislead the computer opponent in trial by trial manner. Subjects were also presented with catch trials in which they were instructed to press buttons in strict accordance with the direction of presented arrows. In case of inappropriate button pressing in catch trials subjects were monetary penalized. Consequently there were three types of experimental trials: honest claims (HC), deceptive claims (DC), and control catch trials (See Fig. 1).

3 fMRI Data Acquisition, Preprocessing and gPPI-analysis

fMRI Data Acquisition. Magnetic resonance imaging was performed using a 3 Tesla Philips Achieva (Philips Medical Systems, Best, The Netherlands). Structural images were acquired using a T1-weighted pulse sequence (T1W-3D-FFE; repetition time [TR] = 2.5 ms; TE = 3.1 ms; 30° flip angle), measuring 130 axial slices (field of view [FOV] = 240 × 240 mm; 256 × 256 scan matrix) of 0.94 mm

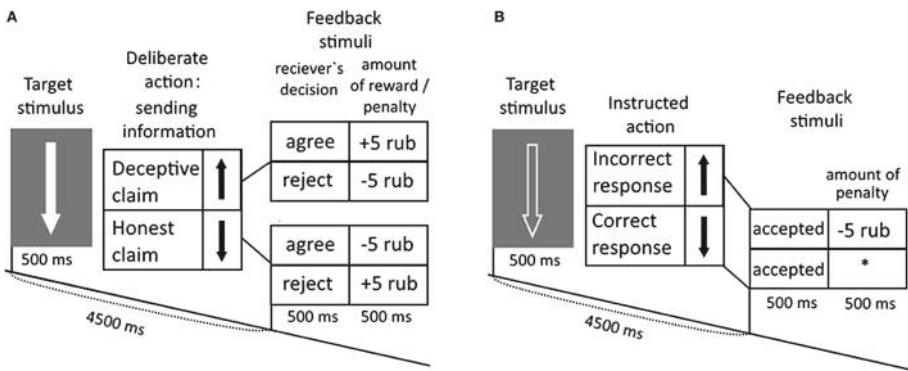


Fig. 1. Experimental task. (A) Subjects were instructed to interact with the opponent by falsely or honestly claiming arrow orientation presented on the monitor. (B) During catch trials, subjects were instructed to press buttons in accordance with the direction of the arrows.

thickness. Functional images were obtained using an echo planar imaging (EPI) sequence (TE = 35 ms; 90° flip angle; FOV = 208 × 208 mm; 128 × 128 scan matrix). Thirty-two continuous 3.5-mm-thick axial slices (voxel size = 3 × 3 × 3.5 mm), covering the entire cerebrum and most of the cerebellum, were oriented with respect to the structural image. The images were acquired using a TR of 2000 ms. Image pre-processing and statistical analyses of the fMRI data were performed using SPM12 software ([Statistical parametric mapping](#)). The data obtained for each subject was spatially realigned to the first functional image. To avoid effects from differences in the time of acquisition for each slice, slice-time correction was applied. The resulting functional images were spatially normalized to a standard stereotactic MNI template (Montreal Neurological Institute) and smoothed (using a Gaussian filter, 8 mm full-width at half-maximum). To prevent head motions of participants “Philadelphia” cervical MRI-compatible collar was used.

gPPI Analysis. Psychophysiological interactions (PPI) were analysed by application of the generalized PPI toolbox [21]. We chose the set of 300 functionally-defined brain regions of interest (ROIs) based on two popular ROI sets (Power’s 264 ROIs and Gordon’s 333 surface parcels) with improved representation of the subcortex and cerebellum [26]. The ROIs were defined as spheres, with a radius of 4 or 5 mm. The ROIs which did not overlap with the group mask from voxel-wise activation analysis [15] were discarded. As a result, 207 ROIs were used in further analysis. GLM for each participant were created in accordance with the following procedure. The blood oxygen level-dependent (BOLD) time series were extracted from each of 207 selected ROIs and further processed for creating PPI-regressors. At the deconvolution step, the neuronal activity underlying the observed BOLD changes within each ROI was mathematically estimated [11]. Estimated parameters for neuronal activity were then multiplied by vectors

describing experimental “on-times” that corresponded to events of interest with zero durations. Resulting vectors were subsequently convolved with a hemodynamic response function [6]. In addition to PPI-regressors (i.e., psychophysiological interaction terms), the GLM contained the following regressors, which were used as ignored variables: (1) six regressors that modeled BOLD signal changes induced by DC, HC, Catch and subsequent feedback stimuli (such as those in conventional subtractive GLM analysis, typically described as psychological variables); (2) trials without responses and wrong button presses in Catch trials; (3) motion parameters; and (4) the time series corresponding to BOLD signal changes within the selected ROI to exclude context-dependent hemodynamic changes. Both PPI and BOLD regressors (used as ignored variables in the current analysis) were modeled with zero durations. To reveal changes in functional coupling between DC and HC conditions, t-contrast between consistent PPI-regressor coefficients were calculated. To get task-related functional connectivity matrices across 207 ROIs the resulted beta coefficients were symmetrized by averaging lower and upper diagonal elements [8].

4 Topological Analysis of PPI Data

A connectivity matrix based on gPPI analysis, which was described in previous chapter, has been constructed for each individual, then averaged across persons. These matrices contain information about PPIs observed during deception process. From the mathematical point of view connectivity matrix could be considered as a weighted graph. Weighted graphs are harder to analyze than binary ones. Usually some mapping from weighted to binary representation is used for analysing [22]. One of the common practices is to use some threshold taking only nodes with the significant connections (or weights in weighted matrix) [12]. As a result we need to work with a sparse matrix, which is always good, but selecting a significant level is always problematic because of the boundary values. One approach to overcome this difficulty is to take several thresholds. However, in this approach we still discard a lot of the information contained in the edge weights. To overcome this it was proposed to use filtrations, which record the results of every possible binarization of the network or dense subset of the binarizations, along with the associated threshold values [12]. The use of filtration retains all of the information about the original weighted networks. Structures, appeared during filtration, can be investigated with the framework of simplicial complex, which could describe interaction of arbitrary large subgroup of nodes (regions of interest). In algebraic topology there is a well developed mathematical apparatus for quantitative analysis of such simplicial complexes. Simplicial complexes framework can be used for capturing higher-order interactions and has already been successfully applied in a series of papers on brain connectivity [3, 12, 17]. One of the advantages pointed out by the authors of these works [3, 12, 28] is the ability to study structures involved in interactions at different levels through the arrangement of strongly connected groups of nodes. This provides us with the language for describing interplay between strong and weak connections in

the system. All-to-all connected brain regions are called cliques. They form complex structure of simplicial complices, inside which so-called topological cavities ([27]) of different dimensions appear, around which information may flow. These cavities link regions from densely connected pattern in long loops and allow us to capture evolution of these loops through the predefined parameter at the different levels of filtration. These loop-like structures were reported as crucial features in the human brain structural architecture [28].

In present study, we utilized two semi-positive weighted matrices. One was related to greater PPI parameters of deceptive condition to honest condition ($DC > HC$) and the other related to its reversed contrast ($HC > DC$).

As a filtration we used dense subset of thresholds to construct a sequence of binary graphs, each included in the next. For each binary graph at each level of threshold all cliques were computed, which resulted in a set of filtration levels and corresponding cliques. Duplicated cliques appearing at lower levels of filtration were eliminated. After application of topological tools to this structure of PPI data we received clique sets and topological loops of dimension one and two. Firstly, we analysed the size and structure of cliques at each level of filtration for connectivity matrix. Secondly, loops or topological cavities size one and two were examined.

For demonstration of the whole process, small toy example of weighted networks is provided in Fig. 2.

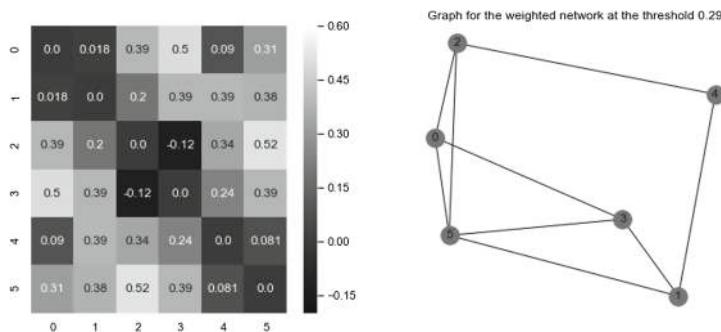


Fig. 2. Small example of weighted network from gPPI data with only 6 nodes (left) and graph representation for one threshold

Filtration process along with analysis of simplices and computation of cycles is shown in Fig. 3. We started merging nodes from the highest weights (strongest connections) to the lowest. During this process, if the connection appears, it will never disappear, but it will be included in different structures during the evolution. Here only the case with 6 nodes was considered, but the procedure was exactly the same for full network of 207 nodes. We began with the highest filtration level (0.52 in this small matrix (see Fig. 2)). At this level we observed separate nodes. Then continued with the threshold [0.49] and at this level two connections

appeared between nodes [0–3] and [2–5]. At level 0.39 link [1–4] was added. Two-node connections are called 1-cliques. The interesting part begins from level 0.29, at this level links [0–2], [0–5], [1–5], [1–3] appeared with several 2-cliques (triadic connections) and 1-cycle, consisting of intervals [1, 4]+[1, 5]+[2, 4]+[2, 4]. At level 0.09 two extra links appeared, which led to destruction of the 1-cycle and appearance of the new 2-cycle or the void, bordered by 2-cycles (triangles). Filtration plays key role in the analysis of evolution of topological features. At each level of filtration we can receive a topological description of a structure in terms of cliques and cavities (in the language of algebraic topology they are also named homology numbers or betti numbers). But with the introduction of filtration, we can track the evolution of these structures, that's why this analysis is named persistent homology. This process can be visualized with the use of so-called barcode plots. For our toy network such plot is represented at the lower right corner in Fig. 3. This plot has three blocks: one for connected components, one for 1-dimensional cycles and one for 2-dimensional cycles. At the part with connected components (Dim 0) we can see 6 bars started at level 0 (separate nodes) and then at each level of filtration we can track changes in connectivity. At the 1-cycle part we can see that 1-cycle started with the level of filtration 3 and died at level 4, and then at level 5 one 2-cycle or void appeared. It is only one way for persistent homology representation, it can also be done with persistence diagram, where each element corresponds to one cycle with the birth and death time or even with persistence landscapes [5]. Usually these representations and different functions from them are used for description of properties of complex structures, like we use statistics for random processes. But in our case we are not so interested in statistics, we are more interested in nodes involved in different structures. We extracted all cycles with the information about the nodes, participating in cycles, and after that analysed them. Now there is a variety of different software to perform topological data analysis, such as (JavaPlex, PHAT, Perseus, Dionysis, TDA, RIPSER, GUDHI, DIPHA, etc.). It makes topological data analysis available along with other methods. Roadmap for persistent homology computation can be found in the paper [24] with the analysis of performance and scalability. We used JavaPlex [29] for specific reasons. In JavaPlex we get an annotated barcode collection as an output, where each cycle is accompanied by a list of nodes, involved in the cycle. It is also very simple to use. As an input for JavaPlex we used precomputed cliques at each level of filtration.

Taking into account previous PPI-findings demonstrating involvement of the brain regions located in the prefrontal cortex and caudate nucleus in deception execution [14, 32], in the current analysis we focused on two ROIs located bilaterally in the head of caudate nuclei and two prefrontal regions within the left inferior frontal gyrus. Statistically significant involvement of these four regions in deception execution was observed in our previous above mentioned studies. In accordance with that, all cliques and loops were filtered in such a way, that only those were taken into account which include at least one of the considered regions. Main results are summarized in the next section.

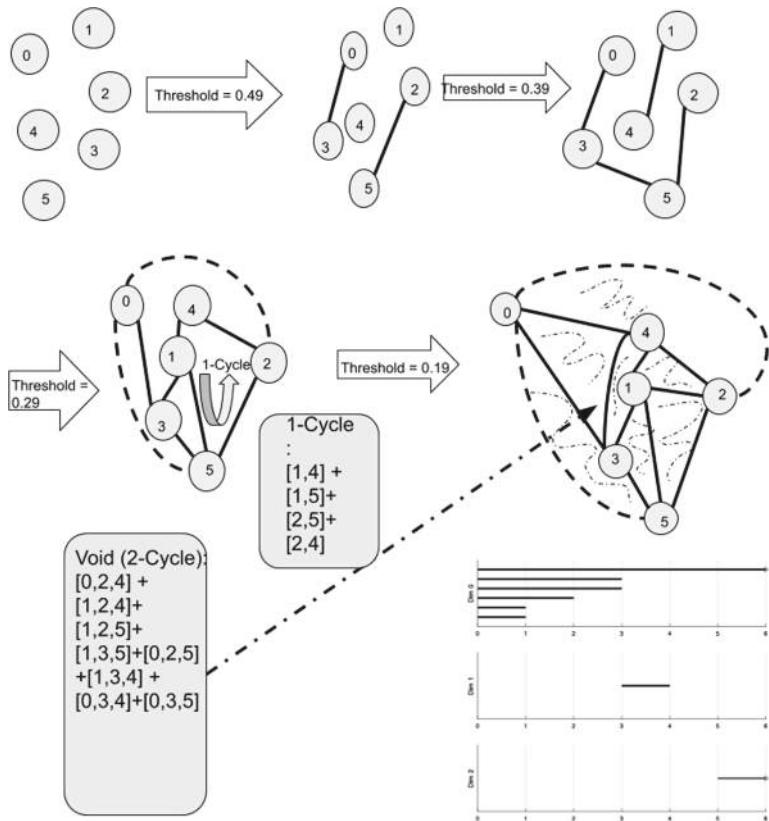


Fig. 3. Process of filtration and simplicial complexes evolution for small weighted network in Fig. 2. In the lower right corner barcode collection with the persistent connected components and cycles of dimension 1 and 2 is provided.

5 Results and Discussion

For each subject over a group of 24 people, gPPI connectivity matrix based on 207 ROIs time series was computed. For topological data analysis we used the connectivity matrix averaged over the group. So as an input for analysis we had 207×207 matrix.

Clique Analysis. For clique analysis we filtered only the cliques with the predefined nodes, specifically two bilateral caudate nuclei and two prefrontal regions within the inferior frontal gyrus. We had focused on those brain regions since they demonstrated significant changes in PPI at the group level analysis when deceptive actions were compared with honest ones. Therefore current analysis describes the organization of functional interactions revealed in the previous research. When deceptive actions were compared with honest ones, the left caudate nucleus was the most frequently observed participant in selected cliques.

This result is in accordance with the previous findings, demonstrating an increase in local activity and psychophysiological interactions of the left caudate nucleus with the left inferior frontal gyrus associated with the execution of deceptive actions [14,32]. These findings exhibit key role played by this area in brain maintenance of deception. It is also in line with our hypothesis that involvement of the left caudate nucleus in brain processing of deception reflects the process of activation of error detection mechanism, the first neurophysiological signature of which was demonstrated in [4]. In contrast to this, the right caudate nucleus was most frequently observed in the revealed cliques associated with honest actions in comparison with deceptive ones. It has to be noted that cliques, as well as cycles, associated with honest actions, were observed only at the very late stages of filtration. It highly likely renders these effects as trivial ones, or even driven by noise, by virtue of the fact that such cliques and cycles were selected at the relatively low values of PPI parameters. It is also supported by an absence of significant changes of the BOLD signals or PPI parameters for honest actions at the group level statistics in our previous studies [14–16]. Nevertheless, this observation for the first time points to possible differential roles of left and right caudate nuclei in processing of deceptive and honest actions, so it can be further investigated in future research. It is important to note that such involvement of the right caudate nucleus specifically in honest actions was not observed in previous studies devoted to the analysis of local BOLD changes or pairwise functional interactions. Therefore, the application of such quantitative topological parameters as the frequency of involvement of particular brain node in cliques formation provides valuable new information regarding the activity of brain network underlying the process of deception.

Figure 4 shows barcodes plot with the information about connective components and cycles for two weighted matrices. It can be seen that in case when deception prevails over honest actions, connected components merge significantly faster, all cycles appeared at the lower levels of filtration. This implies stronger connections between the ROIs in case of deception.

Loops Analysis. Analysis of persistent homology at a relatively early stage of filtration (high values of PPI parameters) revealed three stable 1-dimensional loops, which appeared in the middle of filtration and lasted for two levels. They were associated only with the execution of deceptive actions in comparison with honest ones. The first cycle included nodes in the following brain regions: left supplementary motor area (SMA), left middle frontal gyrus (Brodmann area (BA10), left and right inferior frontal gyri (IFG), left insula and temporal pole, right inferior temporal gyrus, left postcentral and right precentral gyri (BA3) and left lingual gyrus (BA18). This cycle largely reproduced the results of recent PPI studies [14,32], supporting the notion that functional role of those brain areas in deception execution is associated with executive control in general, as well as with action selection, action inhibition, reward expectation and decision making in particular. The second revealed cycle included left caudate nucleus, left superior frontal gyrus, left paracentral gyrus, right middle frontal gyrus, left and right postcentral gyri and left SMA. Taking into account the known role

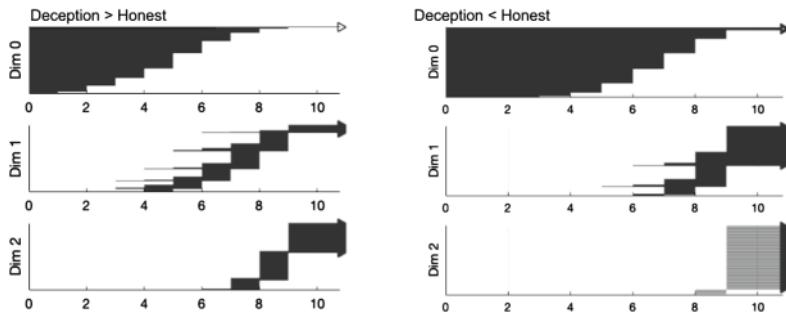


Fig. 4. Barcodes

of paracentral lobule in motor control [13], this functional loop with caudate nucleus and prefrontal brain regions can be considered a good candidate for the functional cycle supporting reaction to incompatible button-to-action mapping in case of execution of deceptive actions. This finding additionally corroborates the hypothesis of involvement of error detection mechanism in deception. And finally, it was earlier hypothesized that deception execution relies on an ability to perceive the mental state of another person, which is supported by ThOM brain network. In accordance with that, the third revealed cycle included the left superior medial frontal gyrus, dorsal anterior cingulate cortex, left caudate nucleus, left angular, right supramarginal gyri and the right insula. The majority of those regions is usually considered as nodes of the brain network responsible for representing cognitive and affective mental states of another person or self [2]. The revealed cycle supports the hypotheses stressing the impact of ThOM-related processes in deceptive behavior. To summarise, the central role of the left caudate nucleus in brain processing of deception was observed at the level of cliques. But the analysis of higher-order topological properties substantially specified functional organization of these cliques by revealing three cycles of interaction of brain regions exhibiting parallelism in functioning of possible brain mechanisms involved in deceptive behavior.

Acknowledgments. We gratefully acknowledge financial support of Saint-Petersburg State University (project ID 35544669), N.P. Bechtereva Institute of the Human Brain of the Russian Academy of Sciences and financial support of Institute of Information and Computational Technologies (Grant AR05134227, Kazakhstan).

References

1. Abe, N., Greene, J.D., Kiehl, K.A.: Reduced engagement of the anterior cingulate cortex in the dishonest decision-making of incarcerated psychopaths. *Soc. Cogn. Affect. Neurosci.* **13**(8), 797–807 (2018)
2. Abu-Akel, A., Shamay-Tsoory, S.: Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia* **49**(11), 2971–2984 (2011)

3. Bassett, D.S., Zurn, P., Gold, J.I.: On the nature and use of models in network neuroscience. *Nat. Rev. Neurosci.* **19**(9), 566–578 (2018). <https://www.nature.com/articles/s41583-018-0038-8>
4. Bechtereva, N., Gretchin, V.: Physiological foundations of mental activity. In: International review of neurobiology, vol. 11, pp. 329–352. Elsevier (1969)
5. Bubenik, P.: Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16**(1), 77–102 (2015)
6. Cisler, J.M., Bush, K., Steele, J.S.: A comparison of statistical methods for detecting context-modulated functional connectivity in fMRI. *Neuroimage* **84**, 1042–1052 (2014)
7. Debey, E., Ridderinkhof, R.K., De Houwer, J., De Schryver, M., Verschuere, B.: Suppressing the truth as a mechanism of deception: delta plots reveal the role of response inhibition in lying. *Conscious. Cogn.* **37**, 148–159 (2015)
8. Di, X., Reynolds, R.C., Biswal, B.B.: Imperfect (de)convolution may introduce spurious psychophysiological interactions and how to avoid it. *Hum. Brain Mapp.* **38**(4), 1723–1740 (2017)
9. Ding, X.P., Wu, S.J., Liu, J., Fu, G., Lee, K.: Functional neural networks of honesty and dishonesty in children: evidence from graph theory analysis. *Sci. Rep.* **7**(1), 12085 (2017)
10. Gao, J.F., Yang, Y., Huang, W.T., Lin, P., Ge, S., Zheng, H.M., Gu, L.Y., Zhou, H., Li, C.H., Rao, N.N.: Exploring time-and frequency-dependent functional connectivity and brain networks during deception with single-trial event-related potentials. *Sci. Rep.* **6**, 37065 (2016)
11. Gitelman, D.R., Penny, W.D., Ashburner, J., Friston, K.J.: Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. *Neuroimage* **19**(1), 200–207 (2003)
12. Giusti, C., Ghrist, R., Bassett, D.S.: Two's company, three (or more) is a simplex. *J. Comput. Neurosci.* **41**(1), 1–14 (2016). <https://doi.org/10.1007/s10827-016-0608-6>
13. Havel, P., Braun, B., Rau, S., Tonn, J.C., Fesl, G., Brückmann, H., Ilmberger, J.: Reproducibility of activation in four motor paradigms. *J. Neurol.* **253**(4), 471–476 (2006)
14. Kireev, M., Korotkov, A., Medvedeva, N., Masharipov, R., Medvedev, S.: Deceptive but not honest manipulative actions are associated with increased interaction between middle and inferior frontal gyri. *Front. Neurosci.* **11**, 482 (2017). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5583606/>
15. Kireev, M., Korotkov, A., Medvedeva, N., Medvedev, S.: Possible role of an error detection mechanism in brain processing of deception: PET-fMRI study. *Int. J. Psychophysiol.* **90**(3), 291–299 (2013)
16. Kireev, M., Medvedeva, N., Korotkov, A., Medvedev, S.: Functional interactions between the caudate nuclei and inferior frontal gyrus providing deliberate deception. *Hum. Physiol.* **41**(1), 22–26 (2015)
17. Knyazeva, I., Poyda, A., Orlov, V., Verkhlyutov, V., Makarenko, N., Kozlov, S., Velichkovsky, B., Ushakov, V.: Resting state dynamic functional connectivity: network topology analysis. *Biol. Inspired Cogn. Archit.* **23**, 43–53 (2018)
18. Lisofsky, N., Kazzer, P., Heekeren, H.R., Prehn, K.: Investigating socio-cognitive processes in deception: a quantitative meta-analysis of neuroimaging studies. *Neuropsychologia* **61**, 113–122 (2014)
19. Luo, Q., Ma, Y., Bhatt, M.A., Montague, P.R., Feng, J.: The functional architecture of the brain underlies strategic deception in impression management. *Front. Hum. Neurosci.* **11**, 513 (2017)

20. Mar, R.A.: The neural bases of social cognition and story comprehension. *Annu. Rev. Psychol.* **62**, 103–134 (2011)
21. McLaren, D.G., Ries, M.L., Xu, G., Johnson, S.C.: A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage* **61**(4), 1277–1286 (2012). <http://www.sciencedirect.com/science/article/pii/S1053811912003497>
22. Newman, M.E.J.: Analysis of weighted networks. *Phys. Rev. E* **70**, 056131 (2004). <https://doi.org/10.1103/PhysRevE.70.056131>
23. Oldfield, R.C.: The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**(1), 97–113 (1971)
24. Otter, N., Porter, M.A., Tillmann, U., Grindrod, P., Harrington, H.A.: A roadmap for the computation of persistent homology. *EPJ Data Sci.* **6**(1), 17 (2017)
25. Saxe, R.: Theory of mind (neural basis). *Encycl. Conscious.* **2**, 401–410 (2009)
26. Seitzman, B.A., Gratton, C., Marek, S., Raut, R.V., Dosenbach, N.U., Schlaggar, B.L., Petersen, S.E., Greene, D.J.: A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum. *bioRxiv* p. 450452 (2018)
27. Sizemore, A.E., Giusti, C., Kahn, A., Vettel, J.M., Betzel, R.F., Bassett, D.S.: Cliques and cavities in the human connectome. *J. Comput. Neurosci.* **44**(1), 115–145 (2018). <https://doi.org/10.1007/s10827-017-0672-6>
28. Sizemore, A.E., Phillips-Cremins, J.E., Ghrist, R., Bassett, D.S.: The importance of the whole: topological data analysis for the network neuroscientist. *Netw. Neurosci.* **3**(3), 656–673 (2018). https://doi.org/10.1162/netn_a.00073
29. Tausz, A., Vejdemo-Johansson, M., Adams, H.: JavaPlex: a research software package for persistent (co)homology. In: Hong, H., Yap, C. (eds.) *Proceedings of ICMS 2014. LNCS*, vol. 8592, pp. 129–136 (2014). <http://appliedtopology.github.io/javaplex/>
30. Volz, K.G., Vogeley, K., Tittgemeyer, M., von Cramon, D.Y., Sutter, M.: The neural basis of deception in strategic interactions. *Front. Behav. Neurosci.* **9**, 27 (2015)
31. Wang, Y., Ng, W.C., Ng, K.S., Yu, K., Wu, T., Li, X.: An electroencephalography network and connectivity analysis for deception in instructed lying tasks. *PLoS One* **10**(2), e0116522 (2015)
32. Yin, L., Weber, B.: I lie, why don't you: neural mechanisms of individual differences in self-serving lying. *Hum. Brain Mapp.* **40**(4), 1101–1113 (2019)
33. Lin, X., Fu, G., Sai, L., Chen, H., Yang, J., Wang, M., Liu, Q., Yang, G., Zhang, J., Zhang, J., et al.: Mapping the small-world properties of brain networks in deception with functional near-infrared spectroscopy. *Sci. Rep.* **6**, 25297 (2016)

Urban Networks and Mobility



Functional Community Detection in Power Grids

Xiaoliang Wang^{1,2} , Fei Xue¹ , Shaofeng Lu¹ , Lin Jiang² ,
and Qigang Wu^{1,2}

¹ Xi'an Jiaotong-Liverpool University, Suzhou, People's Republic of China
fei.xue@xjtlu.edu.cn

² University of Liverpool, Liverpool, UK

Abstract. Community detection algorithm is broadly applied in amount of studies to partition networks. But not all available methods are equally suitable for power grids. This paper proposes the concept of functional community structure based on functionality of the network. And a novel partitioning algorithm is presented by upgrading the Newman fast algorithm of community detection. The coupling strength is therefore proposed to replace conventional adjacency matrix to represent the relationship between nodes in networks. The electrical coupling strength (ECS) is defined to better reflect electrical characteristics between any two nodes in power grids. Furthermore, to consider the functionality of node type distribution, power supply strength (PSS) is proposed based on ECS only from generation nodes to load nodes to evaluate the impact of different node type distribution in the power supply. Moreover, modularity is redefined as power supply modularity based on PSS to evaluate the partitioning performance of power grids. Finally, considering the functionality of power grids. The Newman fast algorithm is upgraded based on power supply modularity.

Keywords: Complex network · Community detection · Functional community · Topological community · Newman fast algorithm

1 Introduction

Recently, as the power systems thrive, the scale of power grids increases constantly, meanwhile power network structures become ever more complex, many research efforts have been concentrated on structural analysis of the power grid. The complexity of the power system leads to the increasing risk of equipment outage and voltage collapse. An appropriate power grid partitioning could provide an effective and manageable distributed control strategy which can efficiently detect system faults and recover most electrical functionality when contingency events occur [1–4].

With the development of complex network theory, community detection methods could also be used to determine various types of community structure in different systems, such as biological systems, power grids, social networks and any type of communication networks that can be represented by a set of nodes and edges [5, 6]. Community detection can divide large-scale networks into some sub-networks whose

scales are smaller and easy to control. This network partitioning strategy can enhance the robustness of large-scale systems [7].

Some recent investigations have developed analysis and evaluation of power grid infrastructures based on complex network theory [8], such that the nodes of the network also represent power generation, transmission substation and load, whereas the branches correspond to transmission lines. Community detection approaches can be applied in power network as a complement for conventional power grid partitioning methods. In [9], a node similarity index for allocating each bus to community sharing maximum similarity has been proposed. However, in this model, the community detection method is mainly based on the pure topological features as undirected and unweighted models and does not take into consideration the function of community, and thus fails to fully reflect the electrical characteristics of power grids. Because of the increasing complexity of community detection problems, heuristic algorithms are implemented to obtain high quality solutions with a reduced runtime. In [10], two genetic algorithms (improved genetic algorithm and generational genetic algorithm) have been modified to solve community detection problems in power systems. Through analyzing the performance of two genetic algorithms, the results show that genetic algorithms are quick and powerful methods to detect communities in large scale power networks. Nevertheless, complex electrical properties also are neglected in this paper, which cannot fully reflect the functionality of power networks. Besides, in [10–12], modularity Q [13, 14] assists as a benchmark to evaluate the partitioning results. However, Q is not specially designed for power systems and does not appropriately reflect the electrical characteristics of power grids.

In conclusion, in most existing studies based on complex network theory, the definition of a community is all based on the density of line distribution among nodes which could be regarded as a **topological community structure**, and the results are questionable for power grid functionality. Furthermore, some power grid partitioning methods consider the functionality of power grids, but the features of community structure in network science are not well utilized. Therefore, this paper aims to develop a novel power network partitioning method, which combines community structural characteristics and functionality of power grids. A new concept which is called **functional community structure** is therefore proposed to develop an optimum partitioning algorithm that can identify community structure from the perspective of network functionality.

The main contributions of this paper are:

- (1) Functional community structure is proposed in comparison with topological community structure.
- (2) Considering the node type distribution, ECS will be extended at PSS to evaluate the impact of different node type distribution during power supply.
- (3) Modularity is redefined as electrical modularity Q_E based on ECS. And the power supply modularity Q_S is further defined to evaluate the network performance based on PSS and Q_E .
- (4) Newman fast algorithm is upgraded to detect functional community structures based on electrical modularity and power supply modularity for power grid partitioning.

2 Functional Community in Power Grids

2.1 Electrical Coupling Strength (ECS)

The essential function of networks is to transmit physical or informational quantities between nodes. In conventional network model, adjacency matrix A_{ij} is usually utilized to represent the connection of network topology [13, 14], which is as follow:

$$A_{ij} = \begin{cases} 1, & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

In a weighted network, A_{ij} is equal instead to the weight of corresponding edge [15]. However, adjacency matrix which does not comprehensively reflect the essential function of networks, only indicates the connection between nodes and neglects the functionality of transmission among buses. Therefore, we establish an extended weighted network model to indicate the transmission capability between any two nodes. Meanwhile, we define this functionality as coupling strength between two nodes in the network. The exact definition of coupling strength may be altered for various types of networks which possess diverse functions. Furthermore, the adjacency matrix is updated to an extended adjacency matrix in which the element corresponding to any two nodes is the coupling strength between them. The extended adjacency matrix is written as follows:

$$\text{Ex}A_{ij} = \text{coupling strength of connection from } i \text{ to } j \quad (2)$$

The functionality of transmission between any buses can be reflected by coupling strength no matter they are directly or indirectly connected (**any non-diagonal element is non-zero**). As a kind of complex network, the coupling strength in power networks is named as electrical coupling strength (ECS), which can reflect the electrical characteristics of power grids during power transmission. To reasonably represent the transmission capability between buses, ECS is determined by transmission capacity [16] and equivalent impedance [17]. The ECS between buses i and j is:

$$E_{ij} = \frac{C_{ij}}{Z_{ij}^e} \quad (3)$$

Where C_{ij} is the power transmission capacity when power is injected at bus i and withdrawn at load bus j [16]:

$$C_{ij} = \min_{l \in L} \left(\frac{P_l^{\max}}{|f_l^{ij}|} \right) \quad (4)$$

Where f_l^{ij} is the power transfer distribution factor (PTDF) on line l when a unit of power injected at bus i and withdrawn from bus j ; P_l^{\max} is the power flow limit

regarding the transmission line l . Z_{ij}^e is the equivalent impedance between bus i and bus j , the definition is [17]:

$$Z_{ij}^e = z_{ii} - 2z_{ij} + z_{jj} \quad (5)$$

Where z_{ii} , z_{ij} and z_{jj} are corresponding elements in the impedance matrix of the network. The scale of C_{ij} and Z_{ij}^e may be quite different, which may result in the value of ECS become only sensitive to C_{ij} or Z_{ij}^e . Therefore, these two quantities are normalized based on their average value as follow:

$$\bar{C}_{ij} = \frac{C_{ij}}{\bar{C}} ; \bar{Y}_{ij} = \frac{Y_{ij}}{\bar{Y}} = \frac{1}{\bar{Z}_{ij}^e} \quad (6)$$

Where Y_{ij} is the reciprocal of Z_{ij}^e . \bar{C} and \bar{Y} are the average values of transmission capacity and equivalent admittance. Then, ECS is further upgraded to adjust the influence of these two components by changing their proportions as Eq. 7. Where a and b are proportions coefficients whose sum is equal to 1.

$$\bar{E}_{ij} = |\alpha \bar{Y}_{ij} + j\beta \bar{C}_{ij}| \quad (7)$$

2.2 Power Supply Strength

Reference [18] has concluded that the generation-load balance of individual partitions can be considered as a major concern in network partitioning for power system restoration. In [19], the researchers proposed that a uniform and dispersed distribution of loads can reduce vulnerability of power grids. However, the corresponding distribution of generations of nodes and loads nodes was not considered in community detection in network science. In [20], the researchers proposed a method of allocation of generation nodes based on identified network topology and load nodes. Therefore, node type distribution is inseparable for power network partitioning.

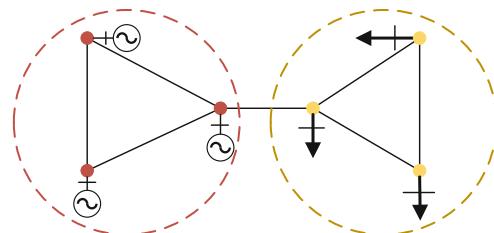


Fig. 1. The impact of node type distribution on power grids partitioning

Figure 1 is an example to explain the effect of node type distribution on power grid partitioning. If structural characteristics of the network are only considered, this network may be divided into two communities as Fig. 1. In this case, all generator buses are in the same community, and all load buses are in the other community. It is obvious that this partitioning result is not meaningful for the function of power grids which are to transmit power from the generation bus to the load bus.

Considering the functionality of power supply networks, ECS is redefined as power supply strength (PSS) to indicate the impact of different node type distribution during power supply. The power supply strength (PSS, S_{gd}) can be expressed as:

$$S_{gd} = \frac{T_g^d}{Z_g^d} \quad (8)$$

Where T_g^d represents the valid capacity for power transmission from generation bus g to load bus d . T_g^d is equal to the minimum value of the capacity at generation bus, load bus and the capacity among generation bus and load bus:

$$T_g^d = \min[T_g, T_{gd}, T_d] \quad (9)$$

Note that T_g is the capacity at generation bus g ; T_d is the capacity at load bus d ; T_{gd} is the power transmission capacity depending on network connection:

$$T_{gd} = \min_{l \in L} \left(\frac{P_l^{\max}}{|f_l^{gd}|} \right) \quad (10)$$

In Eq. 8, Z_g^d is equivalent impedance from bus g to bus d . The definition is similar with Eq. 5 ($Z_g^d = z_{gg} - 2z_{gd} + z_{dd}$). Then, T_g^d and Z_g^d are normalized as:

$$\overline{T_g^d} = \frac{T_g^d}{\overline{T}}; \overline{Y_g^d} = \frac{Y_g^d}{\overline{Y}} = \frac{\frac{1}{Z_g^d}}{\overline{Y}} \quad (11)$$

Afterwards, two coefficients d and 1 are utilized to change the proportions of transmission capacity and equivalent admittance from generations nodes to loads nodes. PSS can be expressed as:

$$\overline{S}_{gd} = \left| \delta \overline{Y}_l + j \lambda \overline{T}_g^d \right| \quad (12)$$

Based on the above discussion, PSS is defined to describe the power supply ability between generations buses and loads buses based on the structural and functional characteristic of power network. The PSS matrix is expressed as:

$$S_{gd} = \begin{cases} \overline{S}_{gd}, & \text{Connection from generation bus } g \text{ to load bus } d \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

Different from conventional topological community detection which is based on the high internal density of line or connection distribution, functional community structure in power grids is based on the superior internal density of PSS distribution. Note that in Eq. 13, only elements between generation nodes and load nodes are non-zero. All other elements are zero. This is different from the conventional adjacency matrix and different from the ECS matrix.

3 Power Supply Modularity

In conventional community detection methods, Newman proposed modularity Q to evaluate partitioning performance. The higher the modularity, the better the partition. The modularity Q is defined as [13, 14]:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad (14)$$

Where, A_{ij} is element in adjacency matrix. m is equal to $\frac{1}{2} \sum_{ij} A_{ij}$ which represents the number of edges in network. k_i (k_j) is the degree of vertex i (j). c_i (c_j) is the community which vertex i (j) belongs to.

The δ -function is 1 if nodes i and j are in same community, otherwise it is 0. The conventional algorithm considers nodes with more intensive connections as community; but we consider nodes with more intensive PSS as community. To explain this point, an example is illustrated in Fig. 2.

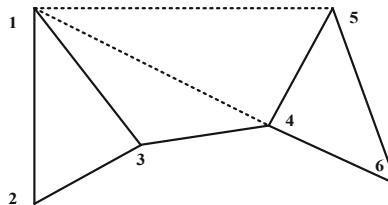


Fig. 2. Functional community detection via Newman fast algorithm.

In Fig. 2, solid lines are the direction connection; dashed lines represent the non-direct connection. There is a direct connection between bus 1 and bus 2, which mean that the corresponding elements of the adjacency matrix among these two buses are not zero. There is no direct connection among bus 1 and bus 5, and the element of adjacency A_{15} is zero. However, for the power network, some physical quantity can still be transmitted between two nodes even if there is no direct connection between them. The ECS between bus 1 and bus 5 is E_{15} that is not zero. Moreover, the magnitude of E_{15} possibly will be greater than E_{12} .

The purpose of this study is to detect functional community structure according to the density of coupling strength. Therefore, we proposed electrical modularity Q_E wherein ECS and Newman fast algorithm are integrated, to detect functional community structure in power network. The electrical modularity can be expressed as:

$$Q_E = \sum_{ij} \left[\frac{E_{ij}}{2M} - \frac{E_i}{2M} \frac{E_j}{2M} \right] \delta(c_i, c_j) \quad (15)$$

Where M is total electrical coupling strength of the whole power grid system. Because the scale of E_{ij} and E_{ji} is equal, the magnitude of M is half of $\sum_{ij} E_{ij}$. E_i (E_j) is ECS degree at bus i (j), which is equal to $\sum_j E_{ij} \left(\sum_i E_{ij} \right)$.

As mentioned in Sect. 2, considering the features of power supply network, partitioning should also consider node type distribution. Then the electrical modularity is further extended as power supply modularity Q_S by PSS. We exploit power supply modularity Q_S for clustering power supply network with Newman fast algorithm. Power supply modularity Q_S is defined as follows:

$$Q_S = \sum_{gd} \left[\frac{S_{gd}}{2N} - \frac{S_g}{2N} \frac{S_d}{2N} \right] \delta(c_i, c_j) \quad (16)$$

Where S_{gd} is power supply strength. N is sum of PSS in network, which is defined as $\frac{1}{2} \sum_{gd} S_{gd}$. S_g (S_d) is defined as PSS degree of generation (load) bus:

$$S_g = \sum_d S_{gd}; S_d = \sum_g S_{gd} \quad (17)$$

4 Power Network Partitioning Algorithm

Considering the electrical characteristics and functionality of power grids, the Newman fast algorithm is redesigned based on power supply modularity, which can be described as fellows. N is the number of buses in power grids.

Algorithm 1 Improved Newman fast algorithm

```

1: initialize power network with N communities;
2: calculate the power supply modularity  $Q_S$ ;
3: while the number of communities is not 1 do
4:   if there is direct connection between two communities
5:     then
6:       group two communities randomly;
7:     else
8:       the communities cannot be grouped together;
9:     end if
10:    calculate the increments of electrical modularity  $\Delta Q_S$ ;
11:    select the partitioning with maximum  $\Delta Q_S$ ;
12:    recalculate  $Q_S$  according to the result of partitioning;
13:    conserve the number of communities (The maximum number of mergers
is N-1);
14: end while

```

5 Case Study

5.1 Experiments on Different Test Systems

To demonstrate the effectiveness of the proposed partitioning algorithm, we perform experiments on the IEEE-118 system, IEEE-300 system and one Italian power network. The relationship between power supply modularity Q_S and the number of communities is shown in Fig. 3 for the IEEE-118 system. The best partitioning result is three communities when power supply modularity Q_S is 0.0752. Table 1 shows the specific results of each communities, furthermore, the topology of partitioning is shown in Fig. 4, where the different colors indicate different communities.

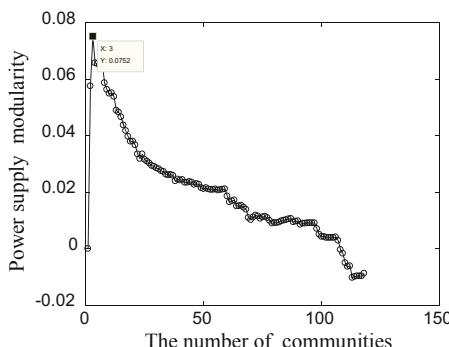
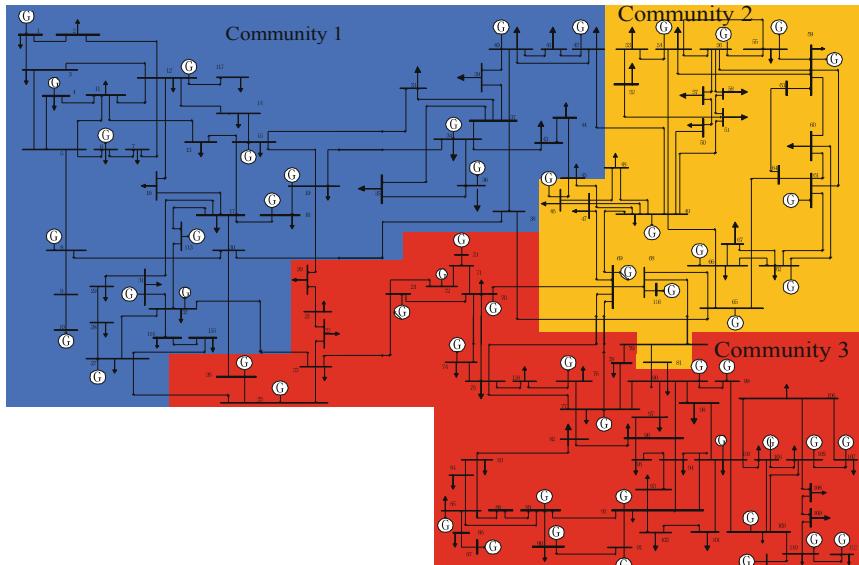
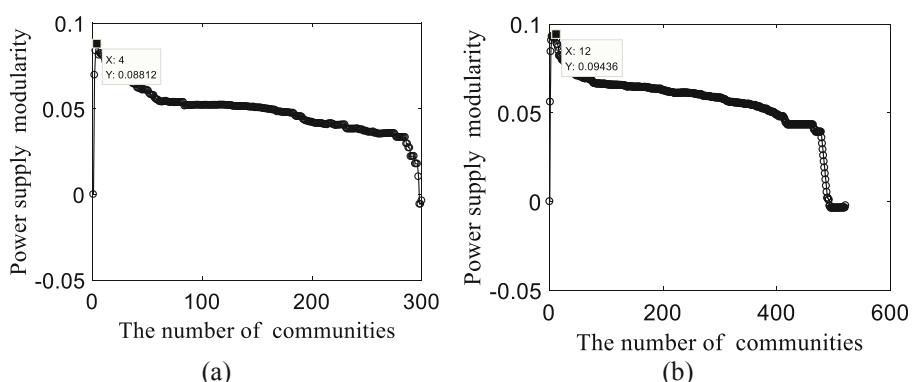


Fig. 3. The Q_S for different number of communities in IEEE-118 system.

Table 1 IEEE 118-bus system partitioning results

Community	Bus number
1	1–19, 27–45, 113–115, 117
2	46–69, 81, 116
3	20–26, 70–80, 82–112, 118

**Fig. 4.** IEEE-118 system partitioning results based on power supply modularity.**Fig. 5.** The Q_s for possible partitioning results in (a) IEEE-300 bus system; (b) Italian power network.

Furthermore, the proposed partitioning algorithm is applied to IEEE-300 and Italian power system to prove the feasibility in large-scale power networks. Figure 5 shows different Q_S for possible partitioning results in IEEE 300-bus system and Italian power network. The Q_S reaches to maximum 0.08812 when IEEE 300-bus system is partitioned into four communities. The Italian power network consists of 521 buses, 159 generators and 679 branches. Figure 5 shows that the optimal partitioning result for this system is 12 communities when Q_S is equal to 0.09436.

5.2 Comparison Against Previous Power Grid Partitioning Methods

To illustrate the feasibility of the proposed method, we compared the improved Newman fast algorithm against previously published methods. In [9], Chen et al. defined the similarity which is utilized to reflect the nodes' relationship in a community. The more a node is similar with nodes in a community, the greater its probability to join the community. However, in this way, complex electrical properties and functionality of power network are neglected, and the power grid is simplified as an undirected and unweighted network. In [9], the partitioning algorithm is applied to the IEEE 118-bus system, so we use this system again here to make comparisons in terms of the power supply modularity.

Furthermore, in [9], the conventional modularity is viewed as an index which can evaluate the performance of partitioning. Nonetheless, the conventional modularity is not designed specifically for power systems, which cannot be fittingly manifest the electrical characteristics and functionality of power networks. The IEEE 118-bus system is partitioned into 8 communities in [9]. The definitions of modularity between the two methods we tested here are entirely different. Therefore, directly comparing their values is not meaningful. The power supply modularity better indicates network functionality, so we compared the two partitioning results instead by power supply modularity Q_S .

The Q_S of partitioning results in [9] is equal to 0.0662. As Fig. 3 shows, the Q_S is equal to 0.0752 with 3 communities. It is obvious to find that power supply modularity of our partitioning result is larger than the other one. Moreover, to ensure that each community has at least one generator, in [9], the researchers pre-divide the grid into 10 initial communities with one generator in each community and assign each load node to its nearest generator which takes no account of electrical characteristics. However, Q_S is based on power supply strength which can indicate the interaction between generation bus and load bus.

6 Conclusion

In conclusion, in this paper, we propose a power grid partitioning method which works by detecting functional community structures from the complex network perspective. Considering the electrical characteristic of power supply networks, PSS is determined to replace conventional adjacency matrix to represent the relationship between nodes in power grids. Further, based on PSS, the modularity is redefined as power supply modularity which is implemented to evaluate the partitioning performance of power

grids. We found that PSS reveals relations among different nodes from a new perspective. Simulation results obtained using the IEEE 118-bus system, IEEE 300-bus system and Italian power network verify the applicability of the proposed strategy in power network partitioning. Our partitioning method performed well by comparison against other methods. Our results show that power supply modularity Q_S can better reflect the functionality of power grids. In addition, the concept of functional community structures can provide a fresh perspective for complex networks, which can be expanded into different network systems. The functionality of a network is representing one of the important features of the network. Different networks have other functionalities. Therefore, functional community detection is also promising in other network systems.

References

1. Jia, Y.W., Xu, Z.: A direct solution to biobjective partitioning problem in electric power networks. *IEEE Trans. Power Syst.* **32**(3), 2481–2483 (2017)
2. Golshani, A., Sun, W., Sun, K.: Advanced power system partitioning method for fast and reliable restoration: toward a self-healing power grid. *IET Gener. Transm. Distrib.* **12**(1), 42–52 (2018)
3. Areffifar, S.A., Mohamed, Y.A.R.I., EL-Fouly, T.H.M.: Comprehensive operational planning framework for self-healing control actions in smart distribution grids. *IEEE Trans. Power Syst.* **28**(4), 4192–4200 (2013)
4. Li, J., Liu, C.C., Schneider, K.P.: Controlled partitioning of a power network considering real and reactive power balance. *IEEE Trans. Smart Grid* **1**(3), 261–269 (2010)
5. Fortunato, S.: Community detection in graphs. *Phys. Rep. Rev. Sect. Phys. Lett.* **486**(3–5), 75–174 (2010)
6. Zarandi, F.D., Rafsanjani, M.K.: Community detection in complex networks using structural similarity. *Phys. Stat. Mech. Appl.* **503**, 882–891 (2018)
7. Pahwa, S., Youssef, M., Schumm, P., Scoglio, C., Schulz, N.: Optimal intentional islanding to enhance the robustness of power grid networks. *Phys. Stat. Mech. Appl.* **392**(17), 3741–3754 (2013)
8. Pagani, G.A., Aiello, V.: The power grid as a complex network: a survey. *Physica A* **392**(11), 2688–2700 (2013)
9. Chen, Z.Q., Xie, Z., Zhang, Q.: Community detection based on local topological information and its application in power grid. *Neurocomputing* **170**, 384–392 (2015)
10. Guerrero, M., Montoya, F.G., Banos, R., Alcayde, A., Gil, C.: Community detection in national-scale high voltage transmission networks using genetic algorithms. *Adv. Eng. Inf.* **38**, 232–241 (2018)
11. Sanchez-Garcia, R.J., Fennelly, M., Norris, S., Wright, N., Niblo, G., Brodzki, J., Bialek, J. W.: Hierarchical spectral clustering of power grids. *IEEE Trans. Power Syst.* **29**(5), 2229–2237 (2014)
12. Guo, J.Y., Hug, G., Tonguz, O.K.: Intelligent partitioning in distributed optimization of electric power systems. *IEEE Trans. Smart Grid* **7**(3), 1249–1258 (2016)
13. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69**(6), 066133 (2004)
14. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. *Phys. Rev. E* **70**(6), 066111 (2004)

15. Newman, M.E.J.: Analysis of weighted networks. *Phys. Rev. E* **70**(5), 056131 (2004)
16. Bompard, E., Napoli, R., Xue, F.: Analysis of structural vulnerabilities in power transmission grids. *Int. J. Crit. Infrast. Prot.* **2**(1–2), 5–12 (2009)
17. Arianos, S., Bompard, E., Carbone, A., Xue, F.: Power grid vulnerability: a complex network approach. *Chaos* **19**(1), 013119 (2009)
18. Ganganath, N., Wang, J.V., Xu, X.Z., Cheng, C.T., Tse, C.K.: Agglomerative clustering-based network partitioning for parallel power system restoration. *IEEE Trans. Industr. Inf.* **14**(8), 3325–3333 (2018)
19. Kim, D.H., Eisenberg, D.A., Chun, Y.H., Park, J.: Network topology and resilience analysis of South Korean power grid. *Physica A* **465**, 13–24 (2017)
20. Guo, W.Z., Wang, H., Wu, Z.P.: Robustness analysis of complex networks with power decentralization strategy via flow-sensitive centrality against cascading failures. *Physica A* **494**, 186–199 (2018)



Comparing Traditional Methods of Complex Networks Construction in a Wind Farm Production Analysis Problem

Sara Cornejo-Bueno¹, Mihaela Ioana Chidean¹, Antonio J. Caamaño¹,
Luís Prieto², and Sancho Salcedo-Sanz^{3(✉)}

¹ Department of Signal Processing and Communications,
Universidad Rey Juan Carlos, Fuenlabrada, Madrid, Spain

² Iberdrola, Bilbao, Madrid, Spain

³ Department of Signal Processing and Communications,
Universidad de Alcalá, Alcalá de Henares, Madrid, Spain
sancho.salcedo@uah.es

Abstract. This work presents a comparison between two methods for complex networks construction (cross-correlation and Mutual Information based), in the assessment of wind speed prediction efficiency at different wind farms in Spain. The approach is accomplished at mesoscale, for wind speed prediction data provided by the Weather Research and Forecasting (WRF) numerical model versus the actual wind speed measurements in the wind farms. Some important differences are found in the complex networks obtained, and the corresponding global measures from them, such as the *betweenness* and *closeness* centrality. We have found out that the mutual information method better captures nonlinear relationships of the problem, obtaining complex networks with fewer spurious links than the cross-correlation based method.

Keywords: Climate networks construction · Wind farms · Complex networks construction · Wind power prediction

1 Introduction

Many natural and social problems can be described and solved in terms of complex networks (CN) [1]. The application areas of CN analysis are increasingly diverse: from studies related to the expansion of epidemics in a certain region [2], to investigations in criminology [3], among many others. In recent years, there has been an increase of these contributions in the area of climatology and climate studies [4–6]. The application of CN theory to the study of the processes that take place in the climate system is englobed in the field of Climate Networks. This novel approach to the study of climate allows identifying climate interdependencies and relationships [7], for example, it has been employed in different

attempts to refine the prediction of complex variability modes of El Niño Southern Oscillation [6,8]. Other authors have applied these theories to the study of spatial and temporal variability of rainfall and extreme events of precipitation [9,10].

Nodes in a climate network are defined by grid points from a set of underlying global climatic data [11]. Due to the continuity of the physical fields, the neighboring grid points should be dynamically correlated. It can be thus assumed that climate is represented by a grid of oscillators where each one represents a dynamical system changing in a complex way [12]. Traditionally all these networks have been constructed based on correlation functions (cross-correlation based method), but recently other clustering algorithms have been employed in order to avoid the problem of spurious links appearance [13], a common issue in CNs obtained from linear autocorrelation methods [14].

In this paper we present a comparison between two different methods for CN construction in a problem of wind speed prediction efficiency in wind farms. First, we consider the most classical method for CN construction, based on the use of linear cross-correlation. On the other hand, we compare it with an alternative method which employs the Mutual Information between the times series in different nodes to construct the CN. We will show that the Mutual Information method is able to capture linear and nonlinear relationships between time series [15] better than the one based on cross-correlation. The comparison is established in terms of the wind power perspective of many wind farms located in the Iberian Peninsula. The proposed approach gives us the opportunity of assessing the dependence between the geographical location of the wind farm and the accuracy of the wind prediction numerical methods. Also, it allows evaluating the importance of the geographical areas where the information is regulated and transmitted.

The remainder of the paper has been structured as follows: next section describes both CN construction methods considered in this paper. Section 3 briefly describes the statistical measures considered to compare the CN construction approaches. Section 4 describes the problem tackled and the available data from real wind farms in Spain. Section 5 discusses the results obtained in this paper, focussing on the differences between the CNs obtained with the different construction methods considered. Finally, Sect. 6 closes the paper by giving some final remarks and conclusions on the results obtained in this work.

2 Network Construction Methods Compared

In this section we describe the CN construction methods discussed in this study, specifically the cross-correlation and Mutual Information-based CN construction approaches.

2.1 Cross-Correlation CN

We describe here the traditional cross-correlation method for CN construction, based on the estimation of linear relationships in the network, obtained by computing the cross-correlation function (γ) for each pair of nodes (i, j):

$$\gamma_{ij} = \frac{1}{n} \sum_{t=1}^n (x_t^i - \bar{x}^i)(x_t^j - \bar{x}^j) \quad (1)$$

where x_t^k is the wind speed prediction error at time t in node k , and n is the length of the time series considered (4300 h in our case).

In the cross-correlation function method, a *link strength* is then established as:

$$S_{ij} = \frac{\gamma_{max} - \bar{\gamma}}{\sigma_\gamma} \quad (2)$$

where $\bar{\gamma}$ and σ_γ stand for the mean and standard deviation of γ . As we explain later, it will be used in the getting of links instead of the cross-correlation value.

2.2 Mutual Information-Based CN

The Mutual Information CN construction approach uses the calculation of non-linear statistical interdependencies from mutual information (M_{ij}), another measure widely used in different problems of Science and Engineering [7, 16]. It can be interpreted as the amount of information in excess and generated by the wrong assumption that the time series that are comparing are independent [5]. The Mutual Information can be expressed by

$$M_{ij} = \sum_{\mu\nu} p_{ij}(\mu, \nu) \log \frac{p_{ij}(\mu, \nu)}{p_i(\mu) \cdot p_j(\nu)} \quad (3)$$

where $p_i(\mu)$ is the probability density function (PDF) of the time series for node i , $p_j(\nu)$ the PDF relative to the time series of node j and, $p_{ij}(\mu, \nu)$ is the joint PDF of a pair of time series relative to nodes i, j . Those probability densities are calculated using an histogram approach as in [5]. M_{ij} is a symmetric measure ($M_{ij} = M_{ji}$) of the degree of statistical interdependence of the time series in nodes i and j ; if they are independent: $p_{ij}(\mu, \nu) = p_i(\mu) \cdot p_j(\nu)$ and thus $M_{ij} = 0$ [7]. Logarithms to base 2 are used, so the unit of measurement of Mutual Information is the bit. The reader is suggested to read reference [5] for more details.

To obtain the links between the nodes for both construction methods, some restrictions are imposed. A spatial threshold of 300 km has been established and associated with each wind farm, so that we can find a representative number of wind farms to correlate within this *radius of influence*. Otherwise, the probability of appearance of spurious links would be greater since the local correlations between physical fields, such as wind speed, usually decay within a length scale [5]. For this reason it makes no sense to analyze correlations between wind farms

located too far away. Also, a statistical threshold is needed. It is obtained by the sum of the mean of the link strength/mutual information (when appropriate) plus u times the standard deviation of the link strength/mutual information respectively,

$$U = \bar{S}_{ij} + u \cdot \sigma_{S_{ij}} \quad (4)$$

We choose $u = 3$ because from that threshold onwards all the links in the network are statistically significant at 99%. Thus, only those pair of nodes whose S_{ij}/M_{ij} (in each case) exceeds both thresholds (the spatial one and the statistical), will be considered as significantly linked. Then, all the emergent links are collected in the adjacency matrix.

3 Network Measures

From the adjacency matrix of the obtained networks, we can calculate all the measures involved in this study, such as the degree distribution (P_k), the betweenness and closeness centrality etc., among others possible ones. The closeness and betweenness centrality are global measures. Closeness centrality (CC_v) is calculated as the inverse sum of the distance from a node to all other nodes in the graph. It is computed as follows:

$$CC_v(i) = \left(\frac{A_i}{N - 1} \right)^2 \frac{1}{C_i} \quad (5)$$

where A_i is the number of reachable nodes from node i (not counting i), N is the number of nodes in the network, and C_i is the sum of distances from node i to all reachable nodes. CC_v is large if the node v is topologically close to the rest of the network.

On the other hand, the betweenness centrality (BC_v) takes into account the number of shortest paths that go through a node, as a mediator for the information transport. It can be expressed by

$$BC_v = \sum_{s,t \neq v} \frac{n_{st}(v)}{N_{st}} \quad (6)$$

where $n_{st}(v)$ is the number of shortest paths from s to t that pass through node v , and N_{st} is the total number of shortest paths from s to t .

Another measure that is calculated in this study is the clustering coefficient C_i , that quantifies the tendency of the nodes to cluster:

$$C_i = \frac{2E_i}{k_i(k_i - 1)} \quad (7)$$

That is, the clustering coefficient of a node i in the network with k_i edges which connects with k_i other nodes. E_i is the number of edges that actually exist between these k_i nodes and $k_i(k_i - 1)/2$ gives the total number of possible edges. The average of the individual C_i 's gives the clustering coefficient of the whole network (C) [1].

4 Problem Tackled: CN in Wind Farm Production Analysis

We consider 171 wind farms in Spain and their wind speed prediction error, i.e. the difference between wind speed prediction data calculated with a numerical method (Weather Research and Forecasting, WRF) and the real wind speed measured in each wind farm. The WRF is a next-generation mesoscale numerical weather prediction system, designed for both atmospheric research and operational forecasting applications [17]. This numerical model has been used in energy applications, e.g. to compare the surface wind simulation and wind energy estimations when forced by different initial and boundary conditions [18]. For the database, we consider a 2 h prediction time-horizon. Around six months (4300 h) are encompassed by the time series of wind speed prediction errors for the 171 wind farms. In Fig. 1, the geographical distribution of the nodes (wind farms considered) is shown. We use the wind speed prediction error for our analysis since the power of the wind passing through the cross-sectional area of the wind turbine varies with the cube (the third power) of the wind speed v [19]. Thus we can also evaluate the results in terms of wind power error.

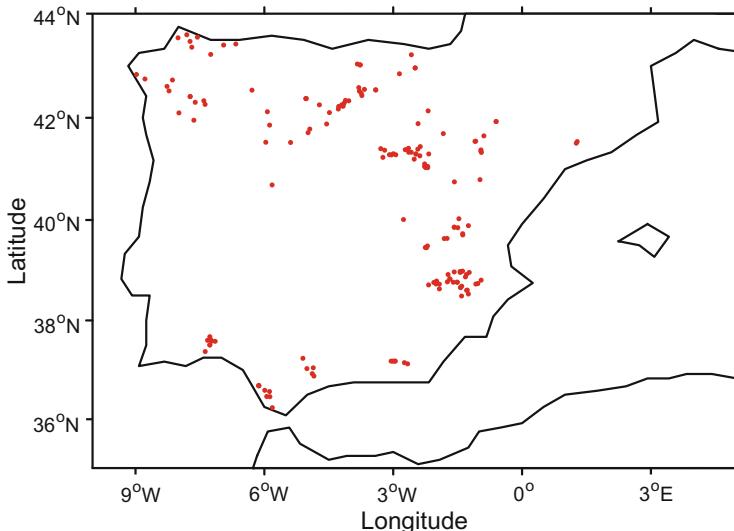


Fig. 1. Wind farms in the Iberian Peninsula considered for this study.

5 Experiments and Results

In this section we analyze the results obtained with the cross-correlation approach (S_{ij}) versus the obtained with mutual information (M_{ij}), both methods have been described in Sect. 2. The CN construction is carried out for the 171

wind farms and their wind speed prediction error. The network measures calculated are those described in Sect. 3. The imposed threshold U is operated with $u=3$ for both construction methods in order to get links between nodes with statistical significance (see Sect. 2.2). Otherwise, a CN with a small threshold is not expected to contain meaningful information for the analysis since it is more likely to generate spurious links.

In a first mesoscale comparative analysis, we can observe that the network obtained by S_{ij} contains a larger amount of links than the one obtained by M_{ij} (Fig. 2). This fact has been also revealed by the local and global clustering coefficient. The local clustering coefficient distributions for each network is represented in the inset of each figure in Fig. 2.

It can be seen that the local clustering coefficient distribution is more flattened in the S_{ij} CN than in the M_{ij} , what is related to the lower number of connections in the second case. The same effect can be detected by means of the global clustering coefficient: the higher the edge density, the larger the global clustering coefficient, as can be seen in Table 1.

Table 1. Edge densities and global clustering coefficients at the imposed thresholds.

	$\rho_{S_{ij}}$	$\rho_{M_{ij}}$
	0.0260	0.0194
C	0.7153	0.6457

Other differences can be evaluated in terms of the degree distribution (P_k) of the networks considered. Figure 3 shows that the M_{ij} CN is more likely to have low-grade nodes as opposed to the S_{ij} network. This could support the fact that CN from M_{ij} construction generate less false positives than with the S_{ij} method. Also, it seems that the behaviour of the nodes' individual connectivity slightly changes from one method to another.

There are many useful measures that reveal important geographical zones as information regulators [5]. Let us continue the study with these other types of global measures, such as betweenness and closeness centrality. Note that the approach remains at mesoscale, since the spatial threshold is still 300 km. Closeness centrality (CC_v) differs substantially between both methods i.e., while we find the wind farm number 1 in Fig. 4(a) the most central node in the S_{ij} construction, we have two nodes in the M_{ij} CN: 2 and 3 as the most central nodes (Fig. 4(b)). The latter are located at Burgos (Castilla y Leon), whereas the first one is at Albacete (Castilla-La Mancha). This denotes a spatial change from one network to the other. Not only the node degree changes from the S_{ij} CN to the M_{ij} CN, but also the structure in terms of the transport of information, obtaining different *focus* nodes in both cases. The same situation can be seen in the betweenness centrality measure (BC_v). In S_{ij} CN, we obtained the node 4 (Fig. 5(a)) as the most regulator node, located again at Albacete. On the other hand, we obtained the node 5, located at Murcia, as the most regulator one in

the M_{ij} construction. These differences found are consistent, since BC_v heavily depends on the local existence or non-existence of a small number of edges in the network. Also, the PDF follows a flat tailed behaviour in the case of BC_v , and, for CC_v , we have a normal distribution in both CN, as it happened in previous studies [5, 20].

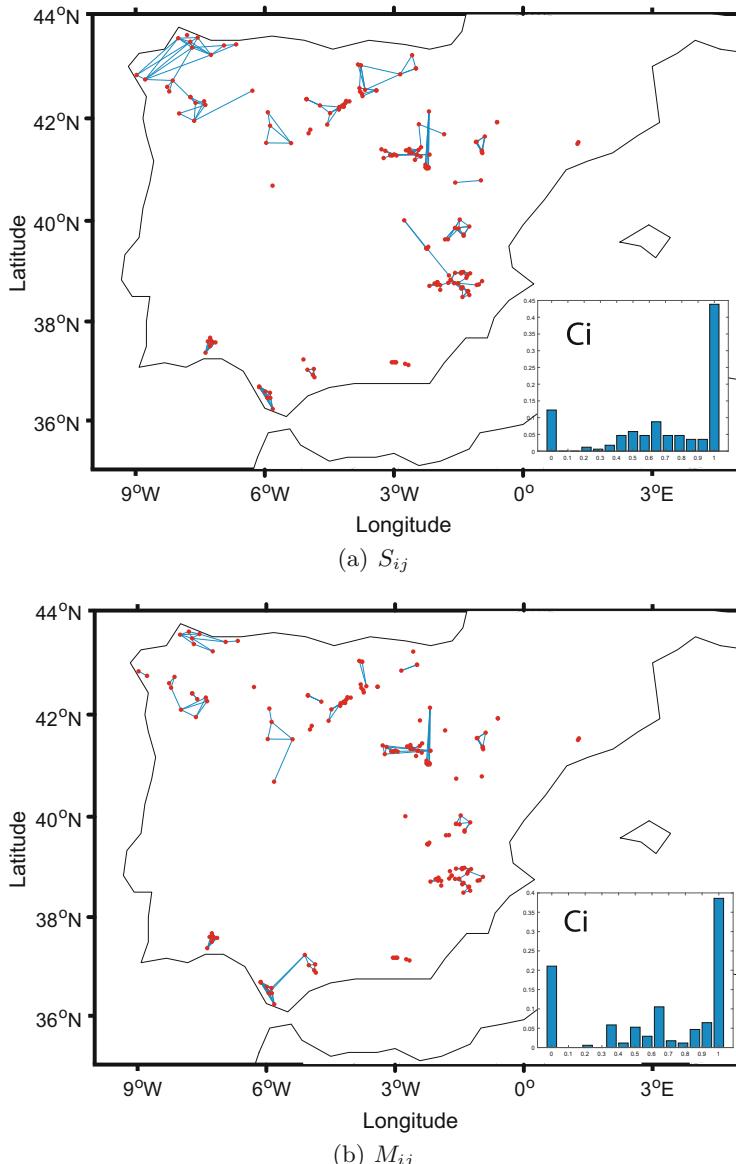


Fig. 2. CNs obtained with cross-correlation function and its corresponding S_{ij} (a), and from mutual information (b).

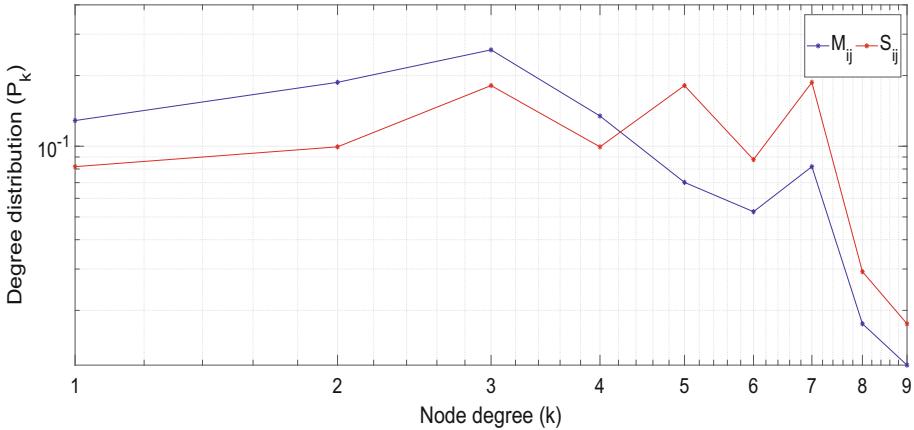


Fig. 3. Degree distribution (P_k) for the S_{ij} and M_{ij} constructions at the imposed thresholds.

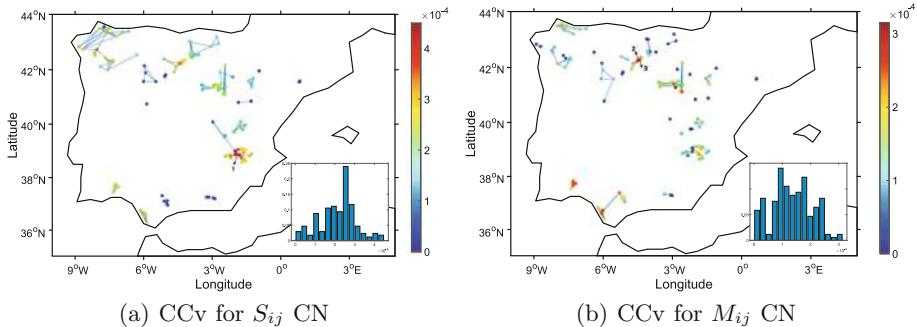


Fig. 4. Closeness centrality (CC_v) for both type of networks and its PDF (in the inset).

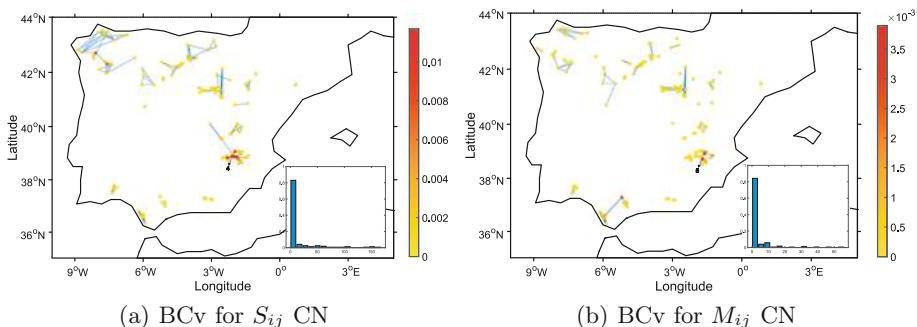


Fig. 5. Betweenness centrality (BC_v) for both type of networks and its PDF (in the inset).

6 Conclusions

In this paper we have carried out an evaluation of the differences between two traditional complex network (CN) construction methods in a problem of wind speed prediction efficiency at different wind farms in Spain. Specifically the construction of CN using the linear cross-correlation method, and the nonlinear Mutual Information measure is analyzed and compared. For both CN construction methods, we have imposed statistical and spatial thresholds so that the networks obtained are statistically significant. We find similar behaviours in global measures as the *closeness* and *betweenness* centrality with respect to other previous works in the literature. These measures allow identifying areas of network importance where the information is transported inside the network. The inclusion of topological mesoscale bounds in the CN construction is consistent with the fact that local correlations between physical fields usually decay within a length scale. Also it reduces the number of spurious links that can appear between wind farms located far away. The application of CN analysis to climate and renewable energy can give us new insights regarding the performance of infrastructures such as wind farms.

Acknowledgments. This research has been partially supported by the Ministerio de Economía y Competitividad of Spain (Grant Ref. TIN2017-85887-C2-2-P).

References

- Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002)
- Karrer, B., Newman, M.E.: Competing epidemics on complex networks. *Phys. Rev. E* **84**(3), 036106 (2011)
- Wang, H., Wang, Z., Li, J., Wei, Q.: Criminal behavior analysis based on complex networks theory. In: IEEE International Symposium on IT in Medicine and Education, Jinan, vol. 1, pp. 951–955 (2009)
- Yamasaki, K., Gozolchiani, A., Havlin, S.: Climate networks around the globe are significantly affected by El Niño. *Phys. Rev. Lett.* **100**(22), 228501 (2008)
- Donges, J.F., Zou, Y., Marwan, N., Kurths, J.: Complex networks in climate dynamics. Comparing linear and nonlinear network construction methods. *Eur. Phys. J. Special Topics* **174**, 157–179 (2009)
- Ludescher, J., Gozolchiani, A., Bogachev, M.I., Bunde, A., Havlin, S., Schellnhuber, H.J.: Improved El Niño forecasting by cooperativity detection. *Proc. Natl. Acad. Sci. USA* **110**(29), 11742–11745 (2013)
- Deza, J.I., Masoller, C., Barreiro, M.: Distinguishing the effects of internal and forced atmospheric variability in climate networks. arXiv preprint [arXiv: 1311.3089](https://arxiv.org/abs/1311.3089) (2013)
- Tsonis, A.A., Swanson, K.L.: Topology and predictability of El Niño and La Niña networks. *Phys. Rev. Lett.* **100**(22), 228502 (2008)
- Naufan, I., Sivakumar, B., Woldemeskel, F.M., Raghavan, S.V., Vu, M.T., Lioung, S.Y.: Spatial connections in regional climate model rainfall outputs at different temporal scales: application of network theory. *J. Hydrol.* **556**, 1232–1243 (2017)

10. Boers, N., Rheinwalt, A., Bookhagen, B., Barbosa, H.M.J., Marwan, N., Marengo, J., Kurths, J.: The South American rainfall dipole: a complex network analysis of extreme events. *Geophys. Res. Lett.* **41**(20), 7397–7405 (2014)
11. Donges, J.F., Zou, Y., Marwan, N., Kurths, J.: The backbone of the climate network. *Europhys. Lett.* **87**(4), 48007 (2009)
12. Tsonis, A.A., Swanson, K.L., Roeber, P.J.: What do networks have to do with climate? *Bull. Am. Meteorol. Soc.* **87**(5), 585–596 (2006)
13. Chidean, M.I., Bulnes, J.M., Bargueño, J.R., Caamaño, A.J., Salcedo-Sanz, S.: Spatio-temporal trend analysis of air temperature in Europe and Western Asia using data-coupled clustering. *Glob. Planet. Change* **129**, 45–55 (2015)
14. Guez, O.C., Gozolchiani, A., Havlin, S.: Influence of autocorrelation on the topology of the climate network. *Phys. Rev. E* **90**(6), 06281 (2014)
15. Davis, K.F., D'Odorico, P., Laio, F., Ridolfi, L.: Global spatio-temporal patterns in human migration: a complex network perspective. *PLoS ONE* **8**(1), 1–8 (2013)
16. Hlinka, J., Hartman, D., Vejmelka, M., Runge, J., Marwan, N., Kurths, J., Palus, M.: Reliability of inference of directed climate networks using conditional mutual information. *Entropy* **15**, 2023–2045 (2013)
17. National Center for Atmospheric Research. <https://www.mmm.ucar.edu/weather-research-and-forecasting-model>
18. Carvalho, D., Rocha, A., Gómez-Gesteira, M., Santos, C.S.: WRF wind simulation and wind energy production estimates forced by different reanalyses: comparison with observed data for Portugal. *Appl. Energy* **117**, 116–126 (2014)
19. Memon, Z.A., Sahito, A.A., Leghari, Z.H., Shaikh, P.H.: Output voltage characteristics of wind energy system considering wind speed and number of blades. *Sindh Univ. Res. J. (Sci. Ser.)* **2**, 281–284 (2016)
20. Goh, K.I., Oh, E., Jeong, H., Kahng, B., Kim, D.: Classification of scale-free networks. *Proc. Nat. Acad. Sci.* **99**(20), 12583–12588 (2002)



Quantifying Life Quality as Walkability on Urban Networks: The Case of Budapest

Luis Guillermo Natera Orozco¹(✉), David Deritei¹, Anna Vancso²,
and Orsolya Vasarhelyi¹

¹ Department of Network and Data Science,
Central European University, Nádor utca 9, Budapest 1055, Hungary
Natera_Luis@phd.ceu.edu

² Corvinus University of Budapest, Budapest, Hungary
<https://networkdatascience.ceu.edu/>

Abstract. Life quality in cities is deeply related to the mobility options, and how easily one can access different services and attractions. The pedestrian infrastructure network provides the backbone for social life in cities. While there are many approaches to quantify life quality, most do not take specifically into account the walkability of the city, and rather offer a city-wide measure. Here we develop a data-driven, network-based method to quantify the liveability of a city. We introduce a life quality index (LQI) based on pedestrian accessibility to amenities and services, safety and environmental variables. Our computational approach outlines novel ways to measure life quality in a more granular scale, that can become valuable for urban planners, city officials and stakeholders. We apply data-driven methods to Budapest, but as having an emphasis on the online and easily available quantitative data, the methods can be generalized and applied to any city.

Keywords: Walkability · Urban networks · Urban development · Life quality

1 Walkability and Liveable Cities

During the 20th century, most cities have evolved to accommodate a car-centric vision [1], allocating a privileged amount of urban space to motorized traffic [2,3]. From a liveability perspective, this situation is suboptimal because the automobile infrastructure dominates and defines the walkable area, increasing car traffic, air pollution and deteriorating walkable conditions.

The concept of walkability is an important factor to consider in connection with liveability. Liveability refers to an environment from an individual perspective [4] which includes “a vibrant, attractive and secure environment for people to live, work and play and encompasses good governance, a competitive economy, high quality of living and environment sustainability” [5]. Thus in a

liveable city, there must be an emphasis not only on sustainable transportation and built environment to reduce the harm on nature [6, 7] but also encouraging citizens to walk for supporting their physical and mental well-being [8]. However, improving walkability is more complex than we would think. Walking should be an available, safe and well-connected mode of transportation, but as Speck put it well, it should be interesting and comfortable as well, to have a feeling of the streets as ‘outdoor living rooms’ [9].

The pedestrian infrastructure that sustains walkability in a city can be described as a network [10]. This approach has been useful to identify street patterns [11, 12] and its evolution [13, 14], measure the morphology of cities [15], and how the streets connectivity impacts on pedestrian volume [16].

The various approaches to create a walkability index or so-called walk score consider mainly the following components: safety and security [17, 18]; convenience, attractiveness and public policy [9, 19], connectedness [20], but also reckon with the land use mix and residential density of the certain area [21]. Another approximation rather accents the importance of its effect on air pollution, health problems, travel costs and even on the sense of community [22]. Thus measuring walkability not only captures the propensity to walk in a city but also includes the components a liveable city must have and support, under the umbrella of sustainability.

There are good examples of how sustainable city development initiatives tackle growing inequalities with data-driven approaches. Long Island used city data to analyze which amenities are needed to increase the quality of life in a newly built environment [23], other cities are investing in smart technologies to develop public transport, connecting spatially discriminated areas [24, 25].

Since the number of components which should be taken into consideration in creating a walkability index is high, the types of data are also mixed and thus difficult to integrate. While the information on connectedness, security, residential density, etc. is quantitative and in general easily available, gaining opinion about attractiveness, convenience, or even about the feeling of security is more complicated. Here we propose to use a data-driven approach as a proxy to quantify life quality, making it reproducible and easily expanded to include different data sources. We apply our methods to Budapest, but as having an emphasis on the online and easily available quantitative data, the methods can be generalized and applied to any city.

2 Data

We work with three different data sources: networks, points of interest and city attributes. The pedestrian network and points of interest were acquired using OSMnx [26], a python library to download and construct networks from OpenStreetMap (OSM). The data contained in OSM is of high quality [27, 28] in terms of correspondence with municipal open data [29] and completeness: More than 80% of the world is covered by OSM [30].

The majority of points of interest were downloaded from OpenStreetMap, from different classification keys (amenity, tourism, shop, office, leisure) using

OSMnx [26]. We filtered the points of interest using the districts' demarcation [31], to get only the data within Budapest boundaries, having, as a result, more than 39,000 data points. We complement the data sets with secondary data sources as specialized directories of doctors and childcare facilities (see appendix).

We categorize the points of interest in six main categories: (I) Family friendliness (Access to education and daycare, and family support services), (II) Access to health care and sport facilities, (III) Art and culture (e.g.: museums, exhibitions), (IV) Nightlife (e.g.: bars, restaurants), (V) Environment (air quality and access to green areas), and (VI) Public Safety. The points of interest and secondary data sources are available at https://github.com/naturaluis/Budapest_LQI.

The district-level data (population and crimes) were obtained from the Hungarian Police's public database, calculated based on the number of crimes committed in public places 100 thousand per capital in 2018 [32]. Population data is coming from the 2016 micro-census conducted by the Hungarian Statistical Bureau [33]. We took into account the air pollution, this data set coming from National Air Pollution Measurement Network [34], containing the geolocation of the air quality stations and different measures (annual median concentration of carbon monoxide, nitrogen dioxide, and PM10 dust).

Accuracy of the Life Quality Index (LQI) model highly depends on how comprehensive the distribution of listed services. We use OSM as our key data source, but to achieve a more comprehensive and country-specific database we collect publicly available data from various Hungarian websites for each category (See Appendix A for databases and sources).

The network contains all the sidewalks and pedestrian designated infrastructure, it is conceptualized as undirected, nonplanar and primal network [10]. The pedestrian network is described as a weighted graph, with its adjacency matrix $W = w_{ij}$ where the weight w_{ij} contains the length between i and j if connected, and 0 otherwise.

We assigned properties to the nodes of the network, matching the nodes with their corresponding districts, then assign nodes as attributes based on the district level data (population and crimes, see Sect. 3.2). For the pollution data, we calculated the corresponding Voronoi cells, for the air quality stations, and matched the nodes with them, we divided the pollution by the number of nodes in each corresponding cell and assigned the value to the nodes (See Sect. 3.3). For the edges, we encoded their length ℓ_{ij} along with the traversal time Tt_{ij} between nodes i and j calculated as $Tt_{ij} = \frac{\ell_{ij}}{ps}$ where ps is the pedestrian speed as a constant rate of 5 km/h.

3 Quantifying Life Quality

The life quality of a person is largely subjective and hard to quantify. However, it is both intuitive and has been scientifically shown that the environment and

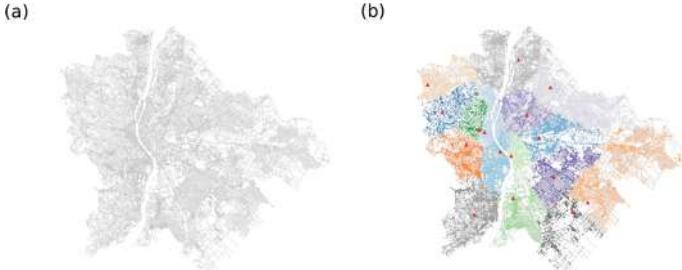


Fig. 1. (a) Network representing Budapest pedestrian structure. The network was built following a primal approach, where the edges are sidewalks and pedestrian infrastructure, and nodes are intersections. (b) The graph-Voronoi tessellation of the Budapest network, generated using a subset of 15 parks as seeds. The color of the nodes represents the cell they belong to and the highlighted red dots are the seeds of each cell. The distance measure between two points is defined as the weighted shortest path on the graph, the weights being the average time required to cross a given edge.

personal well-being strongly correlate [35]. Thus using environmental factors as proxies, life quality and livability becomes quantifiable [36].

The main environmental factors we consider in our model are: the availability of services and amenities, the quality of the infrastructure, environmental factors and safety. The goal of our model is to quantitatively characterize the immediate environment of residents in the space of factors that affect life quality.

The fundamental framework of our model and our calculations is the network representation of Budapest's pedestrian infrastructure. The nodes of the network represent intersections, while links are sidewalks and pedestrian infrastructure. The output of our model is an index, that characterizes every node of the Budapest network, giving a high-resolution quality-landscape of the city. The index is ultimately a number aggregated from multiple sub-categories, and its main value is highlighting inequalities and relative deficiencies within the city.

The final value of the index is a weighted sum, characterizing every node (intersection) in the network:

$$Q_i = w^{services} \tilde{Q}_i^{services} + w^{safety} \tilde{Q}_i^{safety} + w^{environment} \tilde{Q}_i^{environment} \quad (1)$$

In the equation i represents an individual node in the network. The “tilde” above the Q terms means that the values of the different category indices are normalized within the category. The weights w assigned to every term are arbitrary and are highly context-dependent. We include the weights used for producing the results of this paper in Appendix B. All terms of the equation are discussed in the following sections.

3.1 The Services Index: $Q^{services}$

The number quantifying each node in terms of how well it is connected with amenities and services is a weighted sum of sub-categories as well.

$$Q_i^{services} = \sum_c w^c Q_i^c \quad (2)$$

where, c denotes categories (family, culture, health, sport, and nightlife), and w^c the importance (weight) of category c . Some categories, like family, have further subcategories. Even though we have also had data and made a separate analysis on tourism, its effects on life quality of the residents are ambiguous, so we decided to omit it from the index.

What sets categories apart is that they incorporate different sets of amenities, with a few overlaps. The details of the categorization of amenities are included in the Appendix.

For every service/amenity class we have a given set of points of interest (POI) along with where the amenities of that class are available, with exact geo-location. We assign every POI of a given amenity class (e.g supermarket, pharmacy, school, etc.) to the nearest node on the infrastructure network. Each set of POIs organically generates a spatial partitioning of the city with one partition per POI. The partition of a POI is the set of all the nodes from which that particular POI can be reached faster than any other POI of the same class.

Mathematically these partitions are called graph-Voronoi cells [37,38], where every node of a cell is assigned to its closest seed (POI). Distance, in this case, is not euclidean or geometric distance, but the distance on the network, where we use the weighted shortest path between two nodes as the distance measure. The weight of links is a temporal parameter encoding the average time required to cross the represented street from one end to the other, thus the weight is a simple product of average speed and length of the street. This is in principle very similar to the way navigation systems find routes between points. For an example of a graph-Voronoi partitioning see Fig. 1(b).

To assess how well connected a node is to amenities we consider the following factors:

- How important is an amenity - weight (w_a)
- How long does it take to reach the amenity - time to reach (t_{ia})
- Relatively how many nodes (or people) does the amenity share with - exclusivity (P_a)

From the three factors, the latter two are calculated using the city infrastructure network. The index for an amenity class, from the perspective of node i , is proportional with its importance (weight, see Appendix B) and it is inversely proportional with the time to reach the closest POI from i and with the degree of exclusivity.

$$q_i^a = \frac{w_a}{(P_a + 1)(t_{ia} + 1)} \quad (3)$$

There can be certain singular cases when a Voronoi cell is empty ($P_{a=0}$, i.e. no residents in the area) or the node i in question is right at the POI ($t_{ia} = 0$). To avoid anomalies in the index we added 1 to both parameters.

The index of one category is proportional to the sum of its amenity-indices (calculated in (3)). To treat this number on the right scale (in practice we can get very large and very small numbers) we take the natural logarithm of the sum across amenities.

$$Q_i^c = \log\left(\sum_a q_i^a\right) : \quad (4)$$

As we have mentioned earlier the final services index is the weighted sum of the indices of the sub-categories.

$$Q_i^{services} = \sum_c w^c Q_i^c$$

Finally we normalize the values of $Q^{services}$ so its values are comparable to the other values of the final Q Eq. (1):

$$\tilde{Q}_i^{services} = \frac{Q_i^{services} + |min(Q^{services})|}{max(Q^{services}) + |min(Q^{services})|} \quad (5)$$

3.2 Safety Index: Q^{safety}

The safety index is calculated across districts based on the number of crimes committed per one hundred thousand residents. Since the highest resolution data available to us was on the district level, every node i in the same district will have the same safety index value. The crime index:

$$Q_i^{crime} = \frac{N_{crime}^{district}}{n_i^{district}}$$

where $N_{crime}^{district}$ is the number of crimes committed in a district in a year, and $n_i^{district}$ is the number of nodes in the district. The safety index is one minus the normalized crime index.

$$\tilde{Q}_i^{safety} = 1 - \frac{Q_i^{crime}}{max(Q^{crime})} \quad (6)$$

3.3 Environmental Index $Q^{environment}$

The environmental index is made up of two components: air pollution ratio and ratio of natural areas.

Air Pollution Ratio. We use the data provided by Budapest's air pollution measuring stations for the year 2018. For this study, we used the yearly median value of three polluters: carbon monoxide, nitrogen dioxide, and PM10 dust-pollution. As an approximation, we project the geometric Voronoi cells of the measuring stations onto the city map and each node will receive the pollution metrics of the geometrically closest station. We divide these values with the

yearly upper health limit for the given polluter to assess to what degree do these values approximate the health limit. Thus the air pollution index of one node is formalized as follows:

$$C_i = \frac{c_i^{CO}}{c_{limit}^{CO}} + \frac{c_i^{NO_2}}{c_{limit}^{NO_2}} + \frac{c_i^{Pm10}}{c_{limit}^{Pm10}}$$

where $c_{limit}^{CO} = 3000 \text{ g/m}^3$, $c_{limit}^{NO_2} = 40 \text{ g/m}^3$ és $c_{limit}^{Pm10} = 40 \text{ g/m}^3$ a are the yearly upper limits based on [39].

Ratio of Natural Areas. For this index, we have data on the neighborhood level, which is a more granular level of administrative partitioning the city than the districts are. In this case, we project the same index onto every node in the same neighborhood. We consider as natural areas forests, parks and water surfaces (ponds, rivers, etc).

The index:

$$T_i = \frac{R_{water}^{nh(i)} + R_{forest}^{nh(i)} + R_{park}^{nh(i)}}{\max(T)}$$

where $R_x^{nh(i)}$ is the relative surface area of natural area x within the neighborhood that i belongs to ($nh(i)$). In other words, the surface area of a natural area is divided by the number of nodes in the neighborhood and the surface area of the neighborhood. Thus $R_x^{nh(i)} = \frac{T(x)}{T(nh(i))n_{nh}}$, where $T(x)$ is the surface area of x natural area, $T(nh)$ is the surface area of nh neighborhood and n_{nh} is the number of nodes in neighborhood nh . The final environmental index:

$$Q_i^{environment} = \frac{1 + T_i}{1 + C_i}$$

That after a normalization is:

$$\tilde{Q}_i^{environment} = \frac{Q_i^{environment}}{\max(Q^{environment})}$$

4 Results

We quantify life quality in terms of each category (family support, education healthcare, sport, culture, nightlife, environment), and an overall measurement which contains all 6 categories and crime rate normalized by the population for the city of Budapest. Our method allows us to measure life quality for each intersection of the city, which helps to capture within neighborhood inequalities too. Analysis on the category level is beneficial for targeted policy interventions for better service allocation.

Figure 2 shows our overall life quality index (LQI) and by categories. Heatmaps reveal important features of Budapest. Similarly to most European

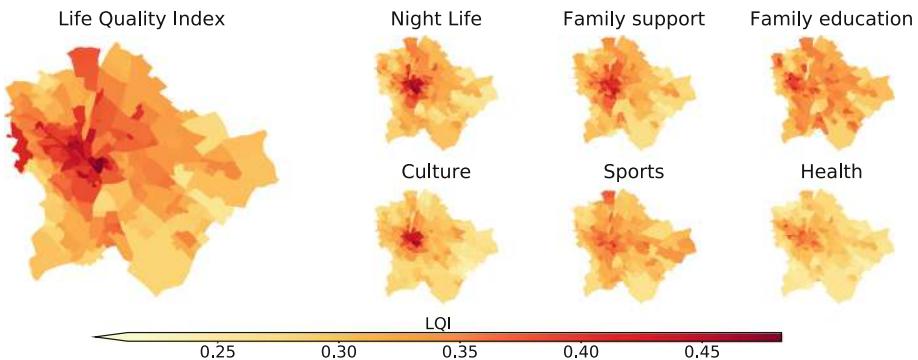


Fig. 2. Budapest neighborhoods, average life quality by categories and aggregated life quality index.

cities life quality is much better in the inner districts [40, 41], especially in the case of Night Life and Culture.

Budapest is divided by the Danube river into two main parts: Buda and Pest. The river does not only serve as a geographical border but due to historical reasons, it also divides the citizens by social status. Hilly Buda, on the West side of the river, used to be the capital of the country, with the residence of the former Hungarian king. On the other side, the mainly flat Pest used to be the agricultural supporter of the aristocrats in Buda [42]. Even though the city has changed dramatically since the Monarchy, the division of Buda and Pest persists, and our life quality index captures it well. However certain services are legally guaranteed to be evenly distributed in the city, such as education and healthcare, for precise modeling one should take into account private care too, which highlights inequalities. So, the traditional division of Buda and Pest is even visible in categories where there should not be that much of a difference (Education, Family Support, Healthcare).

Results also highlight that category LQI-s are highly correlated, less liveable neighborhoods are constant regardless of the amenity category, and well-performing neighborhoods do not change either. It is caused by two main factors: the lack of amenities and the relatively high walking distances in the suburbs.

The compact city concept focuses on building more sustainable and livable cities while designing practical neighborhoods where citizens can maintain everyday life without a car [43]. Since, the walkability of a neighborhood highly correlates with its liveability [44] and the suburbs in Budapest do not show any compact city design features, both long distances and the lack of amenities effects suburban habitats lives negatively.

4.1 Evaluation

Multiple methods have been developed to evaluate the accuracy of quality of life metrics: Scholars used expert validation with geographic visualization [45],

[46], correlations with socioeconomic characteristics [47] and surveying citizens' perceptions of the conditions of life [48].

Our evaluation is based on the micro-economical *hedonic approach* of estimating the values of public goods. In a capitalist market, real estate prices reflect the recognition of a neighborhood's characteristics: Prices are formed based on demand, more desirable places are more expensive, due to the underlying assumption of providing a higher quality of life [41]. Estimating neighborhoods life-quality with real estate prices has a long tradition in urban literature [49–51], therefore we adopt this method to evaluate our model.

We collected the average m^2/EUR price for all 23 districts of Budapest in January 2019 [52] and correlated each LQI category averaged by district with it. Figure 3 shows that our overall LQI correlates the most ($R=0.91$) with the real-estate prices. Most of its components have a positive correlation with real-estate prices, except the environment which is calculated based on air pollution and green surface proximity. The life quality (LQI) in Budapest is much higher in densely populated downtown districts, which are lack of green surface and suffers from high air pollution due to heavy traffic.

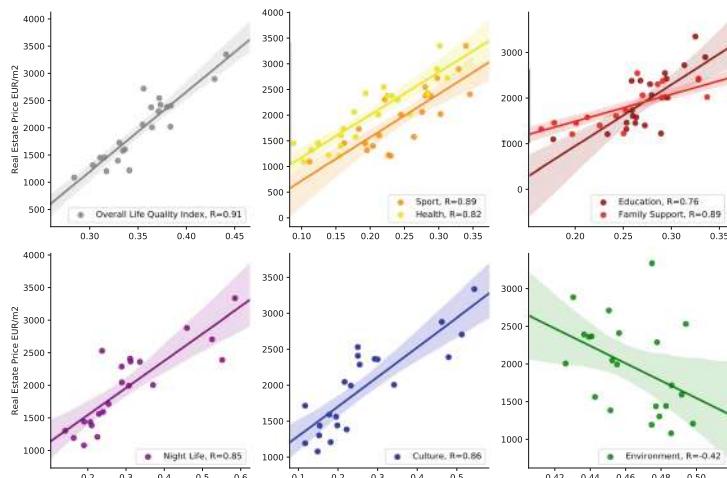


Fig. 3. Districts of Budapest Life Quality Index (LQI) and its components correlated with real estate prices (m^2/EUR)

Summary. Locals of Budapest, like in most European cities, traditionally values downtown areas. The relative closeness to CBD, good access to public transport, and vital city life kept it as a desirable area for living [53]. However, in recent years, the city is facing new challenges: due to gentrification [54] and over-tourism (eg.: Airbnb) real estate prices are sky-rocketing in downtown areas. In contrast with the early 2000-s when (upper) middle-class moved to the suburbs, nowadays, lower-income families and young professionals are leaving the downtown behind in hope for more affordable living.

As our findings show, Budapest is quite centralized and the quality of life highly correlates with real estate prices, which possibly lead to even more inequalities in the future. This spatial discrimination with longer traveling time, less fulfilling environment, and potential segregation reduces the chances of upward mobility and the quality of life of individuals [55].

5 Discussion

We have proposed a methodology to quantify life quality as a function of walkability on urban networks. We have used open data to capture inequalities between neighborhoods and districts in the city. We have shown that the real estate market reflects the life quality that our methods found.

A data-driven approach for quantifying life quality at such a granular level like our proposed method can help decision-makers to tackle social and environmental challenges better. Designing compact, liveable neighborhoods, considering also the upcoming environmental crisis is the number one priority of many cities worldwide.

The use of open data sources and algorithmic approaches adds up towards a systematic framework for understanding urban liveability. Our current approach is not the last word in this development since it does not yet account for multiple other variables, such as the quality of services and infrastructure, and other qualitative variables. To capture the more specific indicator of liveability in different cities it would be necessary to work with more granular and city-dependant data.

We anticipate a future stream of research focused on the use of worldwide open data sets to quantify urban liveability, including longitudinal studies in multiple cities, along with algorithmic modeling, simulations, and machine learning approaches, to first quantify the liveability, propose changes and test them with the ground truth data.

Acknowledgments. The authors wish to thank the experts of KKBK for consultations, and Federico Battiston and Gerardo Iñiguez for comments and discussions on the subject.

Appendix

A Secondary Data Sources

- Sport associations in Budapest [56]
- Kindergartens, daycares, primary and secondary education [57]
- Art and music schools [58]
- Child health services [59]
- Social welfare system (eg.: elderly care) [60]
- Culture centers [61]
- Indoor playgrounds [62]

- Healthcare (hospitals, private and public clinics, specialists) [63]
- Fitness and training facilities [64]
- Outdoor fitness facilities [65]
- Thermal baths and spa [66]
- Playgrounds and parks [67]

B Weights Used in the Calculations

The weights of the different Q indices in the final aggregation as well as in sub-categories highly depends on the context and the nature of the problem. Here we present the values we used to generate the results of this study, that were agreed upon consulting with experts. The weights of the sub-indices from Eq. (1) are of the following values:

$$w^{services} = 0.7$$

$$w^{safety} = 0.1$$

$$w^{environment} = 0.2$$

The category weights used in Eq. (2), aggregating $Q^{services}$ are:

$$w^{family} = 0.3;$$

$$w^{health} = 0.3;$$

$$w^{culture} = 0.15;$$

$$w^{sport} = 0.15;$$

$$w^{nightlife} = 0.1;$$

References

1. Jacobs, J.: *The Death and Life of Great American Cities*. Random House, New York (1961)
2. Gössling, S., Schröder, M., Späth, P., Freytag, T.: Urban space distribution and sustainable transport. *Transp. Rev.* **36**(5), 659–679 (2016)
3. Szell, M.: Crowdsourced quantification and visualization of urban mobility space inequality. *Urban Plan.* **3**(1), 1 (2018)
4. Heylen, K.: Liveability in social housing: three case-studies in Flanders. Paper Presented at the ENHR Conference “Housing in an Expanding Europe: Theory, Policy, Participation and Implementation”, July 2006
5. Shamsuddin, S., Hassan, N.R.A., Bilyamin, S.F.I.: Walkable environment in increasing the liveability of a city. *Proc. Soc. Behav. Sci.* **50**(167–178), 169 (2012)
6. Campbell, S.: Green cities, growing cities, just cities?: urban planning and the contradictions of sustainable development. *J. Am. Plan. Assoc.* **62**(3), 296–312 (1996)
7. Jabareen, Y.: Planning the resilient city: concepts and strategies for coping with climate change and environmental risk. *Cities* **31**, 220–229 (2013)
8. Frank, L.D., Sallis, J.F., Conway, T.L., Chapman, J.E., Saelens, B.E., Bachman, W.: Many pathways from land use to health: associations between neighborhood walkability and active transportation, body mass index, and air quality. *J. Am. plan. Assoc.* **72**(1), 75–87 (2006)

9. Speck, J.: Walkable City: How Downtown Can Save America, One Step at a Time. North Point Press, New York (2012)
10. Porta, S., Crucitti, P., Latora, V.: The network analysis of urban streets: a primal approach. *Environ. Plan. A* **33**(5), 705–726 (2006)
11. Barthélémy, M., Flammini, A.: Modeling urban street patterns. *Phys. Rev. Lett.* **100**(13) 138702 (2008)
12. Louf, R., Barthélémy, M.: A typology of street patterns. *J. R. Soc. Interface* **11**(101), 20140924 (2014)
13. Strano, E., Nicosia, V., Latora, V., Porta, S., Barthélémy, M.: Elementary processes governing the evolution of road networks. *Sci. Rep.* **2**(1), 296 (2012)
14. Barthélémy, M., Bordin, P., Berestycki, H., Gribaudi, M.: Self-organization versus top-down planning in the evolution of a city. *Sci. Rep.* **3**(1), 2153 (2013)
15. Boeing, G.: The morphology and circuitry of walkable and drivable street networks. In: Modeling and Simulation in Science, Engineering and Technology, pp. 271–287, Birkhäuser, Cham (2019)
16. Hajrasouliha, A., Yin, L.: The impact of street network connectivity on pedestrian volume. *Urban Stud.* **52**(13), 2483–2497 (2015)
17. Daniele, Q., Aiello, L.M., Schifanella, R., Davies, A.: The digital life of walkable streets. In: Proceedings of International World Wide Web Conference (2015)
18. Silva, J.P., Akleh, A.Z.: Investigating the relationships between the built environment, the climate, walkability and physical activity in the Arabic peninsula: the case of Bahrain. *Cogent Soc. Sci.* **4**(1), 1–21 (2018)
19. Krambeck, H.V.: The global walkability index (Doctoral dissertation). Massachusetts Institute of Technology (2006)
20. Southworth, M.: Designing the walkable city. *J. Urban Plan. Dev.* **131**(4), 246–257 (2005)
21. Carr, L.J., Dunsiger, S.I., Marcus, B.H.: Walk scoreTMas a global estimate of neighborhood walkability. *Am. J. Prev. Med.* **39**(5), 460–463 (2010)
22. Stephen, M.: A Better Urban Design of Cities is Closely to Sustainable Planning. Rutledge Urban Reader Series, United States (2004)
23. Childs, S.: A Case for Data-Driven City Planning Neighborhood Knowledge Supports Resilient Communities. <https://medium.com/citiesense/a-case-for-data-driven-city-planning-8cd7d9332a> (2018)
24. Kaushik, V.: How Technology Empowers Data-Driven Urban Planning? <https://www.getrevue.co/profile/TGIC/issues/how-technology-empowers-data-driven-urban-planning-59814> (2017)
25. Fitzgerald, M.: Data-Driven City Management A Close Look At Amsterdam Smart City Initiative. MIT Sloan Management Review (2016)
26. Boeing, G.: OSMnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Comput. Environ. Urban Syst.* **65**, 126–139 (2017)
27. Haklay, M.: How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. *Environ. Plan. A* **37**(4), 682–703 (2010)
28. Girres, J.F., Touya, G.: Quality assessment of the French OpenStreetMap Dataset. *Trans. GIS* **14**(4), 435–459 (2010)
29. Ferster, C., Fischer, J., Manaugh, K., Nelson, T., Winters, M.: Using OpenStreetMap to inventory bicycle infrastructure: a comparison with open data from cities. *Int. J. Sustain. Transp.* 1–10 (2019)
30. Barbosa-Filho, H., Barthélémy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tommasini, M.: Models and applications: human mobility (2017)

31. Budapest districts. <https://data2.openstreetmap.hu/hatarok/index.php?admin=9>
32. Official Site of the Hungarian Police. <http://www.police.hu/a-rendorsegrol/statisztikak/bunugyi-statisztikak>. Data Available at request
33. Hungarian popultaion data 2016. <https://www.ksh.hu/mikrocenzus2016/>
34. National Air Pollution Measurement Network, Automatic Measurement Network, Hungary. <http://levegominoseg.hu/manualis-merohalozat>. Data Available at request
35. Rosow, I.: The social effects of the physical environment. *J. Am. Inst. Plann.* **27**(2), 127–133 (1961)
36. Kahneman, D., Krueger, A.B.: Developments in the measurement of subjective well-being. *J. Econ. Perspect.* **20**(1), 3–24 (2006)
37. Erwing, M.: The graph Voronoi diagram with applications. *Netw.: Int. J.* **36**(3), 156–163 (2000)
38. Deritei, D., Lazar, Z.I., Papp, I., Jarai-Szabo, F., Sumi, R., Varga, L., Regan, E.R., Ercsey-Ravasz, M.: Community detection by graph Voronoi diagrams. *New J. Phys.* **16**, 063007 (2014)
39. <http://levegominoseg.hu/jogszabalyok>
40. Hohenberg, P.M., Lees, L.H.: *The Making of Urban Europe 1000–1950*. Harvard University Press, Cambridge (1986)
41. Brueckner, J.K., Thisse bc, J.F., Zenou bd, Y.: Why is central Paris rich and downtown Detroit poor?: an amenity-based theory. *Eur. Econ. Rev.* **43**(1), 91–107 (1999)
42. Gyani, G., Kover, G.: Magyarorszag tarsadalomtortenete a reformkortol a masodik vilaghaboriig. 2. jav. kiad. Osiris, Budapest (1998)
43. Dittmar, H., Ohland, G.: *The new transit town: best practices in transit-oriented development*. Island Press (2012)
44. Rogers, S.H., Halstead, J.M., Gardner, K.H., Carlson, C.H.: Examining walkability and social capital as indicators of quality of life at the municipal and neighborhood scales. *Appl. Res. Qual. Life* **6**(2), 201–213 (2011)
45. Rinner, C.: A geographic visualization approach to multicriteria evaluation of urban quality of life. *Int. J. Geogr. Inf. Sci.* **21**(8), 907–919 (2007)
46. Gavrilidis, A.A., et al.: Urban landscape quality index planning tool for evaluating urban landscapes and improving the quality of life. *Proc. Environ. Sci.* **32**, 155–167 (2016)
47. Talen, E.: Pedestrian access as a measure of urban quality. *Plan. Pract. Res.* **17**(3), 257–278 (2002)
48. Santos, L.D., Martins, I.: Monitoring urban quality of life: the porto experience. *Soc. Indic. Res.* **80**, 411 (2007)
49. Roback, J.: Wages, rents, and the quality of life. *J. Polit. Econ.* **90**(6), 1257–1278 (1982)
50. Blomquist, G., Berger, M., Hoehn, J.: New estimates of quality of life in urban areas. *Am. Econ. Rev.* **78**(1), 89–107 (1988)
51. Lora, E., Powell, A.: A new way of monitoring the quality of urban life. *IDB working paper series*, 272 (2011)
52. Budapest Real Estate Statistics. <https://www.ingatlannet.hu/statisztika/Budapest>
53. Cassiers, T., Kesteloot, C.: Socio-spatial inequalities and social cohesion in European cities. *Urban Stud.* **49**(9), 1909–1924 (2012)
54. Garcia, B.: Cultural policy and urban regeneration in Western European cities: lessons from experience, prospects for the future. *Local Econ.* **19**(4), 312–326 (2004)

55. Gobillon, L., Selod, H., Zenou, Y.: The mechanisms of spatial mismatch. *Urban Stud.* **44**, 2401–2427 (2007)
56. Sport associations in Budapest. https://sportmegoldasok.hu/klub_budapest.ht. Accessed 01 Jan 2019
57. Kindergartens & daycare, primary education, secondary education. <http://budapest.imami.hu/iskolak-ovodak-bolcsodek/>. Accessed 01 Jan 2019
58. Art and music schools. <http://budapest.imami.hu/iskolak-ovodak-bolcsodek/muveszeti-es-sportiskolak>. Accessed 01 Jan 2019
59. Pediatricians, Gynecologists. <http://budapest.imami.hu/egeszseg>. Accessed 01 Jan 2019
60. Social Welfare. <http://szocialisportal.hu/intezmenykereso>. Accessed 01 Jan 2019
61. Culture centers, Budapest. <https://holmivan.valami.infomuvelodes-kultura-klub-lista-122muvelodesi-kozpont>. Accessed 01 Jan 2019
62. Indoor playgrounds. <http://budapest.imami.hu/szolgaltatok/szulinapi-party>. Accessed 01 Jan 2019
63. Healthcare (hospitals, private and public clinics, specialists). <https://www.budapestinfo.eu/gyogyitas>. Accessed 01 Jan 2019
64. Fitness and Training facilities. <https://www.budapestinfo.eu/sportolas/fitnesstermek>. Accessed 01 Jan 2019
65. Outdoor fitness facilities. <https://www.google.com/maps/d/u/0/viewer?msa=0&hl=en&ie=UTF8&t=h&ll=47.481378772238784%2C19.130302999999913&spn=1.329955%2C3.757785&source=embed&mid=1qIGN-mnKuGrvNxo0ENmACm78RP8&z=11>. Accessed 01 Jan 2019
66. Thermal and Spa. <https://www.budapestinfo.eu/furdozes>. Accessed 01 Jan 2019
67. Playgrounds and Park. <https://zoldkalausz.hu/>. Accessed 01 Jan 2019



A Network Theoretical Approach to Identify Vulnerabilities of Urban Drainage Networks Against Structural Failures

Paria Hajiamoosha^(✉) and Christian Urich

Monash University, Wellington Rd., Clayton, VIC 3800, Australia
{paria.hajiamoosha,Christian.urich}@monash.edu

Abstract. This paper compares two different representations (primal-mapping and dual-mapping) and applies a range of topological metrics to analysis the vulnerability of an urban drainage network (UDN) against structural failures, which is examined in the Elster Creek catchment. Based on the node degree distribution, the properties of the dual graph are similar to a scale-free network while the primal graph behaves like a random network. Further, the results show that the structure of the dual graph has better connectivity and redundancy compared to the structure of the primal graph. To identify vulnerabilities this paper test's a new centrality metrics, that modifies betweenness centrality based on the UDN's performance. This new metric ranks the most vulnerable conduit in the UDN. To validate the ranking a hydrodynamic model is used as a reference. The results show the significance of the structural metric in identifying critical components of the UDN and suggest a dual representation is an appropriate method for investigating vulnerabilities of an UDN against structural failures.

Keywords: Graph modelling · Network science · Urban drainage network · Vulnerability

1 Introduction

Urban drainage networks (UDNs) are a merit part of urban infrastructure that convey stormwater and protect people, their assets, and the environment from flooding [1]. Although, design, management, and maintenance practises have been continuously improved over the years, there are several functional (e.g. climate changes and population growth), and structural issues (e.g. aging infrastructure) that affect the performance of UDNs [2, 3]. More frequent structural failures caused by blockages and pipe collapse, are increasing the vulnerability [2, 3] and lead to increased flooding. A better understanding of the structural and the functional complexity of UDNs, particularly in large cities, to identify critical components of UDNs is therefore key to improve the performance and reduce the vulnerability.

Network science has received more and more attention to better understand the underlying structure and function of man-made and natural systems ranging from urban infrastructure networks (e.g. road networks and transport systems, power grids,

and urban water networks) to social networks [5, 6]. Network science has been applied to investigate structure, connectivity, evolution, vulnerability, and robustness of urban infrastructure systems [4, 9, 12–15].

To analyse the underlaying properties of an urban infrastructure network requires the mapping of the network as a graph. The most intuitive method to translate an urban infrastructure network to a graph is known as primal-mapping. In this way, the spatial layout of the urban infrastructure network is used as a basis for the generated graph [6, 7, 16–19]. The primal graph representation of urban infrastructure networks has been successfully applied to investigate evolution, vulnerability, and resilience of the networks. For example, Agathokleous et al. modelled a Water Distribution Network (WDN) as a primal graph to assess the network's vulnerability. The authors show that the topology of the WDN during different operating conditions (different pressure) can predict the WDN's behaviour and help to optimise it and minimise the vulnerability [9]. However, primal representation is not enough to describe complexity an urban infrastructure network, and is not able to give us much information about structure of the network and the patterns [7, 8, 16].

In this regard, researchers have in recent years successfully applied a dual-mapping approach to analyse urban transport systems and road networks. Instead of using the spatial layout of the infrastructure to describe the links and nodes, a dual mapped graph representation uses a property of the network to form nodes and links. This representation is based on the information space of the network [7, 8, 16–19]. The dual-mapping approach makes the most effective use of the networks' information to optimise process according to the main aims of researches [16–19, 24]. In literature, the dual representation has been employed to model physical networks (mostly road networks, and a limited number of urban water networks) [7, 8, 16–19, 24]. These studies showed that such an approach can reveal a networks' evolution and topological patterns [7, 8, 16–19]. For example, Yang et al., and Krueger et al. showed that the patterns governing the evaluation of man-made networks are similar to patterns found in natural scale-free networks e.g. rivers [7, 26].

Building on this research, the aim of this paper is to explore if a dual representation of an urban drainage network (UDN) can give insights into its vulnerability, by identifying its most critical elements.

In this paper, we classify an UDN using primal and dual representation, and analyse and compare the structure of the network using graph's metrics such as node degree, degree distribution, and clustering. Additionally, we apply a centrality metric to investigate the vulnerability of the network, and to identify the most critical elements in the network for both models in relation to urban flooding. Finally, we compare the results of both representations to find critical elements and investigate the vulnerability. To justify these results, we use a hydrodynamic model for the UDN.

2 Methodology

2.1 Generate Graphs and Their Topological Analysis

Urban Drainage Network. UDNs like other urban water networks and urban infrastructure networks (road networks), are planar graphs. In addition, UDNs are considered as a gravity driven system, in which flow has a specified direction in the system. The system is therefore considered as a directed graph based on the flow direction. A directed graph (digraph) is a network in which all links have a specific direction from one node to others [5, 6].

Primal-Mapping Representation. A primal-mapping graph can be used to convert an infrastructure spatial map into a graph and analyse its properties. In a primal-mapping graph, links (edges) are usually considered as network segments, and the intersections of the segments are mapped as nodes [4, 7, 8, 16–19, 25]. In this step, the UDN is mapped into a primal graph, in which the nodes and the links represent junctions and conduits (pipes), respectively. To generate a directed primal graph the UDN is directed based on the water's flow direction in the pipe (see Fig. 1).

Dual-Mapping Representation. A dual-mapped graph is a representation of a network where nodes are network segments (pipes), which are merged based on a common attribute and links are intersections [7, 8, 16–19, 24]. In a dual representation, the main problem is to determine which segments belongs to the same node. Two popular methods have been used to dual-map road networks. The first one is the intersection continuity negotiation (ICN) method, which is based on the geometrical properties of the primal graph. The ICN method merges aligned road segments to a node. Although, this method works well on street networks, it often causes misleading outcomes for complex geometries. [7, 16–19]. The other method is the street name approach (SN), which is based on the information space of the network. The SN approach works based on the simple principle that two contiguous street segments, with the same street name, merge to one node [7, 16–19]. The Hierarchical Intersection Continuity Negotiation (HICN) combines the ICN and SN approaches which is used in this paper. For the ICN method, geometrical data is used to generate a directed graph (based on the pipes' elevation), so that multiple contiguous pipes, which can merge, are represented as one node if the pipes follow the same flow direction. Similar to the SN approach the pipe diameter [7] is considered as criterion for merging pipes to segments to organize the network into functional units based on flow capacity and the modelled pipe blockages. Using these approaches we aim to analyse the impact of structure, flow capacity, and their changes on the components of the UDN. Furthermore, we connect the structure of the network (blockage of pipes) to its function (flood) which leads to changes of the flood volume. Figure 1 shows the primal-mapping approach versus the dual-mapping approach used in this paper.

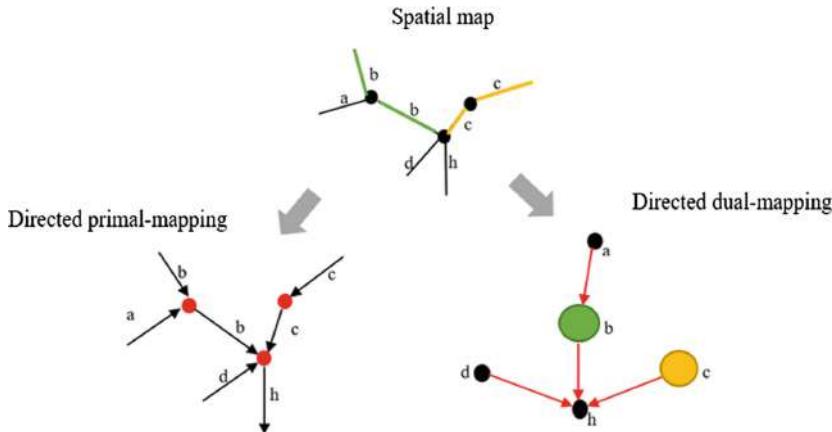


Fig. 1. The primal-mapping approach versus the dual-mapping approach, top: spatial map, bottom – left: directed primal-mapping, and bottom – right: dual-mapping approach.

Graph Metrics. To quantify the structure, connectivity, vulnerability and robustness of graphs [5, 6, 20] this paper applies following graph matrices (see Table 1).

Table 1. Graph metrics that applied here.

Graph property	Application	Definition	Equation
Graph density (q)	Indicate sparseness or dense-connectivity	Fraction between total and maximum possible links of a graph [4, 7, 11–14]	$q = \frac{2L}{N(N-1)}$
Clustering (C)	Connectivity	Degree of a node for tending to cluster together in a graph [6]	$C_i = \frac{2L_i}{k_i(k_i-1)}$
Average clustering coefficient ($\langle C \rangle$)	Connectivity	Average of C_i for all nodes of graph, and shows probability that two neighbours of a node connect to each other [6]	$\langle C \rangle = \frac{\sum_{i=1}^N C_i}{N}$
Meshedness coefficient (MC)	Indicator of cycles and loops density	Ratio of actual number of loops per maximum possible number of loops in a graph [4, 11, 14]	$MC = \frac{L-N+1}{2N-5}$
Transitivity (T) or global clustering coefficient	Connectivity	The fraction between the total number of triangles $N\Delta$ and the number of connected triples N_3 in the network [4, 7, 10, 13–15]	$T = \frac{3N\Delta}{N_3}$
Node degree (k)	Structural measurement	Number of links of a node in a graph [4, 7, 11–15, 24]	–
Average node degree ($\langle k \rangle$)	Connectivity	Average value of the node degree of a graph [4, 7, 11–15]	$\langle k \rangle = \frac{2L}{N}$
Node degree distribution (p_k)	Structural measurement and indicate type and performance of graph	Probability that a node has degree k in a graph [4, 7, 11–15, 24]	$p_k = \frac{N_k}{N}$

Indicators for Connectivity. Connectivity is an important factor in investigation the network's vulnerability [4, 7, 14]. Following indicators have been identified in literature:

- Average node degree distribution which determines the structure of the network. Urban water networks generally have a structure between tree-like (for $\langle k \rangle \leq 2$) and 2D grid (for $\langle k \rangle \leq 4$) [4].
- Graph density, which indicates how well the nodes are connected in a graph, and ranges between zero (for sparse graphs) and one (for dense graphs) [4, 6, 7].
- Average clustering coefficient computes how neighbours of nodes connect to each other. In the other words, it measures local link density of a graph [6]. This metrics also measures modularity of a graph, and $\langle C \rangle < 0.1$ shows that modular organisation of the graph is weak [7].
- Meshedness coefficient and transitivity have an important role in the quantification of redundancy by computing the percentage of loops and triangular loops [4].

Node Degree and Degree Distribution. Degree distribution (p_k) has a key role in graph theory since the functional form of p_k determines the type of graphs which is important in understanding many network phenomena like robustness [6, 16].

2.2 Urban Drainage Network Vulnerability Index

To investigate the vulnerability of the UDN and to identify the most important nodes a centrality metric is applied according to the topological structure of the network [9–11]. Betweenness centrality is a measurement of centrality. This metric is defined as a fraction between the number of shortest paths between two nodes (u and j) that include node i, and the number of shortest paths between two nodes (u and j) [9–11].

The importance of a node has a different meaning for different network types. In this case, the network has been generated from an UDN. The UDN is a gravity system that includes multi sources (input points in upstream) to collect runoff, and a few targets (output points downstream) to drain the runoff out of the system. In the other words, all source nodes are linked to one of the targets. Based on the performance of the UDN, we modify the betweenness centrality and introduce the Urban Drainage Network Betweenness Centrality (UDNBC) defined by the betweenness centrality of node i as the number of shortest paths between source nodes and target nodes that include node i. Therefore, this metrics indicates the most important pipes, which drain high volume of runoff, and their failure leads to more flood volume in the system. This measurement is applied to rank the nodes and find the most critical nodes in the network.

2.3 Hydrodynamic Simulation

The assess the impact of pipe failure in the UDN and link the structure of the network with its function a hydrodynamic model is used as benchmark for the results of the graph theory model. The simulation is undertaken by EPA Storm Water Management Model (SWMM 5.1), which is a simulation tool to model, analyse, and design storm water runoff, sewer, and combined networks for single rainfall event, or long-term

simulation of runoff [23]. The function of the UDN is simulated under a storm event as long-term rainfall series. Pipes are failed one by one by reducing their diameter to 0. The flood volume is computed and saved. The pipes are ranked based on their impact on the flood volume. These results are used as a reference for the degree of importance of nodes in the network.

2.4 Case Study Area

The model is applied to the Elster Creek catchment (Fig. 2). It is located in the southeast of Melbourne, Australia and has an area of approximately 45 km² and includes the municipalities Glen Eira City, Bayside City, Kingston City, and City of Port Phillip. The catchment mainly consists of residential areas with a population of about 100000 households and has had many large floods in recent years [21, 22]. Figure 2 illustrates Elster Creek catchment.

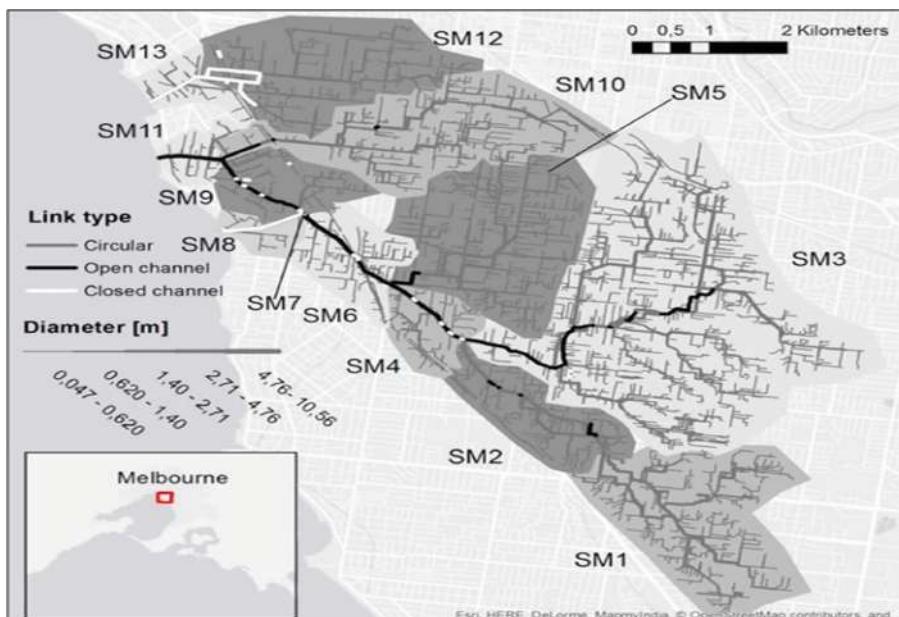


Fig. 2. Case study area

3 Result and Discussion

3.1 Topological Analysis of the Graphs

The Elster Creek urban drainage network has been converted to different primal and dual graph. Table 2 represents the structural metrics of these graphs.

Table 2. Topological metrics of primal and dual graph

Graph type	N	L	$\langle k \rangle$	q	MC	$\langle C \rangle$	T
Primal graph	10008	10305	2.059	0.00021	0.0149	0.0033	0.102
Dual graph	5003	5239	2.094	0.00042	0.0237	0.0057	0.121

For the primal graph, the number of links and nodes is approximately twice the number of links and nodes of the dual graph. This is the results of the merged nodes in the dual graph. The average node degree of both graphs is near 2, identifying the structure as tree-like with rare loops.

Based on the graph density value, primal and dual graph are sparse. This means disconnecting a few links disconnects the graph [4, 14]. In the dual graph, MC, $\langle C \rangle$, and T are greater than the same metrics in the primal graph. That means the percentage of loops (more alternative path between two nodes) in the dual graph is higher than the percentage of loops in the primal graph. Therefore, the dual graph has better connectivity and redundancy compared to the primal graph. Moreover, the average clustering coefficient of the primal and the dual graph is lower than 0.1, which means that the modular organisation of the graphs is weak.

Node Degree and Degree Distribution. Beside the topological metrics, the node degree distribution was analysed for the primal and the dual graph. We consider total k (not k_{in} and k_{out}) for the graphs, and for computing the degree distribution (see Fig. 3). Maximum total degree was 31 and 7, and the majority of nodes had $k = 1$ and $k = 2$ in the dual and the primal graph, respectively. To analyse the behaviour of the graphs, we fitted a power-low function¹ to the node degree distribution of the graphs. In the dual graph, the power low exponent was $\gamma = 2.89$ ($2 < \gamma < 3$). This range is within the scale-free regime, and most real networks are in this regime [6], so that the dual graph is scale-free and follows ultra-small world with the large hubs [6]. In the primal graph, however, γ was greater than three ($\gamma = 3.65$). This graph is in the random network regime and follows the small world with hubs that is not significantly large [6]. Moreover, the properties of the primal graph are similar to the properties of a random graph [6]. According to the results, we also fitted a binomial² and Poisson³ function to primal graph (Fig. 3–down), these distributions, however, did not indicate a better match with the graph. Therefore, although the primal graph behaves like a random graph, it did not follow binomial and Poisson distributions. The degree distributions of the graphs are illustrated in Fig. 3.

¹ $p_k \approx k^{-\gamma}$.

² $p_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}$.

³ $p_k = e^{-\langle k \rangle} \frac{\langle k \rangle^k}{k!}$.

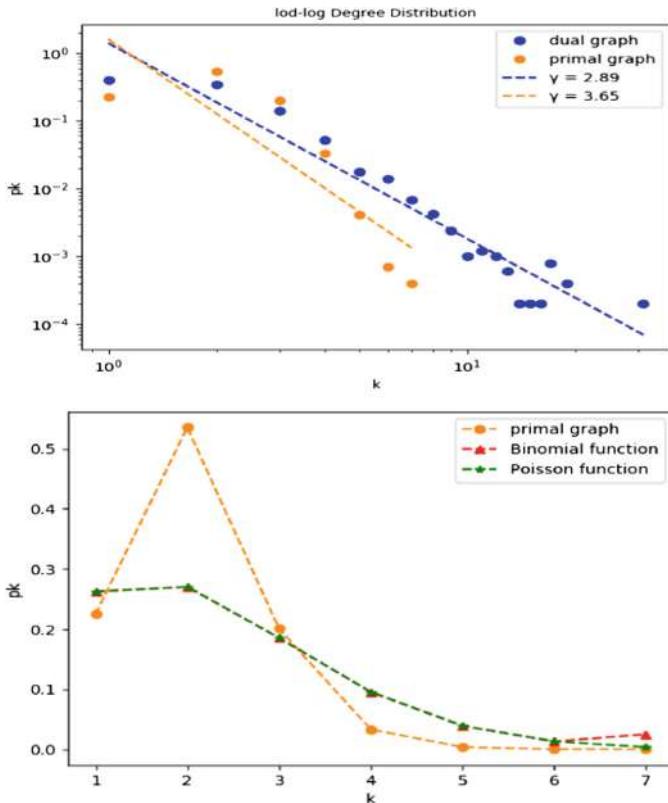


Fig. 3. Node degree distribution; top: log-log degree distribution, down: fitting binomial and Poisson distribution with primal graph distribution.

3.2 Urban Drainage Network Betweenness Centrality (UDNBC)

The UDN has three output points or targets. According to UDNBC definition, we computed the number of shortest paths between all nodes and targets. Most nodes were only connected to one target. Nodes with a high UDNBC are more important in the network. To compare UDNBC of the primal and the dual graph, we applied feature scaling normalization method⁴, which gives a value between 0 (for lowest UDNBC) and 1 (for highest UDNBC). In Fig. 4, the normalized measurement of UDNBC perfectly matched for both the graphs, especially in $UDNBC > 0.4$, and the slope of fitted line with data is near 1 (0.9797). The UDNBC of the primal graph was between 0 and 7182, while it was between 0 and 3642 in the dual graph. In both graphs, the UDNBC of the nodes was divided into the four section. The first section included the greatest values of UDNBC in range 3000–4000 for the dual graph and 6000–8000 for the primal graph (normalized value > 0.9) which indicate nodes further downstream.

⁴ $UDNBC' = \frac{UDNBC - UDNBC_{min}}{UDNBC_{max} - UDNBC_{min}}$.

The UDNBC ranged between 2000–3000, 1000–2000 and 0–1000 in the dual graph, and between 4000–6000, 2000–4000 and 0–2000 in the primal graph for second (normalized value > 0.6), third (normalized value > 0.4) and fourth (normalized value < 0.4) section, respectively.

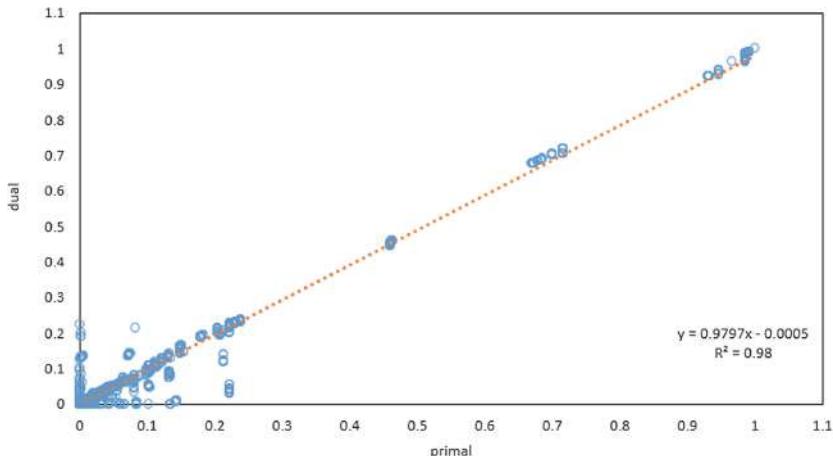


Fig. 4. UDNBC of the dual graph vs UDNBC of the primal graph

We applied UDNBC to rank the conduits based on their importance in the UDN. The degree of the importance of each conduit in an UDN depends on the volume of flooded water where the conduits were completely blocked. Therefore, we expected nodes with a higher UDNBC to cause more flooding. During this process, the maximum increase of flooded volume, is more than 1% (1.1%), and we classified conduits into four parts included equal and more than 1%, 0.5%–1%, 0.1%–0.5%, and less than 0.1%. In this case, therefore, most critical conduits belonged to the first part ($\geq 1\%$).

In Fig. 5, the results of UDNBC were compared to the results of the flooded volume. In the first section of UDNBC, all nodes increased the flood volume by more than 0.5%. Moreover, 100% of nodes, which raised the flood's volume by more than 1% (most critical conduits), were located in this section in both graphs. In addition, the flood volume of the nodes increased by 0.1%–1%, and 0.1%–0.5% in the second and the third section, respectively. Finally, all nodes that increased flooding by less than 0.1%, are in the fourth section. Therefore, the UDNBC can be consider as a trustable metrics to identify critical components in UDNs.

We further tested the ability of the dual graph to identify the most critical elements of the UDN compared to the primal graph. The results indicated that the performance of the dual graph and the primal graph were similar in determining the most critical components of the UDN (increasing flood volume equal or more than 1%). In this case, however, the dual representation is more insightful because it can investigate the impacts of structural changes on the dynamics of functional failures of the UDN. Table 3 shows the portion of nodes of each section in increasing the flood volume in the primal and the dual graph.

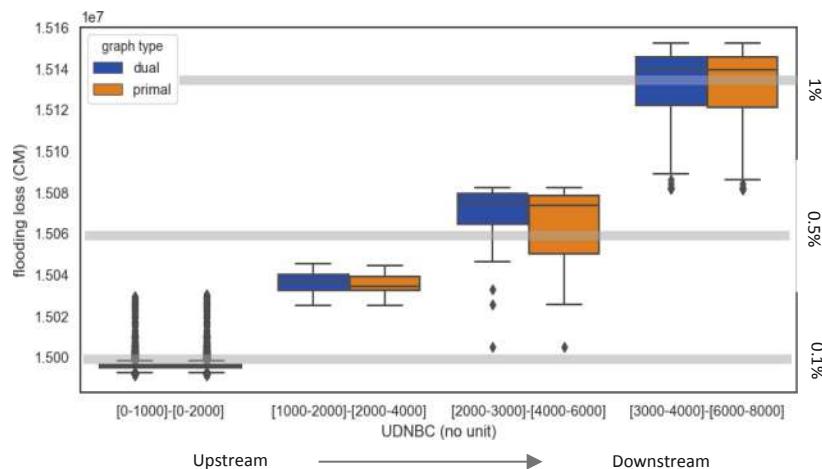


Fig. 5. Comparison of the outcomes of the graph model to hydrodynamic model and their classifications.

Table 3. The portion of nodes of each section in increasing flood volume in the primal and the dual graph.

UDNbc section	Type of graph	Section boundary	Increase of flood volume			
			1% \leq	0.5%–1%	0.1%–0.5%	0.1% \geq
1	Dual	3000–4000	100%	32%	0	0
	Primal	6000–8000	100%	36%	0	0
2	Dual	2000–3000	0	68%	0.1%	0
	Primal	4000–6000	0	64%	0.1%	0
3	Dual	1000–2000	0	0	0.35%	0
	Primal	2000–4000	0	0	0.35%	0
4	Dual	0–1000	0	0	99.55%	100%
	Primal	0–2000	0	0	99.55%	100%

4 Conclusion

This paper has analysed primal-mapping and dual-mapping for modelling the vulnerability of UDNs, and introduced a new centrality metrics based on the UDNs performance. In the first section, the structure of the dual and primal graph has been investigated using graph metrics to quantify the structure, connectivity, vulnerability and robustness of the networks. The connectivity indicators showed that both graphs were tree-like and sparse. Additionally, the dual graph had more connectivity and robustness compared to the primal one. The node degree distribution of the dual graph was a scale-free, while the primal graph behaved like a random graph.

In the second section, we introduced a new centrality metrics UDNBC to measure the degree of importance of each node in the UDN's graph to the vulnerability. The results show that the UDNBC is a reliable indicator to identify the critical elements that led to high flood volume. Moreover, when analyzing the vulnerability of the network in relation to pipe failure and flooding, we found that UDNBC have successfully detected the most critical conduits of the UDN and showed the same results for both representations. Therefore, the UDNBC as a measurement for finding critical elements in UDN does not depend to type of the graph (primal or dual). In addition, the results confirms that a dual graph representation of the UDN is an appropriate method for modelling the vulnerability of an UDN like the primal representation as well. Furthermore, the dual representation that proposed in this paper (using pipe diameter considered as criterion of merging segments) has several benefits in understanding the structure of UDNs. The main benefit is that the graph representation can be applied as a dynamic model for investigating the impact of changes of the pipe diameter (flow capacity) on performance and vulnerability of UDNs based on graph theory (e.g. changes of the pipe diameter lead to change structural metrics of dual graph like degree distribution).

Acknowledgement. The Australian Government, Department of Education and Training - Research Training Program (RTP) funded this work.

References

1. Zhang, C., Wang, Y., Li, Y., Ding, W.: Vulnerability analysis of urban drainage systems: tree vs. loop networks. *Sustainability* **9**, 397 (2017)
2. Mugume, S., Butler, D.: Evaluation of functional resilience in urban drainage and flood management systems using a global analysis approach. *Urban Water J.* **14**(7), 727–736 (2017)
3. Mugume, S., Gomez, D., Fu, G., Farmani, R., Butler, D.: A global analysis approach for investigating structural resilience in urban drainage systems. *Water Res.* **81**, 15–26 (2015)
4. Yazdani, A., Jeffrey, P.: Applying network theory to quantify the redundancy and structural robustness of water distribution systems. *Water Resour. Plann. Manag.* **138**(2), 153–161 (2011)
5. Newman, M.: Networks: An Introduction. Oxford Scholarship Online (2010). <https://doi.org/10.1093/acprof:oso/9780199206650.001.0001>
6. Barabasi, A.: Network Science. Cambridge University Press, Cambridge (2016)
7. Krueger, E., Klinkhamer, C., Urich, C., Zhan, X., Rao, P.: Generic patterns in the evolution of urban water networks: evidence from a large Asian city. *Am. Phys. Soc.* **95**(3) (2017)
8. Yazdani, A., Jeffrey, P.: Water distribution system vulnerability analysis using weighted and directed network models. *Water Resour. Res.* **48**(6) (2012). <https://doi.org/10.1029/2012WR011897>
9. Agathokleous, A., Christodoulou, C., Christodoulou, S.: Topological robustness and vulnerability assessment of water distribution networks. *Water Resour. Manag.* **31**, 4007–4021 (2017)
10. Girvan, M., Newman, E.: Community structure in social and biological networks. *PNAS* **99**(12), 7821–7826 (2002)

11. Ulusoy, A., Stoianov, I., Chazerain A.: Integrating graph theory and hydraulic model-based measures for the analysis of WDN resilience. In: 1st International WDSA/CCWI 2018 Joint Conference; Kingston, Ontario, Canada (2018)
12. Torres, J., Duenas-Osorio, L., Li, Q., Yazdani, A.: Exploring topological effects on water distribution system performance using graph theory and statistical models. Am. Soc. Civil Eng. (2016). [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000709](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000709)
13. Zimmerman, R., Zhu, Q., Dimitric, C.: A network framework for dynamic models of urban food, energy and water systems (FEWS). Environ. Prog. Sustain. Energy **37**(1), 122–131 (2017)
14. Hwang, H., Lansey, K.: Water distribution system classification using system characteristics and graph-theory metrics. Water Resour. Planning Manag. **143**(12), 04017071-1–0401707-13 (2017)
15. Herrera, M., Abraham, E., Stoianov, I.: A Graph-theoretic framework for assessing the resilience of sectorised water distribution networks. Water Resour. Manag., 1685–1699 (2016)
16. Masucci, A., Stanilov, K., Batty, M.: Exploring the evolution of London's street network in the information space: a dual approach. Am. Phys. Soc. E **89** (2014). <https://doi.org/10.1103/physreve.89.012805>
17. Kalapala, V., Sanwalani, V., Clauset, A., Moore, C.: Scale invariance in road networks. Phys. Rev. E **73** (2006). <https://doi.org/10.1103/physreve.73.026130>
18. Porta, S., Crucitti, P., Latora, V.: The network analysis of urban streets: a dual approach. Phys. A **369**, 853–866 (2006)
19. Rosvall, M., Trusina, A., Minnhagen, P., Sneppen, K.: Networks and cities: an information perspective. Phys. Rev. Lett. **94**, 028701 (2005)
20. Boccaletti, A., Latora, V., Morenod, Y., Chavezf, M., Hwang, D.: Complex networks: structure and dynamics. Phys. Rep. **424**, 175–308 (2006)
21. Thrysøe, C., Arnbjerg-Nielsen, K., Borup, M.: Identifying fit-for-purpose lumped surrogate models for large urban drainage systems using GLUE. J. Hydrol. **568**, 517–533 (2018)
22. Olesen, L., Löwe, R., Arnbjerg-Nielsen, K.: Flood damage assessment – literature review and recommended procedure. Cooperative Research Centre for Water Sensitive Cities, Clayton Campus Monash University, Australia (2017)
23. Rossman, L.: Storm water management model user's manual, version 5. Water Supply and Water Resources Division National Risk Management Research Laboratory Cincinnati, OH 45268 (2009)
24. Zhan, X., Ukkusuri, S.: Dynamics of functional failures and recovery in complex road networks (2017). <https://www.researchgate.net/publication/320791885>. Accessed 12 Sep 2018
25. Masucci, A., Smith, D., Crook, A., Batty, M.: Random planar graphs and the London street network. Eur. Phys. J. B **71**, 259–271 (2009)
26. Yang, S., Paik, K., McGrath, G.S., Urich, C., Krueger, E., Kumar, P., Roa, P.S.C.: Functional topology of evolving urban drainage networks. Water Resour. Res. **53**, 8966–8979 (2017)



Mining Behavioural Patterns in Urban Mobility Sequences Using Foursquare Check-in Data from Tokyo

Galina Deeva¹, Johannes De Smedt², Jochen De Weerdt¹,
and María Óskarsdóttir^{3(✉)}

¹ Faculty of Economics and Business, Department of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

² School Management Science and Business Economics Group, University of Edinburgh Business, 29 Buccleuch Place, Edinburgh EH8 9JS, UK

³ Department of Computer Science, Reykjavík University, Menntavegi 1,
101 Reykjavík, Iceland
mariaoskars@ru.is

Abstract. In a study of mobility and urban behaviour, we analyse a longitudinal mobility data set from a sequence mining perspective using a technique that discovers behavioural constraints in sequences of movements between venues. Our contribution is two-fold. First, we propose a methodology to convert aggregated mobility data into insightful patterns. Second, we discover distinctive behavioural patterns in the sequences relative to when in the day they were formed. We analyse sequences of venues as well as sequences of subcategories and categories to discover how people move through Tokyo. The results indicate that our methodology is capable of discovering meaningful behavioural patterns, that can be potentially used to improve urban mobility.

Keywords: Mobility · Urban analysis · Sequence classification · Behavioural constraints · Supervised learning

1 Introduction

Understanding how humans move around their habitats in their daily lives is important for urban planning and commuting, as well as the study of mobility and other urban phenomena [3]. With the advancement of new technologies, tools and data sources, the research in this area is no longer subject to large surveys and long collection times [5, 11]. Instead, the millions of social media users worldwide are volunteering vast amounts of geographic information that is high in quality and usability, thus providing opportunities for research on and design in urban environments. Such digital data sources known as location based social networks—or LBSN—have become an important source for urban and mobility analysis in the last years. In particular, check-in data from location technology platform Foursquare have been used in various studies to discover

universal patterns in mobility, for venue recommendation and sociology, to name a few [7–9].

In this paper, we propose a methodology to convert aggregated data on movements between venues in a city (thus not requiring individual sequences) into insightful patterns. Our approach consists of three stages: (i) conversion of the data into mobility networks, (ii) random walk-based pseudo sequence generation, and (iii) sequence classification and pattern discovery in the form of constraints using the state-of-the-art sequence classification technique called iBCM. Key advantages of our approach include scalability to real-world data sets, accurate and concise pattern discovery useful for urban planning and mobility, and its privacy-friendly nature given that individual sequence data is not required.

We look at a longitudinal mobility data set from a sequence mining perspective. The data set, which originates from Foursquare¹, contains movements of individuals across time in the megacity Tokyo. The movements can be viewed as sequential data where each location is an item in the alphabet of venues and a sequence comprises an ordered list of locations that describe people's behaviour when moving through the city. The sequence mining perspective has been previously adopted in the context of finding common routines and venues that frequently appear together using LDA [4, 10]. However, our approach is different: first, because of a novel methodology for pseudo sequence generation, and second, because we apply a novel and powerful sequence mining technique called the interesting Behavioural Constraint Miner or iBCM [2]. The technique is capable of discovering expressive and concise patterns using behavioural constraint templates, such as simple occurrence, looping and position in a sequence. In addition, the technique can discover the absence of particular behaviour, such as that two items never occur together, which is interesting for the understanding of urban mobility patterns. The output of iBCM is a set of features that represent behavioural constraints in the sequences. Subsequently, supervised analytics techniques can be applied to the feature set to discover meaningful patterns, interesting signals and to classify the sequences.

The goal of this study is to discover distinctive and discriminating behavioural constraints in mobility sequences during different periods of the day. In the context of urban mobility we take a fine-grained look at the venues to see how individuals travel through a city. We apply iBCM to commute sequences to discover particular behaviour in certain time periods. We apply the technique in a supervised setting with the goal of classifying the sequences with respect to the part of day. For urban growth and dynamics, we consider, on the one hand, the subcategory and, on the other hand, the category of the venues. Thereby, we analyse sequences of categories, i.e., whether they are for example Residence, Food, or Travel & Transport, and of subcategories. Interesting behavioural constraint templates in this case could be whether after visiting a museum do people go to a restaurant and whether the restaurant is close to the museum or is public transport needed to get there. For this application, our goal is again to classify the sequences with respect to the part of day and to study the feature sets generated by iBCM to

¹ <https://foursquare.com>.

discover behavioural constraints that are prominent at each time. This look at urban dynamics allows us to study in which context event interactions happen and could also be applied for location intelligence.

Our approach consists of three steps, namely building complex networks of venues and people's movements between them, the generation of pseudo sequences that represent the collective behaviour of the population as it moves through the city, and the subsequent sequence classification and constraint mining. As the data is too aggregated to analyse the mobility sequences of individuals, we extract pseudo sequences from the networks. These pseudo sequences represent the collective movements of the population. In this way we are able to focus on recurrent mobility patterns with common characteristics instead of trying to reproduce likely travel trajectories. [10] Our analyses transition from a fine grained view of movements between venues to an aggregated approach that looks at mobility in the broader sense using categories. This gives alternative perspectives of urban dynamics and behavioural patterns in the context of venue interaction.

In the next section we present the data we used in our study followed by a description of our proposed methodology. In Sect. 4 we present the results of our study and finish with some concluding remarks in Sect. 5.

2 Data Description

We use a mobility data set describing venue interactions in Tokyo. The data was provided for the participants of the Future Cities data challenge². The data contains a list of venues in each city, together with their category and subcategory and a geographic location. There are ten categories and almost five hundred subcategories in total. Moreover, there is information about how users interact with the venues, in the form of movements from one venue to another. When a user checks in at a given venue, an edge is created between the venue where they checked in last—the *from* venue—and the venue where they are currently checked in—the *to* venue. The timing of the check-in is also recorded and assigned to one of five bins, representing the part-of-day, see Table 1. The parts-of-day are morning, midday, afternoon, night and overnight. The data is aggregated at

Table 1. The time periods.

Part-of-day	Start hour	End hour
Morning	6:00	10:00
Midday	10:00	15:00
Afternoon	15:00	19:00
Night	19:00	24:00
Overnight	24:00	6:00

² <https://www.futurecitieschallenge.com/>.

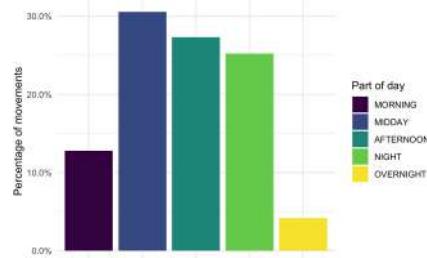


Fig. 1. Relative frequency of movements at different parts-of-day in Tokyo.

a monthly level and therefore each movement also carries a weight indicating the number of check-ins that took place by any user for the given venue pair in the respective month. For our analyses we focus on Tokyo and the months April, May and June 2017. During this three month period, users checked in at a total of 53740 venues, with a monthly average of 47133 venues. There were a total of 1489006 check-ins made in the three months, with an average of 496335 movements per month.

Figure 1 shows the distribution of movements by part-of-day in the three months we considered for Tokyo. Most movements take place at midday, followed by movements in the afternoon and at night. The lowest number of movements happen overnight.

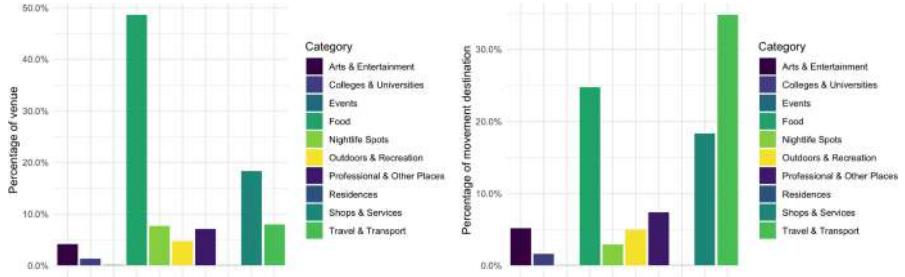
Figure 2 shows the distribution of the venues' categories and the category of the *to* venue of the movements. We can see that most of the venues in Tokyo, which is almost fifty percent, belong to the category Food. Very few of the venues belong to the categories Events and Residences. Looking at the distribution of categories of to movements, we see that most movements end in the Travel & Transport category, followed by Food and Shops & Services categories. Interestingly, even though less than ten percent of the venues belong to Travel & Transport, almost 35% of movements end in that category, and much less movements end in the Food category, even though most venues belong to that category. This reflects that the way people interact with venues does not necessarily depend on the set of venues available for a visit.

3 Methodology

Our proposed methodology consists of three steps which we describe in detail below.

3.1 Complex Network Building

We make complex networks at three levels of granularity: venue (V), subcategory (S) and category (C). Table 2 shows an example of a few venues together with their corresponding subcategory and category.



(a) Relative frequency of the categories of the venues.
 (b) Relative frequency of the categories of the destination venue of the movements.

Fig. 2. Relative frequency of categories for venues and movements.

The venue networks are created using the aggregated movements between venues. To create the subcategory and category networks, we replace the venues by their corresponding subcategory or category, respectively, and aggregate the edges. To distinguish mobility behaviour at different times of the day, we build separate networks for each time period. The time periods are morning (*MO*), midday (*MI*), afternoon (*AF*), night (*NI*) and overnight (*OV*). The definition of the time periods can be seen in Table 1.

Table 2. An example of venues, their subcategory and category.

Venue	Subcategory	Category
Leuven train station	Train Stations	Travel & Transport
Bart's bar	Bars	Nightlife Spots
Eskimo attire	Clothing Stores	Shops & Services
Moskow Meals	Russian Restaurants	Food

We denote the networks using the granularity as a subscript and the part of day as a superscript. For example, \mathcal{N}_V^{MO} is the morning venue network and \mathcal{N}_S^{NI} is the night subcategory network. The venue networks have 58 thousand nodes, the subcategory networks have 456 nodes and the category networks have 10 nodes. All the networks are directed, indicating a movement from one location to another, and weighted by the number of movements in the given time period.

Figures 3 and 4 show category networks for Tokyo in June 2017 for different parts of the day. The first network in Fig. 3 shows people's movements during midday and the second network shows movements in the afternoon. The three networks in Fig. 4 show movements in the morning, night and overnight. Although the networks appear very similar, there are some noticeable differences. For example, at midday Residences have no self loops and there are no movements from Events to Nightlife Spots. In addition, in the afternoon the weights

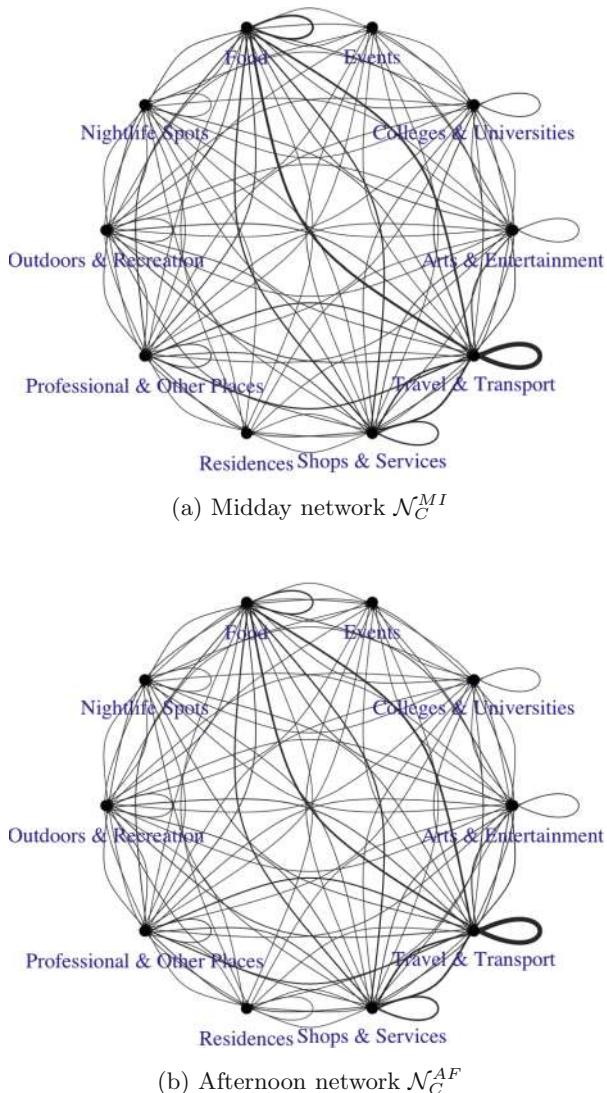


Fig. 3. Tokyo category networks.

on the edges from Food to Travel & Transport and from Travel & Transport to Food are almost the same, whereas at midday the weight on the edge from Travel & Transport to Food is 36% higher than on the edge in the opposite direction. The overnight network is also less densely connected than the other networks as we already observed in Fig. 1.

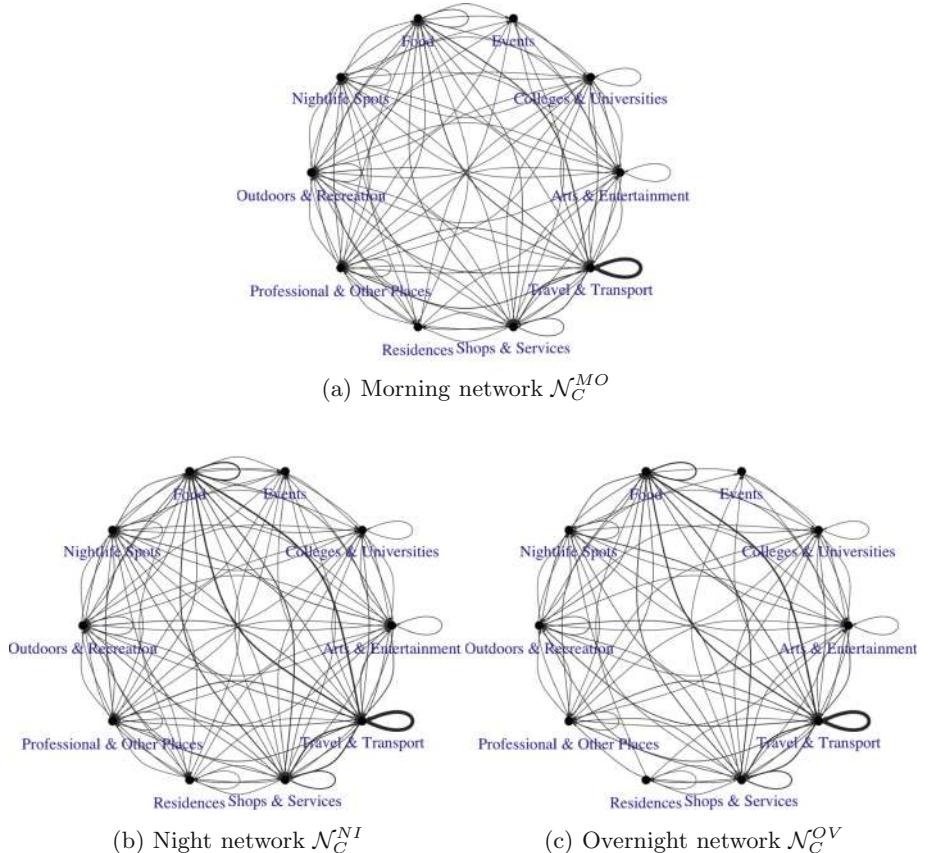


Fig. 4. Tokyo category networks.

3.2 Pseudo Sequence Generation

As we do not have individual mobility sequences, we generate pseudo sequences that portray the collective mobility of the population.

The pseudo sequences are generated by initialising a number of random walks in the networks where edges are weighted by the number of movements. These random walks give a representation of the movements between venues and are used in lieu of sequences. Random walks on networks can be used to extract useful information from networks [6]. In our networks, we know for certain that transitions from A to B and from B to C happen, but we do not know if the order of these transitions is from A to B to C. However, with an assumption that some people have moved more than once around the city, and given that there is a high chance for the location B to be followed by C, we can conclude that some people who moved from A to B continued their way with a movement from B to C. iBCM is fairly robust to this assumption, given that its patterns represent

relations between two items at most. Furthermore, as we generate a considerable number of such random walks, we capture the variability in movements with more common behaviour appearing more frequently in the sequences. Thus, with such pseudo sequences we get an approximation of the average movements which allow us to look at recurrent mobility patterns with common characteristics. We generate random walks with 10, 20 and 30 steps. We chose these lengths as they provide a range that we can test in our experiments. By intuition, someone probably does not visit several venues in the span of a few hours and therefore we chose 10 as a lower bound, however, the sequence mining classification technique might benefit from longer walks when trying to discover recurrent patterns.

From the venue networks \mathcal{N}_V we get venue sequences \mathcal{S}_V in which we replace the venues with their corresponding subcategory and category to obtain the sequences $\mathcal{S}_{V \rightarrow S}$ and $\mathcal{S}_{V \rightarrow C}$. From the subcategory networks \mathcal{N}_S we obtain subcategory sequences \mathcal{S}_S and we replace the subcategories with their corresponding category to obtain the category sequences $\mathcal{S}_{S \rightarrow C}$. Finally, from the category networks \mathcal{N}_C we obtain the category sequences \mathcal{S}_C .

A pseudo sequence for a venue network with the venues in Table 2 could for example be

Leuven train station → Eskimo attire → Moskow Meal → Leuven train station $\in \mathcal{S}_V$

with the corresponding subcategory and category sequences

Train Stations → Clothing Stores → Russian Restaurants → Train Stations $\in \mathcal{S}_{V \rightarrow S}$

Travel & Transport → Shops & Services → Food → Travel & Transport $\in \mathcal{S}_{V \rightarrow C}$.

The first sequence is very detailed since it shows movements between exact locations. As there are 58 thousand venues there is a lot of variability in these sequences. The venue networks furthermore have multiple small components of only one or two nodes, which means that these venues are never together in a sequence with the majority of venues. The second sequence on the other hand shows a more eminent mobility pattern and the same holds for the third sequence, where the pattern is even more high-level.

The sequences are labelled according to which part-of-day network they come from.

3.3 Sequence Classification and Constraint Discovery

iBCM discovers discriminative patterns from sequential data [2]. The behavioural constraints mined by iBCM are based on the Declare language [12]. First, the algorithm generates sequential patterns separately for each class of sequences. Next, each sequence is transformed to a feature vector with binary features indicating whether a certain pattern between two items is present in the sequence. Subsequently, a predictive model is built based on these features.

We look into five classes of sequences that represent the five part-of-day periods described above (morning (*MO*), midday (*MI*), afternoon (*AF*), night

(*NI*) and overnight (*OV*)), see Table 1. iBCM derives binary features from the sequences. First, it splits the sequences in a particular number of windows, for which the patterns are mined, which are subsequently used in a classification algorithm. The search for patterns in different windows provides more specific information about when particular behaviour occurs. This ability of iBCM, however, is not applicable to data used in this study, as it doesn't contain exact timestamps of the movements. Therefore, we do not vary the window parameter and do not take the location in the sequences into account for classification purposes.

Next to the number of windows, another parameter to be varied is the support level (the percentage of sequences that contain a certain sequential pattern). We vary support from 0.4 to 0.8 with 0.1 intervals, thus obtaining 5 different feature sets for each data set. The discovered patterns are then fed into the random forest algorithm, chosen because it performs well with binary features and is often used in sequence classification [1].

4 Results

Table 3 presents an overview of the accuracy and lift results produced by iBCM and random forest, as well as the number of behavioural constraints generated for different pseudo sequences. The results are calculated using 10-fold cross validation. For each data set, we only present the results for the support level that yielded the best accuracy.

It can be observed that the accuracy results obtained by iBCM and random forest are relatively high, ranging from 0.66 to 1.0. Such high accuracy obtained in multi-class classification confirms that iBCM was able to find patterns that were discriminating between the classes.

Some examples of the observed patterns for different sequences are listed in Tables 4 and 5, which illustrate behaviour in April and May 2017. For example, the pattern Existence3(Travel & Transport) in Table 4 indicates that Travel & Transport item occurred at least three times during a certain window in the morning, which confirms the intuition that people tend to use several types of transportation in a sequence to commute to work in the morning. Similarly, we can conclude that at lunchbreak people tend to use transport to get to a restaurant, that they are less likely to visit Colleges & Universities in the afternoon, they go from shops to the train station in the evening, and, finally, that they will to a lesser extent visit restaurants during the night.

Similar patterns occur in Table 5. We can observe that people tend to visit some food places in the morning, followed by commuting to work. Then, during lunchbreak, some people might use transportation to visit Shops & Services. Finally, it can be concluded, that the restaurants don't get visits in the afternoon, Colleges & Universities are not visited in the evening, and Shops & Services are not visited at night.

In terms of the influence of various parameters on classification results, we make the following observations:

Table 3. Classification results for pseudo sequences generated from networks with different parameters.

Sequence	Length of walk	Month	Support	# Constraints	Accuracy	Lift
$\mathcal{S}_{C \rightarrow C}$	10	April	0.5	24	1.0	4.9
$\mathcal{S}_{S \rightarrow C}$	10	April	0.7	23	0.99	4.51
$\mathcal{S}_{C \rightarrow C}$	20	April	0.7	20	0.99	4.9
$\mathcal{S}_{S \rightarrow C}$	20	April	0.8	22	0.87	4.54
$\mathcal{S}_{V \rightarrow C}$	20	April	0.4	36	0.96	3.44
$\mathcal{S}_{S \rightarrow S}$	10	April	0.6	14	0.67	2.74
$\mathcal{S}_{V \rightarrow S}$	10	April	0.5	10	0.74	3.09
$\mathcal{S}_{S \rightarrow S}$	20	April	0.5	31	1.0	4.54
$\mathcal{S}_{V \rightarrow S}$	20	April	0.4	15	0.85	2.30
$\mathcal{S}_{C \rightarrow C}$	10	May	0.7	24	0.86	3.35
$\mathcal{S}_{S \rightarrow C}$	10	May	0.7	24	0.87	3.24
$\mathcal{S}_{V \rightarrow C}$	10	May	0.4	28	0.97	3.4
$\mathcal{S}_{C \rightarrow C}$	20	May	0.8	23	0.99	4.99
$\mathcal{S}_{S \rightarrow C}$	20	May	0.8	23	0.87	4.59
$\mathcal{S}_{V \rightarrow C}$	20	May	0.4	41	0.99	3.39
$\mathcal{S}_{S \rightarrow S}$	10	May	0.6	14	0.66	2.67
$\mathcal{S}_{V \rightarrow S}$	10	May	0.5	10	0.71	1.85
$\mathcal{S}_{S \rightarrow S}$	20	May	0.5	41	0.99	4.59
$\mathcal{S}_{V \rightarrow S}$	20	May	0.4	20	0.85	2.3
$\mathcal{S}_{C \rightarrow C}$	10	June	0.5	24	0.85	4.99
$\mathcal{S}_{S \rightarrow C}$	10	June	0.6	19	0.99	4.53
$\mathcal{S}_{V \rightarrow C}$	10	June	0.4	28	0.97	3.44
$\mathcal{S}_{C \rightarrow C}$	20	June	0.8	22	0.84	4.99
$\mathcal{S}_{S \rightarrow C}$	20	June	0.8	23	0.87	4.54
$\mathcal{S}_{V \rightarrow C}$	20	June	0.4	36	0.99	3.44
$\mathcal{S}_{S \rightarrow S}$	10	June	0.6	19	0.83	4.54
$\mathcal{S}_{V \rightarrow S}$	10	June	0.5	12	0.74	3.09
$\mathcal{S}_{S \rightarrow S}$	20	June	0.5	26	0.99	4.54
$\mathcal{S}_{V \rightarrow S}$	20	June	0.4	18	0.87	2.48

- We generate sequences of three possible lengths: 10, 20 and 30 items. For sequences with length 30 no features were obtained, which is in line with the intuition that no meaningful sequences of so many movements could be possible within one time period. The results for lengths 10 and 20 are similar, however, there is a slight tendency for the accuracy to be higher for longer

Table 4. Examples of constraints derived by iBCM for $\mathcal{S}_{V \rightarrow C}$ sequences of length 20 in May.

Label	Constraint
MO	Existence3(Travel & Transport)
MI	CoExistence(Travel & Transport, Food)
AN	Absence(Colleges & universities)
NI	CoExistence(Shops & Services, Travel & Transport)
OV	Absence(Food)

Table 5. Examples of constraints derived by iBCM for $\mathcal{S}_{S \rightarrow C}$ sequences of length 20 in April.

Label	Constraint
MO	CoExistence(Food, Travel & Transport)
MI	CoExistence(Travel & Transport, Shops & Services)
AN	Absence(Food)
NI	Absence(Colleges & universities)
OV	Absence(Shops & Services)

sequences with the average accuracy of 0.85 and 0.93, respectively. This confirms the expectation that longer sequences could potentially contain more information; however, in the context of this study, there is a clear limitation to a length of pseudo sequences that can still be realistic.

- The sequences with three different levels of granularity are analysed: category, subcategory and venues. The latter didn't yield any results, because the alphabet of possible venues was simply too large (58 thousands), making it challenging to search for patterns in short sequences of length 10–20. The average accuracy results for subcategory and category are 0.83 and 0.93, respectively. Thus, it is possible that more interesting patterns can be extracted when looking at the movements between venues from a more general perspective. However, more experiments are needed to confirm this observation.
- Finally, there is no clear tendency for a certain type of network generation, i.e., with category, subcategory or venue, to produce better results.

5 Conclusion

In this paper, we used a sequence classification technique to discover behavioural patterns that are distinctive in mobility patterns during different parts of the day. The high accuracy levels that we obtained show that the pseudo sequences that we generated from the venue, subcategory and category networks contain constraints that are distinct and discriminative for the various parts of the day. The technique is fairly good at detecting these constraints, which in addition

are intuitive. They provide an understanding of urban mobility and the specific constraints in sequences from each class could be used to improve transport and accessibility in the city. Our proposed methodology is capable of discovering accurate and concise patterns that are useful for mobility and urban planning and is furthermore privacy friendly since individual movements are not required.

For future work, we would like to carry out the same analysis for the other cities that were provided to see if the mobility patterns are universal. If they are not, we could apply the same technique to classify the cities and discover patterns that are unique for each city. In addition, it would be interesting to augment the sequences with spatial information, and thus take into account the distance between two check-in venues and whether they are in the same neighbourhood or not. The main drawback of our research is the absence of actual mobility sequences, and therefore we resort to pseudo-sequences. We would have liked to work with real sequences for our analyses, but the data was aggregated for privacy reasons. However, the pseudo sequences provided a representation of movements between venues that were both logical and plausible in the context of urban mobility throughout the day. It would be very interesting to apply iBCM to individual mobility sequences to see how well they are approximated by the pseudo sequences.

References

1. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
2. De Smedt, J., Deeva, G., De Weerdt, J.: Mining behavioral sequence constraints for classification. *IEEE Trans. Knowl. Data Eng.* (2019)
3. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. *Nature* **453**(7196), 779 (2008)
4. Long, X., Jin, L., Joshi, J.: Exploring trajectory-driven local geographic topics in foursquare. In: Proceedings of the 2012 ACM conference on ubiquitous computing, pp. 927–934. ACM (2012)
5. Martí, P., Serrano-Estrada, L., Nolasco-Cirugeda, A.: Social media data: challenges, opportunities and limitations in urban studies. *Comput. Environ. Urban Syst.* **74**, 161–174 (2019)
6. Masuda, N., Porter, M.A., Lambiotte, R.: Random walks and diffusion on networks. *Phys. Rep.* **716**, 1–58 (2017)
7. Noë, N., Whitaker, R.M., Chorley, M.J., Pollet, T.V.: Birds of a feather locate together? Foursquare checkins and personality homophily. *Comput. Hum. Behav.* **58**, 343–353 (2016)
8. Noulas, A., Scellato, S., Lathia, N., Mascolo, C.: A random walk around the city: new venue recommendation in location-based social networks. In: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, pp. 144–153. IEEE (2012)
9. Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., Mascolo, C.: A tale of many cities: universal patterns in human urban mobility. *PloS One* **7**(5), e37027 (2012)
10. Pianese, F., An, X., Kawsar, F., Ishizuka, H.: Discovering and predicting user routines by differential analysis of social network traces. In: 2013 IEEE 14th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–9. IEEE (2013)

11. Yue, Y., Lan, T., Yeh, A.G., Li, Q.Q.: Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. *Travel. Behav. Soc.* **1**(2), 69–78 (2014)
12. Pesic, M., Schonenberg, H., Van der Aalst, W.M.: Declare: full support for loosely-structured processes. In: 11th IEEE International Enterprise Distributed Object Computing Conference, p. 287. IEEE, October 2007



Temporal Analysis of a Bus Transit Network

Manju Manohar Manjalavil¹ , Gitakrishnan Ramadurai^{1,2} ,
and Balaraman Ravindran^{1,2}

¹ Indian Institute of Technology Madras, Chennai, Tamil Nadu, India
manjum113@gmail.com

² Robert Bosch Centre for Data Science and Artificial Intelligence,
Indian Institute of Technology Madras, Chennai, Tamil Nadu, India

Abstract. Transit networks are essentially temporal with their topology evolving over time. While there are several studies on the topological properties of bus transit networks, none of them have captured the temporal network characteristics. We present a temporal analysis of a bus transit network using snapshot representation. We propose a supply-based weight measure, called the service utilization factor (SUF), and define it as the passenger demand per trip between two bus stops. We evaluate the complex network properties in three weighted cases for a bus network in India, using the number of overlapping routes, passenger demand between routes and SUF as weights. The study network is well-connected with 1.48 number of transfers on average to travel between any two stops over the day. The temporal analysis indicated an inadequate number of services in peak periods and route redundancy across the time periods. We identified the existing and potential hubs in the network, which were found to vary across time periods. The network has strongly connected communities that remain constant across the day. Our conclusions exemplify the importance of temporally analyzing transit networks for improving their efficiency.

Keywords: Bus network · Transit · Snapshot · Temporal · Weighted network · Demand · Service utilization

1 Introduction

Transportation systems are the lifelines of a society, enabling people to take part in various socio-economic activities. Compared to private vehicles, transit systems are more economical, safer and have additional benefits such as fuel savings and reduced carbon footprint. An efficient transit system can attract more personal vehicle users, reducing congestion and pollution. The design of a public transit system is influenced by the geographic, social, and economic conditions in the city. These systems evolve as the city grows and appear with complex structural and dynamical features.

Complex network theory can be used to analyze the transit systems to quantify their topological characteristics and dynamical mechanisms. A transit system can be modeled as a network with stations/stops as nodes and the route connecting any two nodes as edge. Transit networks are geographically constrained by the presence of roads/lines and are smaller in size compared to the virtual networks. Transit systems that have been

analyzed using network theory include the railway [1–3], airline [4–6], subway [7, 8] and bus networks [3, 9–12]. These studies have modeled the transit systems as static networks. In a static network, an edge between two stations exists throughout the observation period. However, in real-world transit systems, presence of an edge between two nodes depends on whether a route connects the corresponding stations in that particular time period. A temporal network is a network in which the topology changes over time. An edge will connect two nodes at a time instant only if there is an ‘interaction’ between those nodes at that instant.

In this paper, we model the bus network of an Indian city to understand how the temporal variations in the bus schedules and passenger flow patterns influence the structure and function of the network. We use the snapshot representation in which the observation period is divided into equal time-windows and all the nodes and edges present in each time window is aggregated as a static network or snapshot for that period. Snapshots have been previously used to analyze communication [12–14] and airline networks [15]. We analyze the topological properties of each snapshot using standard static metrics such as degree, path lengths and centralities [3, 10]. We perform comparative analysis of these snapshots over the day. Even though such a representation cannot fully capture the dynamic characteristics of the network, we can understand how the network behavior changes over the day.

The rest of this paper is organized as follows. In Sect. 2, we describe the data, the network representations, and edge weights. In the next section, we present the results and discuss their implications. In the last section, we summarize the work and briefly discuss future research directions.

2 Methods

2.1 Data

We construct the network using passenger ticketing data. In the bus system under study, most of the tickets are recorded using Electronic Ticketing Machines (ETMs). The ETMs store each ticket as a separate record. Each record has information on the route characteristics such as route number, origin and destination, trip characteristics such as schedule details, trip start and end times and boarding and alighting stage of each passenger, ticket details such as the type, ticket issued time and amount collected, along with other vehicle and crew related details. For this study, we extracted data for a single weekday (Wednesday, 1st November 2017). Due to the absence of working ETMs in all the buses, the data was available for only about 67% of the scheduled trips. Using a projection factor based on the number of scheduled trips and the number of recorded trips in each route, we project the available passenger information for the missing trips.

The ETMs record only the boarding and alighting stage of each passenger where a stage typically includes 2–3 bus stops located within an approximately 2 km stretch. Hence, we map the passenger demand from stage to bus stop level using a points of interest (POI) based mapping procedure [16]. The boarding/alighting information is

not recorded for concession tickets (student and senior citizen passes and daily/weekly/monthly passes). Hence, such passenger trips are not included in the current study.

2.2 Temporal Network

Let t be the time window over which the topology of the network remains unchanged. $G_t = (N_t, E_t)$ represents a snapshot at time t , with N_t nodes and E_t edges. The value of t may vary from a few minutes to hours, days or years, depending on the time period of observation chosen. A_t defines the adjacency matrix corresponding to G_t . A series of snapshots G_1, G_2, \dots, G_T can be used to represent the temporal network over the observation period T .

In the current study, we fix the time window of the snapshots as 1 h. The analysis period is taken as 04:00–23:00 corresponding to the regular bus services, resulting in 19 snapshot networks. A trip is a single run of a bus from origin to destination. We group the trips into different snapshots based on their start times. Let T_S and T_E be the start and end times of a snapshot time window. Then a trip that starts at time t_s belongs to this snapshot if,

$$T_S \leq t_s < T_E \quad (1)$$

We analyze each snapshot network using L-space and P-space representations [10]. In L-space representation, an edge exists between two nodes if there is at least one bus route servicing them consecutively. In P-space, two nodes are connected by an edge if there is at least one common route servicing them. In each space representation, we evaluate the topological properties of an unweighted and three weighted representations of the network. The role of geographical constraints in shaping the transit network are captured using an unweighted network. The first edge weight is route overlap and it is used to understand the influence of overlapping routes along the edges. The next weight is passenger demand, the pattern of which will influence the bus schedules and thereby, the network. In the third weighted network, newly proposed ‘Service Utilization Factor’ (SUF) is the edge weight. We define SUF as the average demand per trip between two stops and measure it using Eq. 2,

$$(SUF)_{ij} = \frac{D_{ij}}{Z_{ij}} \quad (2)$$

where D_{ij} is the demand between stops i and j and Z_{ij} is the number of trips connecting i and j . In L-space, D_{ij} is the number of people in a bus as it travels between the two stops. D_{ij} will include passengers travelling between other stops as well, who have to pass through the considered stops. In P-space, D_{ij} is the actual number of people travelling from stop i to stop j . We use the passenger boarding and alighting data at bus stop level to calculate D_{ij} . The SUF weighted network reveals how the availability of services vis-à-vis passenger demand would impact the network characteristics. The edges are assumed to be undirected in all the representations.

3 Results and Discussions

We analyze the topological properties of the snapshot networks corresponding to the different time periods. Each network is modeled in L-space and P-space, under the different edge weights viz. route overlaps, passenger demand, and service utilization factor. Figure 1 shows that even the basic network features such as number of nodes and edges vary with time. The regular bus services start after 4:00 and almost all the routes are functioning by 6:00, as indicated by the steep growth in the plot. Most of the services end by 22:00, resulting in the fast decline in the number of nodes and edges. From 22:00, a small number of night service buses run between major centers.

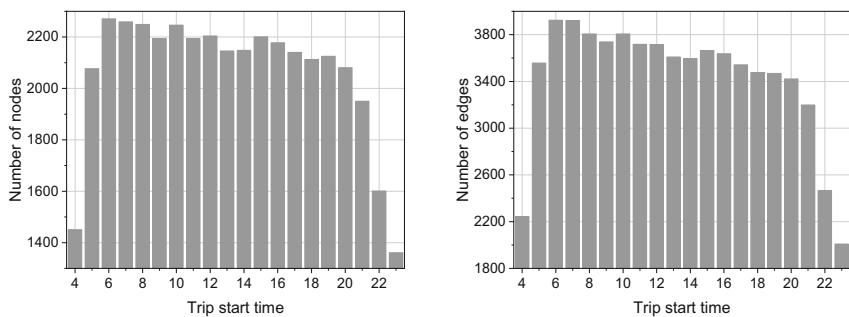


Fig. 1. Number of nodes and edges in the network corresponding to each time period.

We plotted the bus route networks corresponding to four different time periods using the geo-layout algorithm in Gephi 0.9.2 [17]. Figure 2 presents how the network structure changes in different time periods.

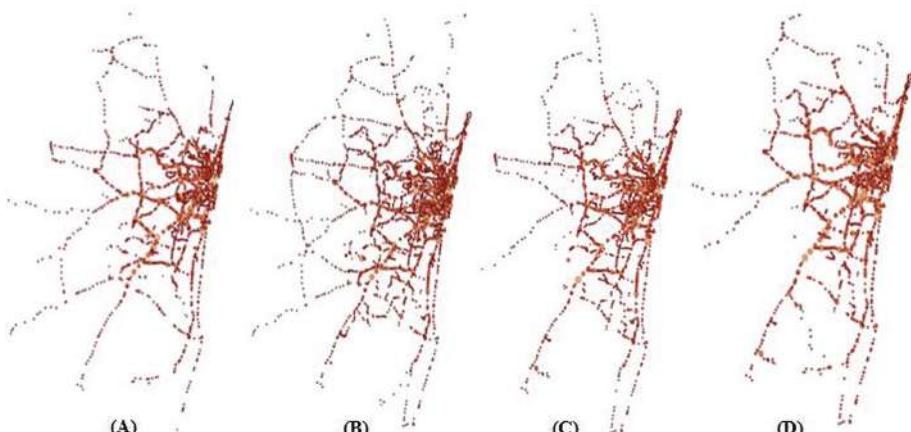


Fig. 2. The bus transit network corresponding to different time periods. (A) 4:00–5:00, (B) 8:00–9:00, (C) 21:00–22:00 and (D) 23:00–4:00.

Visualization of the network gives a broad idea about the areas that are serviced at different time periods. The regular services start around 4:00–5:00 (Fig. 2(A)), with routes spread along all major corridors. The network in Fig. 2(B) corresponding to morning peak period of 8:00–9:00 has almost all the servicing routes, as observed from the greater spread of the network and higher density in the central region. The services start dwindling by 21:00–22:00, as observed from the comparatively less-spread network in Fig. 2(C). Figure 2(D) presents the night service routes functioning from 23:00–4:00.

We identified the communities within the networks using the Gephi implementation of the ‘Fast unfolding of communities in large networks’ [18] algorithm. The modularity value varied within 0.85–0.87 across the time periods. The high modularity values indicate strongly connected communities which have weak connections with other communities. The communities indicate areas serviced by the same set of routes and they remained constant over the periods.

We discuss the topological properties evaluated and significant results in the following sub-sections.

3.1 Degree

Degree of a node is the total number of neighbors it is connected with. In L-space, the average node degree in the static network is 3.58, while it varies between 3.0 to 3.5 in the temporal networks. These values indicate good connectivity within the network with bus routes connecting every bus stop directly to three other bus stops on an average. Figure 3(a) presents the node degree variation across time periods, with higher average degree during the morning peak hours (6:00–8:00). The marginal fall in average degree over the day indicates that more services are run in the morning peak periods. Most offices and schools in the city start by 9:00–10:00. Hence, the morning peak traffic is concentrated within a short time-span, while the evening peaks are more dispersed with schools and offices closing at different time periods.

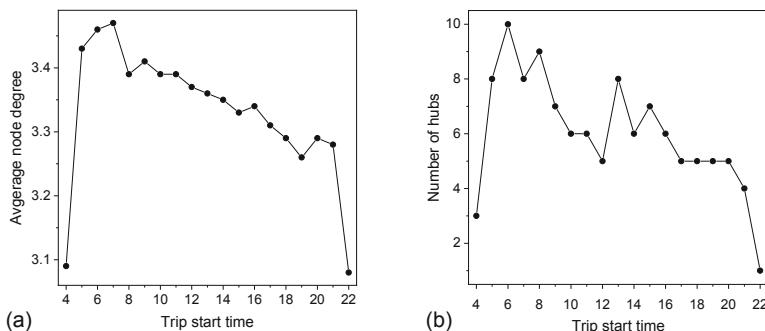


Fig. 3. (a) Average node degree and (b) number of hubs, corresponding to different periods.

In P-space, node degree indicates the number of bus stops reachable from each bus stop, without making transfers. The average node degree in static network is 97, while

it varies between 71 to 92 in the temporal networks. However, most of the node degrees were lower than the average values. The high average values can be attributed to the presence of a few high degree nodes. These nodes are the hubs/terminals, from which routes ply in different directions. For the current network, we define hubs as nodes with average degree greater than 500. Figure 3(b) presents the number of hubs in each time period. The number of hubs is also higher in the morning peak periods, indicating a higher number of services in the morning peak.

We plotted the degree distribution on a double logarithmic scale for both L-space and P-space networks. The degree distributions in L-space varied exponentially, while in P-space, they followed a power law. However, the exponent values varied without any specific trend over the day.

3.2 Strength

Strength of a node is the sum of the edge weights incident on it. The node strength values in the L-space demand weighted network (Fig. 4(a)) gives the passenger demand profile. The demand in the morning peak period is higher than the evening peak period. The evening peak demand is distributed over a larger time period. The strength curve is nearly symmetrical in the L-space SUF weighted network (Fig. 4(b)). The higher strength values corresponding to the peak periods indicate an inadequate number of services in peak periods.

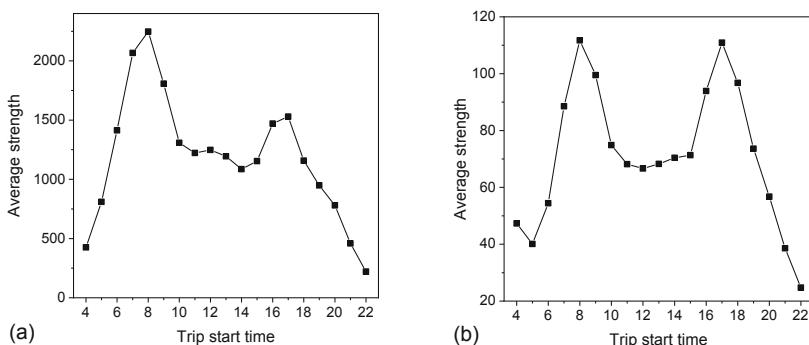


Fig. 4. Average strength values for (a) passenger demand and (b) SUF weighted networks, corresponding to the time periods.

The node strengths are dispersed in all the three weighted L-space networks. Figure 5 shows the coefficient of variation (COV) of strength values. The higher COV variation in the demand weighted network indicates the demand fluctuations in the bus stops over the day. The increase in the COV values at 22:00 in the SUF weighted network indicates the discrepancies in the scheduling of night services. The services are more in routes with low demand and inadequate in routes with higher demand.

Redundant routes are routes in which the number of trips is more than that required to service its demand. For each time period, we select the edges in the L-space network with low SUF values and identify the routes along them. The range of SUF values varied between time periods and hence, we fix the ‘low SUF’ threshold as the value which splits the edge set in half. Table 1 gives the classification of redundancy, using which we assign redundancy level to each route. The percentage of routes with ‘high’ redundancy varied between 0.72%–5.69% in different time periods, while the corresponding value was 4.4% in the static network. The number of trips along highly redundant routes are to be reduced. However, no route is highly redundant across all time periods. We need to identify the time period in which each route is highly redundant and accordingly re-schedule the services. Figure 6 shows the redundant routes in static and three different snapshot networks.

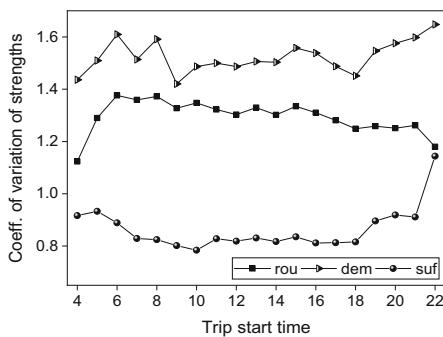


Fig. 5. Coefficient of variation in strength values corresponding to different time periods.

Table 1. Classification of redundancy in routes

Redundancy	Low	Medium	High
Number of edges with low SUF	<=10	10–30	<=30



Fig. 6. Redundant routes in different networks: (a) Static network, (b) 9:00–10:00, (c) 13:00–14:00 and (d) 17:00–18:00.

3.3 Clustering Coefficient

The clustering coefficient evaluates the connectivity among the neighbors of each node. The weighted clustering coefficient is evaluated using Barrat's equation [19]. The clustering coefficients for all the networks in L-space are in the range 0.19–0.24. Low coefficient values indicate lesser number of inter-connections between the routes. Figure 7(a) shows that the coefficient values vary similarly in the unweighted and weighted cases. The edge weights are not correlated with clustering. In P-space, the clustering coefficient values are in the range 0.7–0.8 for all networks. The interconnection between all nodes along a route in P-space leads to the higher coefficient values. In both space representations, coefficient values are lower at 4:00 and after 20:00. The fall in number of services during these periods lead to the fall in interconnected routes.

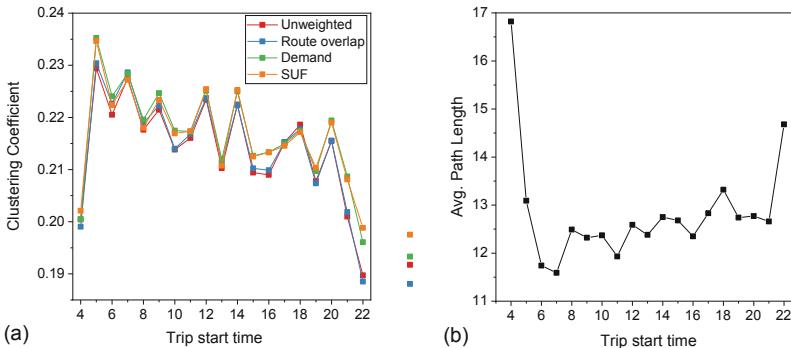


Fig. 7. Clustering coefficient values in each network and (b) path lengths in L-space, corresponding to different time periods.

3.4 Characteristic Path Length

Characteristic path length measures the average number of nodes in the shortest path for all possible node pairs in the network. In L-space, it is the number of bus stops to be passed while travelling between any two bus stops. The path lengths in L-space vary from 11.6 to 16.8 over the day, while the corresponding value in the static network is 11. Figure 7(b) presents the characteristic path length variation across the day. The path lengths are shortest in peak periods due to the presence of higher number of services.

In P-space, characteristic path length quantifies the average number of transfers or route changes required while traveling between any two stops. The path lengths in P-space ranged between 2.35–2.48, indicating that on an average 1.48 transfers are required to travel from a stop to any other stop in the network, irrespective of time of travel.

3.5 Betweenness Centrality

Betweenness centrality quantifies importance of a node based on its connectivity. A more central node will mediate a higher number of shortest paths in the network. Nodes with higher betweenness centrality (C_B) enable easier transfers and thereby faster travel between stops that are not connected by direct routes. Such nodes are potential candidates to be developed as hub. We evaluated the C_B value of all the nodes in L-space, in each time period. For each node, we identified the number of time periods in which they have a high C_B value. The stops with high betweenness centrality across all time periods are potential hubs. We compared these stops with corresponding stops identified from static network based on higher C_B values. Figure 8 presents the potential hubs. The 7 green nodes are potential hubs identified from the static network, the 6 blue nodes are potential hubs identified from temporal networks, and the 13 red nodes are potential hubs common in both static and temporal cases. Few of the hubs (green) identified from the static network were absent in temporal case, implying that they are not central over the whole day. Also, a few new stops (blue) that are central over the whole day were identified from the temporal analysis. Since the hubs should act as efficient transfer points across the day, it is important to identify them using a temporal analysis.



Fig. 8. Potential bus stops that can be developed into hubs.

4 Conclusion

We have presented a temporal analysis of a bus transit network, in two space representations and under four different edge weights. The snapshot networks revealed how the structure of the network is varying over time, influenced by the routes servicing in each time-period. The average degree across time periods indicate a network with good connectivity over the day. Hubs/terminals are major bus stops which act as the starting point of a number of routes and are integral to the network, as they handle a large volume of passengers. Most of the hubs are present in all the snapshots, while some were prominent in certain time periods only.

We captured the demand profile in the network using the strength values in the demand weighted network. The strength values in the SUF weighted network indicated inadequate number of services in the peak periods. More detailed investigation using the SUF strengths revealed the presence of redundant routes in the network. These are routes along which the number of services is high in proportion to the demand along them. The redundant routes were found to be different across time periods. The observations from SUF strengths advocate a need to re-analyze the service schedules, considering the passenger demand along each route, in different time periods.

Large clustering coefficient and small characteristic pathlengths across time periods are indicative of small world network property in P-space networks. Low route interconnections and small number of transfers across time periods indicate a well-connected route network. We identified the potential hubs in the network using the betweenness centrality measure. We compared these hubs with the potential hubs identified from static network. Some of the stops were common in both cases, while some stops were not central across the day. While considering stops for bus stops improvement plans, it is important to give a higher priority to stops with high centrality value across time periods. The degree-degree correlations and community formations in the network were found to be similar across time periods.

The temporal analysis has given a better understanding of the transit network characteristics. However, since the snapshots aggregate the characteristics within a time window, there is a loss of temporal information. We plan to use a more disaggregate approach to temporally analyze the network. Further, we have analyzed the network over a single day. It will be interesting to see how the network changes over multiple days and seasons.

Acknowledgment. The authors acknowledge the support from Robert Bosch Centre for Data Science and Artificial Intelligence (RBC-DSAI) at IIT Madras.

References

1. Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P., Mukherjee, G., Manna, S.: Small-world properties of the indian railway network. *Phys. Rev. E* **67**(3), 036106 (2003)
2. Kurant, M., Thiran, P.: Extraction and analysis of traffic and topologies of transportation networks. *Phys. Rev. E* **74**(3), 036114 (2006)

3. Soh, H., Lim, S., Zhang, T., Fu, X., Lee, G.K.K., Hung, T.G.G., Di, P., Prakasam, S., Wong, L.: Weighted complex network analysis of travel routes on the Singapore Public Transportation System. *Phys. A Stat. Mech. Appl.* **389**(24), 5852–5863 (2010)
4. Bagler, G.: Analysis of the airport network of India as a complex weighted network. *Phys. A Stat. Mech. Appl.* **387**(12), 2972–2980 (2008)
5. Guimera, R., Amaral, L.A.N.: Modeling the world-wide airport network. *Eur. Phys. J. B Condens. Matter Complex Syst.* **38**(2), 381–385 (2004)
6. Li, W., Cai, X.: Statistical analysis of airport network of China. *Phys. Rev. E* **69**(4), 046106 (2004)
7. Latora, V., Marchiori, M.: Is the Boston subway a small-world network? *Phys. A Stat. Mech. Appl.* **314**(1–4), 109–113 (2002)
8. Seaton, K.A., Hackett, L.M.: Stations, trains and small-world networks. *Phys. A Stat. Mech. Appl.* **339**(3–4), 635–644 (2004)
9. Chatterjee, A., Manohar, M., Ramadurai, G.: Statistical analysis of bus networks in India. *PLoS ONE* **11**(12), e0168478 (2016)
10. Xu, X., Hu, J., Liu, F., Liu, L.: Scaling and correlations in three bus transport networks of China. *Phys. A Stat. Mech. Appl.* **374**(1), 441–448 (2007)
11. Chen, Y.Z., Li, N., He, D.R.: A study on some urban bus transport networks. *Phys. A Stat. Mech. Appl.* **376**, 747–754 (2007)
12. Uddin, S., Piraveenan, M., Chung, K.S.K., Hossain, L.: Topological analysis of longitudinal networks. In: 2013 46th Hawaii International Conference on System Sciences, pp. 3931–3940. IEEE (2013)
13. Braha, D., Bar-Yam, Y.: Time-dependent complex networks: dynamic centrality, dynamic motifs, and cycles of social interaction. In: Gross, T., Sayama, H. (eds.) *Adaptive networks: Theory, models and applications*, pp. 38–50. Springer, Heidelberg (2008)
14. Tang, J., Musolesi, M., Mascolo, C., Latora, V., Nicosia, V.: Analysing information flows and key mediators through temporal centrality metrics. In: Proceedings of the 3rd workshop on Social Network Systems, p. 3. ACM (2010)
15. Mou, J., Liu, C., Chen, S., Huang, G., Lu, X.: Temporal characteristics of the Chinese aviation network and their effects on the spread of infectious diseases. *Sci. Rep.* **7**(1), 1275 (2017)
16. Manjalavil, M.M., Ramaduari, G.: Topological properties of bus transit networks considering demand and service utilization weight measures (Manuscript submitted for publication)
17. Gephi, an open source graph visualization and manipulation software. <https://gephi.org/>
18. Blonde, V.D., Guillaume, J.L., Lambiotte, R., Mech, E.L.J.S.: Fast unfolding communities in large networks. *J. Stat. Mech: Theory Exp.* **2008**(10), P10008 (2008)
19. Barrat, A., Weigt, M.: On the properties of small-world network models. *Eur. Phys. J. B Condens. Matter Complex Syst.* **13**(3), 547–560 (2000)



Modeling Urban Mobility Networks Using Constrained Labeled Sequences

Stephen Eubank^(✉), Madhav Marathe^(✉), Henning Mortveit^(✉),
and Anil Vullikanti^(✉)

Biocomplexity Institute, University of Virginia, Charlottesville, USA
{eubank,marathe,Henning.Mortveit,vsakumar}@virginia.edu

Abstract. Models of urban mobility patterns are key inputs to urban analytics, e.g., planning the transportation infrastructure. Typically, only limited data about mobility is available, due to privacy and data collection challenges. Here, we study the problem of reconstructing mobility traces for all individuals from flow measurements on transportation network links. We formalize this as the *constrained label sequence problem* (denoted by CLS). CLS is the problem of constructing labeled sequences with constraints that correspond to flows. The proposed formulation is quite general and allows us to consider bounds on the total flow through a set of links, instead of an individual link.

We show that CLS is computationally hard to solve exactly, and design an algorithm with rigorous approximation guarantees. We also study the complexity of counting and sampling from the set of mobility traces consistent with such measurements. Finally, we demonstrate our results using transportation flows across zones around Washington DC and data from the American Community Survey Commuting Flows.

1 Introduction

There has been a lot of work on developing detailed models of human mobility in urban regions for transportation infrastructure analysis, vehicular ad hoc networks, and urban planning, e.g., [6, 12, 13]. A mobility trace for an individual is the specific sequence of *locations* that person visits, the time of arrival and departure at the location and the activity performed (e.g. shopping, exercise, lunch). Such mobility models serve as inputs to problems arising in urban analytics, public health epidemiology and transport planning; see [1, 2, 4, 11] for a detailed discussion.

Typically it is only possible to get a small sample or certain aggregate statistics about individual mobility patterns for reasons such privacy constraints. Examples of such data include: flow measurements on links in transportation networks, smart card transactions on public transit, and mobility traces for a small subset of people at a coarse geographical resolution. A well studied source of samples of mobility traces is anonymized call data records (CDR), that give base station level data for each anonymized user [1, 5, 8, 11]. In this setting, the

movement of a user is represented as a sequence of base station IDs. Analysis of such samples has revealed interesting “motifs” in mobility patterns [5]. A different approach has been to use smart card transactions on public transit to get statistics of movements across different nodes and edges of a transportation network, e.g., [13]. Finally, transportation departments typically collect traffic flow statistics on different links, and across transportation analysis zones (TAZs). Each data set is a small fraction of the total data needed for urban analytics applications.

Much of the work on mobility modeling has focused on reconstructing a mobility trace for each individual from such samples or statistics. There has been a lot of work on using data mining and machine learning techniques to translate such statistics to a mobility model for all the individuals in the population. For instance, Yuan et al. [13] use a sequence prediction approach to complete a mobility sequence for an individual from partial observations. However, it is unclear how to use such methods when we only have information on flow measurements across different links. Furthermore, such mobility patterns could be considered in more general networks (i.e., not just transportation networks), by suitable abstractions of “locations”.

Reconstruction of mobility traces can be viewed at an abstract level as the generation of labeled sequences, that satisfy constraints corresponding to such statistics. An abstract way to represent such mobility patterns is to consider each individual trace as a labeled sequence $f(a_{i1}), \dots, f(a_{ik})$, where a_{i1}, \dots, a_{ik} is a sequence of “activities” corresponding to the individual, and f maps each a_{ij} to a label in a set L . The labels in L correspond to locations (e.g., a census block group of a county, or a base station ID). The interpretation is that the activity a_{ij} is performed at location $f(a_{ij})$. Flow measurements might be available on the set of links E of a network $H = (L, E)$ with edges of the form (ℓ, ℓ') connecting the locations in L . Each labeled activity sequence that is mapped on to locations and has ℓ and ℓ' as consecutive locations contributes to one unit of flow on (ℓ, ℓ') . See Fig. 1 for an example.

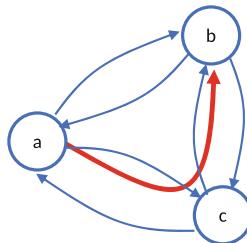


Fig. 1. A mobility trace as a sequence of labels in a directed network $H = (L, E)$ with $L = \{a, b, c\}$ and edge set E all the possible links. The labeled sequence a, c, b is a possible mobility trace.

Our Contributions

1. We formalize the construction of mobility traces from flow measurements as a network labeling problem denoted by CLS. We show that finding a feasible labeling is NP-hard in general, and design a linear programming based rounding algorithm with rigorous bounds on the deviations on the flow constraints.
2. Next, we consider the complexity of counting and sampling from the set of feasible labelings. Since determining if there exists a feasible labeling is NP-hard, the counting problem is naturally hard. However, we show that even if constraints are defined for individual pairs of labels, the counting problem is #P-hard. Finally, we design a simple dynamic program for both the counting and sampling problems, when $|L|$ and k are both constant.
3. We evaluate our method on a commuter flow data set for Washington DC and a set of nearby counties. We find that our linear programming solution gives an almost integral solution, and a simpler rounding, which simply rounds the fractional values, does very well.

2 Preliminaries

Let $P = \{p_1, \dots, p_n\}$ denote a set of individuals. Each $p_i \in P$ has associated a sequence $\text{seq}_i = a_{i,1}, \dots, a_{i,n_i}$; let $\text{Seq} = \{\text{seq}_1, \dots, \text{seq}_n\}$ be the set of all sequences, and $A = \cup_{i=1}^n \{a_{i,1}, \dots, a_{i,n_i}\}$. Let $L = \{\ell_1, \dots, \ell_m\}$ denote a set of labels. A *labeling* is a mapping $f: A \rightarrow L$. In general, some of the sequences may already be partially labeled; let $f_{init}: A_{init} \rightarrow L$ denote the partial labeling for $A_{init} \subseteq A$. If f_{init} is given, our focus is on finding a labeling $f(\cdot)$ consistent with f_{init} , i.e., where $f(a) = f_{init}(a)$ for all $a \in A_{init}$. In order to avoid notational overload, we will not specifically indicate f_{init} as part of the input. All our results will hold for any given partial labeling f_{init} . For a sequence seq_i , we refer to $f(\text{seq}_i) = f(a_{i,1}), \dots, f(a_{i,n_i})$ as a labeled sequence.

Let $H = (L, E)$ denote a network on the set of labels, where we will think of pairs of labels ℓ, ℓ' as edges of H . For a labeling $f(\cdot)$, let $N(\text{Seq}, f, (\ell_1, \ell_2)) = |\{(i, j) : f(a_{i,j}) = \ell_1, f(a_{i,j+1}) = \ell_2\}|$ denote the number of labeled sequences having consecutive terms labeled ℓ_1, ℓ_2 . We will refer to them as labeled sequences passing through the edge (ℓ_1, ℓ_2) ; note that we count multiplicity here. We sometimes denote it by $N(f, (\ell_1, \ell_2))$ when Seq is clear from the context. We extend the definition of $N(\cdot)$ to a set of links of H in the following manner: consider a set $S = \{(\ell_1, \ell'_1), (\ell_2, \ell'_2), \dots, (\ell_k, \ell'_k)\} \subseteq E$ of edges. Then

$$N(f, S) = N(\text{Seq}, f, S) = \sum_{(\ell_i, \ell'_i) \in S} N(f, (\ell_i, \ell'_i))$$

is the number of labeled sequences $f(\text{seq})$ which pass through any (ℓ_i, ℓ'_i) .

Let $\mathcal{S} \subseteq 2^E$ be a collection of sets of edges. For $S = \{(\ell_1, \ell'_1), (\ell_2, \ell'_2), \dots, (\ell_k, \ell'_k)\} \in \mathcal{S}$, we use $F_L(S)$ and $F_U(S)$ to denote lower and upper bounds on flow measurements from input data—this counts the number of mobility sequences in the input which pass through the edge $(\ell, \ell') \in S$. The problem we study is to

	a	b	c
a	1	3	3
b	1	0	2
c	0	3	0

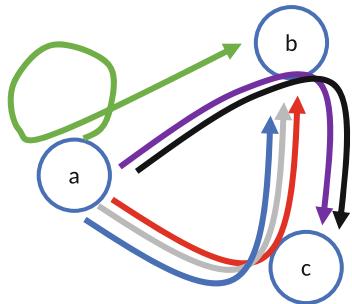


Fig. 2. Example: Seq consists of six sequences with $k = 3$. We have $L = \{a, b, c\}$, and $\mathcal{S} = L \times L$ is the set of all ordered pairs of labels. For each $S \in \mathcal{S}$, we have $F_L(S) = 0$, and $F_U(S)$ is shown in the matrix. The labeled sequences are shown in different colors: (1) $f(\text{seq}_1) = (a, a, b)$ (green), (2) $f(\text{seq}_2) = (a, b, c)$ (purple), (3) $f(\text{seq}_3) = (a, b, c)$ (black), (4) $f(\text{seq}_4) = (a, c, b)$ (blue), (5) $f(\text{seq}_5) = (a, c, b)$ (gray), (6) $f(\text{seq}_6) = (a, c, b)$ (red).

determine a labeling f such that $N(f, S)$ is within the interval $[F_L(S), F_U(S)]$. This is formalized as the CLS problem below.

Constrained Sequence Labeling (CSL) Problem

Given: P , Seq , L , \mathcal{S} , and $F_L(\cdot)$, and $F_U(\cdot)$, as defined above

Compute: a labeling $f: A \rightarrow L$ such that for all $S \in \mathcal{S}$, we have $N(f, S) \in [F_L(S), F_U(S)]$.

Example. Figure 2 illustrates the definitions through an example with $|\text{Seq}| = 6$, $k = 3$, $|L| = 3$, and \mathcal{S} consisting of pairs. Note that seq_1 , seq_2 and seq_3 all contribute to $F_L(a, b)$, $F_U(a, b)$, although at different positions. Finally, if we change $f(\text{seq}_2)$ (the purple sequence) to (a, b, b) instead of (a, b, c) , that would also be a feasible solution for the given bounds on $F_L(\cdot)$ and $F_U(\cdot)$.

In general, we are interested in finding not just one feasible labeling, but sampling from the space of all feasible labelings, or equivalently, count the number of such labelings. This is formalized as the problem below.

Constrained Sequence Label Sampling (CLSSample) Problem

Given: P , Seq , L , \mathcal{S} , and $F_L(\cdot)$, and $F_U(\cdot)$, as defined above

Compute: a labeling $f(\cdot)$ sampled uniformly at random from the set of all feasible labelings, i.e., $\{f: A \rightarrow L : N(f, S) \in [F_L(S), F_U(S)], \forall S \in \mathcal{S}\}$.

A related problem is to count the number of feasible solutions to a given instance of CLS; we refer to this as the CLSCOUNT problem.

Other Variations. In practice, we expect there might be inconsistencies or incompleteness in the input data, and there might not even be a feasible solution. In such cases, we consider approximate versions of various forms, e.g., finding

the largest subset $\text{Seq}' \subseteq \text{Seq}$ such that the flow counts are satisfied for Seq' . Finally, more generally, we would like to sample from the space of all feasible labelings f uniformly at random, instead of just generating one.

3 Hardness

We show here the decision version of CLS, namely deciding if there is a feasible solution $f(\cdot)$ such that $N(f, S) \in [F_L(S), F_U(S)]$ for all $S \in \mathcal{S}$, is NP-complete.

Theorem 1. *Deciding if an instance of CLS has a feasible solution is NP-complete.*

Proof. The proof is by reduction from the independent set problem. An instance consists of a graph $G = (V, E)$, and a parameter K , and the objective is to determine if there is an independent $U \subseteq V$ with $|U| = K$.

We reduce this to an instance of CLS as follows. Let $V = \{v_1, \dots, v_n\}$. The set Seq consists of K sequences $\text{seq}_1, \dots, \text{seq}_K$ with $\text{seq}_i = a_{i,1}, a_{i,2}$. We have $L = V$, and \mathcal{S} consists of (1) sets $S_{u,v} = \{(u, u), (v, v)\}$, for $u, v \in V$, with $(u, v) \in E$, and (2) the set $S_G = \{(v_1, v_1), \dots, (v_n, v_n)\}$. For each set $S_{u,v}$, we have $F_L(S_{u,v}) = 0$ and $F_U(S_{u,v}) = 1$. Finally, we have $F_L(S_G) = F_U(S_G) = K$.

We observe that there is a feasible labeling $f(\cdot)$ if and only if there is an independent set $U \subseteq V$ in G of size K . For the “if” part, suppose $U = \{u_1, \dots, u_K\}$ is an independent set. We construct the labeling f with $f(\text{seq}_i) = u_i, u_i$. Then, we have $N(f, (u, u)) = 1$ for $u \in U$, and 0 if $u \notin U$. This implies $N(f, S_G) = K$. For any $(u, v) \in E$, at most one of $u, v \in U$. This implies $N(f, (u, u)) + N(f, (v, v)) \leq 1$, so that $N(f, S_{u,v}) \in [0, 1]$. Therefore, $f(\cdot)$ satisfies all the constraints.

For the “only if” part, suppose we have a feasible labeling $f(\cdot)$. Since $N(f, S_G) = K$, it must be the case that there are exactly K nodes u such that $N(f, (u, u)) = 1$; let $U = \{u : N(f, (u, u)) = 1\}$. The constraints corresponding to $S_{u,v}$ for each $(u, v) \in E$ imply that $N(f, (u, u)) + N(f, (v, v)) \leq 1$. Therefore, at most one of $u, v \in U$, which implies U is an independent set.

4 Algorithm SeqRound to Construct a Feasible Labeling Algorithm

We describe an algorithm based on linear programming rounding. For simplicity, we assume that $|\text{seq}_i| = k$ for all $\text{seq}_i \in \text{Seq}$, where $k \geq 2$; the method can be extended easily even when the sequences have different lengths. Let $\mathcal{P} = L^k$ be a set of paths of labels. We have a variable $x(P)$ for each $P \in \mathcal{P}$, which is 1 if P is selected. Let $\ell(P)$ denote the set of label pairs (i.e., edges) in P , and let

$$\Delta = \max_{P \in \mathcal{P}} |\{\{S : \ell(P) \cap S \neq \emptyset\}\}|.$$

The following integer program (IP) directly corresponds to the CLS problem.

$$\forall S \in \mathcal{S} : \sum_{P: \ell(P) \cap S \neq \emptyset} x(P) \leq F_U(S) \quad (1)$$

$$\forall S \in \mathcal{S} : \sum_{P: \ell(P) \cap S \neq \emptyset} x(P) \geq F_L(S) \quad (2)$$

$$\sum_P x(P) = |\text{Seq}| \quad (3)$$

$$\forall P \in \mathcal{P} : x(P) \in \mathbb{Z} \quad (4)$$

Lemma 1. *An instance of CLS has a feasible solution if and only if (IP) is feasible.*

Our algorithm SEQROUND involves the following steps.

1. Solve the linear relaxation of (IP), with the constraints (4) replaced by $x(P) \geq 0$
2. Use the rounding algorithm of [7] (Theorem 3) to transform $x(\cdot)$ to an integral solution $\hat{x}(\cdot)$
3. If $\sum_P \hat{x}(P) > |\text{Seq}|$, pick $m = \sum_P \hat{x}(P) - |\text{Seq}|$ arbitrary paths Q_1, \dots, Q_m in \mathcal{P} with $\hat{x}(Q_i) > 0$ (possibly non-distinct), and set $\hat{x}(Q_i) = \hat{x}(Q_i) - 1$ for $i = 1, \dots, m$
4. If, on the other hand, $\sum_P \hat{x}(P) < |\text{Seq}|$, pick $m = |\text{Seq}| - \sum_P \hat{x}(P)$ paths Q_1, \dots, Q_m in \mathcal{P} , and set $\hat{x}(Q_i) = \hat{x}(Q_i) + 1$ for $i = 1, \dots, m$
5. Let $\mathcal{P}' = \{P : \hat{x}(P) > 0\}$. Order the paths in \mathcal{P}' as $P_1, \dots, P_{|\text{Seq}|}$ arbitrarily. Construct the labeling $f(\cdot)$ in the following manner:
 - (a) Let $P_i = \ell_{i1}, \dots, \ell_{ik}$
 - (b) For each $\text{seq}_i = a_{i1}, \dots, a_{ik}$, let $f(a_{ij}) = \ell_{ij}$.

Theorem 2. *The above algorithm runs in time polynomial in $|L|^k$. The labeling $f(\cdot)$ constructed above ensures that: (1) each sequence $\text{seq}_i \in \text{Seq}$ is labeled, and (1) $N(f, S) \in [F_L(S) - 2\Delta - 2, F_U(S) + 2\Delta + 2]$ for all S .*

Proof. We first rewrite the LP relaxation in the following equivalent manner:

$$\forall S \in \mathcal{S} : \sum_{P: \ell(P) \cap S \neq \emptyset} x(P) \leq F_U(S) \quad (5)$$

$$\forall S \in \mathcal{S} : \sum_{P: \ell(P) \cap S \neq \emptyset} -x(P) \leq -F_L(S) \quad (6)$$

$$\sum_P x(P) \leq |\text{Seq}| \quad (7)$$

$$\sum_P -x(P) \leq -|\text{Seq}| \quad (8)$$

$$\forall P \in \mathcal{P} : x(P) \geq 0 \quad (9)$$

Note that constraints (7) and (8) together imply $\sum_P x(P) = |\text{Seq}|$. Let the above LP be expressed as the system $Ax \leq b, x \geq 0$; note that x is indexed by the paths P . Let

$$D = \max_P \left\{ \sum_{i:A_{iP}>0} A_{iP}, - \sum_{i:A_{iP}<0} A_{iP} \right\}$$

be the maximum over all columns of A of the sum of the positive entries and the negative of the sum of all negative entries. Then, the rounding algorithm of [7] gives the integral solution \hat{x} such that $\hat{x}_P \in \{\lfloor x_P \rfloor, \lceil x_P \rceil\}$, and $A\hat{x} \leq b + D\mathbf{1}$, where $\mathbf{1}$ is the all 1's vector. Each variable $x(P)$ appears in $|\{S : \ell(P) \cap S \neq \emptyset\}| + 1 \leq \Delta + 1$ constraints with a positive coefficient, from the definition of Δ . It also appears in the same number of negative coefficient.

Therefore, the solution \hat{x} after step 2 satisfies the following properties:

- (i) For all $S \in \mathcal{S}$, $\sum_{P:\ell(P) \cap S \neq \emptyset} \hat{x}(P) \leq F_U(S) + \Delta + 1$, from constraints (5)
- (ii) For all $S \in \mathcal{S}$, $\sum_{P:\ell(P) \cap S \neq \emptyset} -\hat{x}(P) \leq -F_L(S) + \Delta + 1$, which implies $\sum_{P:\ell(P) \cap S \neq \emptyset} \hat{x}(P) \geq F_L(S) - \Delta - 1$
- (iii) $\sum_P \hat{x}(P) \leq |\text{Seq}| + \Delta + 1$
- (iv) $\sum_P -\hat{x}(P) \leq -|\text{Seq}| + \Delta + 1$, which implies $\sum_P \hat{x}(P) \geq |\text{Seq}| - \Delta - 1$

If it turns out that $\sum_P \hat{x}(P) > |\text{Seq}|$, by property (iii) above, we have $m = \sum_P \hat{x}(P) - |\text{Seq}| \leq \Delta + 1$. In this case, in step (3) of the algorithm, we reduce $\hat{x}(Q_i)$ for m such paths. This does not affect the upper bound corresponding to any set S , but it can change the lower bound, and we have $\sum_{P:\ell(P) \cap S \neq \emptyset} \hat{x}(P) \geq F_L(S) - 2\Delta - 2$. Similarly, if property (iv) above holds, step (4) of the algorithm increases $\hat{x}(P)$ values for m paths, and the upper bounds change to $\sum_{P:\ell(P) \cap S \neq \emptyset} \hat{x}(P) \leq F_U(S) + 2\Delta + 2$. In both the above cases, if $\sum_P \hat{x}(P) \neq |\text{Seq}|$, the adjustment to the paths ensures that all sequences get labeled.

4.1 Speeding up Algorithm SeqRound Using Constrained Flows

Step (2) of SEQROUND involves solving an LP with $|L|^k$ variables (one per path), and $O(|\mathcal{S}|)$ constraints. This takes time polynomial in $\max\{|L|^k, |\mathcal{S}|\}$, which grows very fast with k . While we expect a lower bound of $|\text{Seq}|$ on the running time (since each sequence needs to be labeled), $|L|^k$ could become larger than $|\text{Seq}|$ if k is large. The cost of solving the LP can be reduced significantly by reducing it to a constrained network flow problem, as we discuss below.

Constrained Network Flow Formulation. For simplicity, we assume as before that $|\text{seq}_i| = k$ for all i ; this approach can be extended even if the sequence lengths are different.

1. We create a directed network $G = (V, E)$, where (1) $V = \cup_{i=1}^k V_i$, with $V_i = \{\ell^{(i)} : \ell \in L\}$, and (2) $E = \{(\ell_1^{(i)}, \ell_2^{(i+1)}) : \ell_1, \ell_2 \in L, i = 1, \dots, k-1\}$.

2. Define variables $x(\ell_1^{(i)}, \ell_2^{(i+1)})$, for all $\ell_1, \ell_2 \in L$, and $i = 1, \dots, k - 1$. Solve the following linear program (LP_{flow}), to find a feasible solution:

$$\sum_{\ell_1} x(\ell_1^{(i-1)}, \ell_2^{(i)}) - \sum_{\ell_2} x(\ell_1^{(i)}, \ell_2^{(i+1)}) = 0, \quad \text{for all } \ell \in L, 1 < i < k \quad (10)$$

$$\sum_i \sum_{(\ell_1, \ell_2) \in S} x(\ell_1^{(i)}, \ell_2^{(i+1)}) \in [F_L(S), F_U(S)], \quad \text{for all } \ell_1, \ell_2 \quad (11)$$

$$\sum_{\ell_1, \ell_2} x(\ell_1^{(1)}, \ell_2^{(2)}) = |\text{Seq}| \quad (12)$$

$$x(e) \geq 0, \quad \text{for all } e \in E \quad (13)$$

3. If the objective value of (LP_{flow}) is not feasible, there is no solution to the CLS instance.
4. Decompose the edge flows $x(\cdot)$ into at most $|\text{Seq}|$ flow paths in the following manner, till while there exists $e \in E$ with $x(e) > 0$:
- Find a path $P = \ell_1^{(1)}, \dots, \ell_k^{(k)}$ with $x(\ell_i^{(i)}, \ell_{i+1}^{(i+1)}) > 0$ for all i
 - Set $x(P) = \min_{i=1}^k x(\ell_i^{(i)}, \ell_{i+1}^{(i+1)})$
 - Update x by reducing each $x(\ell_i^{(i)}, \ell_{i+1}^{(i+1)})$ by $x(P)$

Claim. The above algorithm finds a feasible fractional solution $x(P)$ for Step (2) of SEQROUND.

Non-unimodularity. A classical result from polyhedral theory implies that all extreme points of the corresponding polyhedron are integral if the matrix is totally unimodular [9]. The program (P) has a very similar structure as a network flow—in fact, the only difference is the capacity constraint: instead of it being defined for individual edges, it is defined for sets of edges. However, the resulting program is not totally unimodular, as we show below.

Lemma 2. *LP_{flow} is not totally-unimodular, in general.*

5 The CLSCOUNT and CLSSAMPLE Problems

Here, we study the complexity of CLSCOUNT and CLSSAMPLE. Sampling and counting are equivalent in a fundamental sense, due to the “self-reducible” structure of CLS: this implies that an algorithm for CLSCOUNT can be used to solve CLSSAMPLE, and vice versa (approximately); we refer to [10] for discussion on these connections between counting and sampling.

First, Theorem 1 implies that any finite approximation to the number of solutions to an instance of CLS is NP-hard. However, the proof involves instances in which the set S is pretty complex. We show that the counting problem is $\#P$ -hard even when S consists of sets with individual label pairs.

Theorem 3. *The CLSCOUNT problem is $\#P$ -hard even if $|S| = 1$ for all $S \in \mathcal{S}$, and all sequences in Seq have length 4.*

Proof. We show that CLSCOUNT with this restriction on the sets is a generalization of the problem of counting the number of perfect matchings in a bipartite graph, which is $\#P$ -hard—this is referred to as the permanent of the graph (see [10]).

An instance of the permanent problem is a bipartite graph $G = (V = V_1 \cup V_2, E)$, with all edges between V_1 and V_2 . A subset $E' \subseteq E$ is a perfect matching if each node in V is incident to exactly one edge in E' . Counting the number of such perfect matchings is $\#P$ -hard. We reduce this to an instance of CLSCOUNT in the following manner. Let $n = |V_1| = |V_2|$. We define $\text{Seq} = \{\text{seq}_1, \dots, \text{seq}_n\}$, with $\text{seq}_i = a_{i1}, a_{i2}, a_{i3}, a_{i4}$, for $i = 1, \dots, n$. We have $L = V_1 \cup V_2 \cup \{s, t\}$, with a partial labeling f_{init} defined in the following manner: $f_{init}(a_{i1}) = s$, and $f_{init}(a_{i4}) = t$ for all i . We define $\mathcal{S} = \{s\} \times V_1 \cup V_1 \times V_2 \cup V_2 \times \{t\}$. The bounds $F_L(\cdot), F_U(\cdot)$ are defined in the following manner: $F_L(S) = F_U(S) = 1$ for all $S \in \{s\} \times V_1 \cup V_2 \times \{t\}$. For all $e = (u, v) \in E$, we define $F_L(u, v) = 0, F_U(u, v) = 1$. Finally, for all $(u, v) \notin E$, we define $F_L(u, v) = F_U(u, v) = 0$.

Consider any feasible labeling f consistent with f_{init} . Then, for each seq_i we must have $f(a_{i1}) = s, f(a_{i4}) = t$. Since $F_L(s, u) = F_U(s, u) = 1$ for all $u \in V_1$, it must be the case that $f(\text{seq}_i)$ has a distinct label $u_i \in V_1$ for each i . Similarly, each $f(\text{seq}_i)$ has a distinct node in V_2 as the third label. This means each $f(\text{seq}_i)$ is a path s, u_i, v_i, t , where $u_i \in V_1, v_i \in V_2$, and $u_i \neq u_j, v_i \neq v_j$ for $i \neq j$. Further, we must have $(u_i, v_i) \in E$. Therefore, the set of edges $\{(u_i, v_i) : i = 1, \dots, n\}$ is a perfect matching in G , and there is a one to one correspondence between feasible labelings $f(\cdot)$ and perfect matchings in G .

Solving CLSCOUNT and CLSSAMPLE When $|L|$ and $|\text{seq}_i|$ Are All Constant. In this special case, we show that CLSCOUNT can be solved in time polynomial in $|\text{Seq}|$. We then discuss how this can be used to solve CLSSAMPLE using ideas from [10].

Our algorithm DPCount uses a dynamic programming technique. As before, we assume $|\text{seq}_i| = k$ for all i , though this assumption can be removed. Since $|L|$ is a constant, \mathcal{S} has a constant size; let $\mathcal{S} = \{S_1, \dots, S_M\}$.

Algorithm DPCount involves the following steps.

1. We maintain a table $T[\cdot]$ with entries of the form $T[i, x_1, \dots, x_M]$, which indicates the number of partial labelings f of $\text{seq}_1, \dots, \text{seq}_i$, such that $N(\{\text{seq}_1, \dots, \text{seq}_i\}, f, S_j) = x_j$
2. For $i = 1$, $T[1, x_1, \dots, x_M]$ is precisely the number of labelings g of seq_1 , such that for all j : $N(\{\text{seq}_1\}, g, S_j) = x_j$; this can be computed by testing all $|L|^k$ possible labelings of seq_1 .
3. Next, consider $i > 1$, and assume the entries $T[i, x_1, \dots, x_M]$ have already been computed for all x_1, \dots, x_M

- (a) Consider each possible labeling $g : \{a_{i+1,1}, \dots, a_{i+1,k}\} \rightarrow L$ of seq_{i+1} .

There are $|L|^k$ such labelings. For each $S_j \in \mathcal{S}$, and $(\ell, \ell') \in S_j$, recall that $N(\text{seq}_{i+1}, g, (\ell, \ell')) = |\{(i+1, j) : g(a_{i+1,j}) = \ell, g(a_{i+1,j+1}) = \ell'\}|$, which is the number of times the pair (ℓ, ℓ') occurs consecutively in $g(\text{seq}_{i+1})$.

$$\text{Let } g(S_j) = \sum_{(\ell, \ell') \in S_j} N(\{\text{seq}_{i+1}\}, g, (\ell, \ell'))$$

- (b) $T[i+1, x_1, \dots, x_M] = \sum_q T[i, x_1 - g(S_1), \dots, x_M - g(S_M)]$
4. Let $U = \{(x_1, \dots, x_M) : F_L(S_j) \leq x_j \leq F_U\}$. Return $\sum_{(x_1, \dots, x_M) \in U} T[n, x_1, \dots, x_M]$.

Theorem 4. If $|L|$ and k are constants, algorithm DPCOUNT correctly computes the number of feasible labelings in polynomial time.

Proof. First, observe that for any S_j , we have

$$N(f, S_j) = \sum_{(\ell, \ell') \in S_j} N(f, (\ell, \ell')) \leq |S_j|n = O(|L|^2 n),$$

which is $O(n)$ if $|L|$ is a constant. This implies the size of the table $T[\cdot]$ is $O(n^{M+1})$, when $|L|$ and k are constants.

Step (2) of the algorithm tries every possible labeling $g \in L^k$, and evaluates the bound for every S_j ; therefore it takes $O(|L|^k M)$ time. In each round of the recursive step (3), the algorithm considers each possible labeling g of seq_{i+1} , every set S_j (as in Step (2)), and every possible entry $T[i, x_1, \dots, x_M]$. This takes $O(n^{M+1}|L|^k)$ time.

From Counting to Sampling. The exact counts in table $T[\cdot]$ allow for easy sampling, which can be done in an iterative manner. The labeling for seq_n can be chosen based on all the entries $T[n-1, \dots]$, for all possible combinations, and the subsequent labelings are done within that subtree.

6 Empirical Results

We use our algorithm to generate an urban mobility network for the Washington DC area, using county to county travel estimates. Our main objective is to evaluate the performance of our algorithm, and examine the solutions produced.

Dataset and Method. We use the commuter flow dataset from the American Community Survey (ACS) [3] for the counties around Washington DC (Fig. 3). The ACS dataset uses extensive surveys, and then provides population adjusted estimates of number of commuters between any pair of counties for a typical day. Here, we consider 15 counties, indicated by their FIPS codes. We take the set L of labels to be these 15 FIPS codes. We take $k = 3$, and $n = 1.2$ million. The ACS dataset can be represented as a flow matrix $\text{flow}[i, j]$, which gives the flow estimate from FIPS i to FIPS j . In our experiment, we take $\mathcal{S} = \{(i, j) : i, j \text{ are FIPS codes}\}$. For each $S = (i, j) \in \mathcal{S}$, we set $F_U(S) = \text{flow}[i, j]$, and $F_L(S) = \text{flow}[i, j] - 1$.

We use the Gurobi software to implement SEQROUND on this dataset.

1. **Approximation guarantee of SeqRound in practice.** For the above instance, with \mathcal{S} consisting of pairs, the LP solution was half integral, i.e., all the variables were either integral, or some integer plus 0.5. As a result, simply rounding the fractional solutions gives an integral solution, which gives minimal violation of the constraints.

2. **Structure of solution.** For $k = 3$, our solution had 100 types of labeled sequences ℓ_1, ℓ_2, ℓ_3 , which is a small fraction of the 15^3 possible sequence types. Of these, 35 types of sequences involve a repeated label.



Fig. 3. Fifteen counties around Washington DC

7 Conclusions

CLS provides a natural formulation to reconstruct urban mobility traces from different kinds of flow measurements. The formulation is also general enough to incorporate a variety of constraints, allowing it to be useful in a number of applications. The basic setup is relevant not just in urban mobility, but also activity sequences in other types of networks, e.g., on a web graph.

Acknowledgments. This work has been partially supported by the following grants: NSF CRISP 2.0 Grant 1832587, DTRA CNIMS (Contract HDTRA1-11-D-0016-0001), NSF DIBBS Grant ACI-1443054, NSF EAGER Grant CMMI-1745207, and NSF BIG DATA Grant IIS-1633028.

References

1. Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M.: Human mobility: models and applications. *Phys. Rep.* **734**, 1–74 (2018)
2. Barrett, C.L., Beckman, R.J., Khan, M., Anil Kumar, V.S., Marathe, M.V., Stretz, P.E., Dutta, T., Lewis, B.: Generation and analysis of large synthetic social contact networks. In: Winter Simulation Conference, pp. 1003–1014. Winter Simulation Conference (2009)

3. US Census Bureau: American Community survey: Commuting flows. <https://www.census.gov/topics/employment/commuting/guidance/flows.html>. Accessed 11 Oct 2019
4. Eubank, S., Guclu, H., Anil Kumar, V.S., Marathe, M., Srinivasan, A., Toroczkai, Z., Wang, N.: Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184 (2004)
5. Gonzalez, M.C., Hidalgo, C., Barabasi, A.-L.: Understanding individual human mobility patterns. *Nature* **453**, 779–82 (2008)
6. Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., Willinger, W.: Human mobility modeling at metropolitan scales. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, MobiSys 2012, pp. 239–252. ACM, New York (2012)
7. Karp, R.M., Leighton, F.T., Rivest, R.L., Thompson, C.D., Vazirani, U.V., Vazirani, V.V.: Global wire routing in two-dimensional arrays. *Algorithmica* **2**(1–4), 113–129 (1987)
8. Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C.: Unravelling daily human mobility motifs. *J. R. Soc. Interface* **10**(84), 20130246 (2013)
9. Schrijver, A.: Combinatorial Optimization - Polyhedra and Efficiency. Springer, Heidelberg (2003)
10. Sinclair, A., Jerrum, M.: Approximate counting, uniform generation and rapidly mixing Markov chains. *Inf. Comput.* **82**(1), 93–133 (1989)
11. Toch, E., Lerner, B., Ben-Zion, E., Ben-Gal, I.: Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowl. Inf. Syst.* **58**(3), 501–523 (2019)
12. Yoon, J., Noble, B.D., Liu, M., Kim, M.: Building realistic mobility models from coarse-grained traces. In: Proceedings of the 4th International Conference on Mobile Systems, Applications and Services, MobiSys 2006, pp. 177–190. ACM, New York (2006)
13. Yuan, N.J., Wang, Y., Zhang, F., Xie, X., Sun, G.: Reconstructing individual mobility from smart card transactions: a space alignment approach. In: 2013 IEEE 13th International Conference on Data Mining, pp. 877–886. IEEE (2013)

Quantifying Success through Social Network Analysis



A Network Approach to the Formation of Self-assembled Teams

Rustom Ichhaporia, Diego Gómez-Zará^(✉), Leslie DeChurch,
and Noshir Contractor

Northwestern University, Evanston, IL 60208, USA

{rustom.ichhaporia,dechurch,nosh}@northwestern.edu,
dgomezara@u.northwestern.edu

Abstract. Which individuals in a network make the most appealing teammates? Which invitations are most likely to be accepted? And which are most likely to be rejected? This study explores the factors that are most likely to explain the selection, acceptance, and rejection of invitations in self-assembling teams. We conducted a field study with 780 participants using an online platform that enables people to form teams. Participants completed an initial survey assessing traits, relationships, and skills. Next, they searched for and invited others to join a team. Recipients could then accept, reject, or ignore invitations. Using Exponential Random Graph Models (ERGMs), we studied how traits and social networks influence teammate choices. Our results demonstrated that (a) agreeable leaders with high psychological collectivism send invitations most frequently, (b) previous collaborators, leaders, competent workers, females, and younger individuals receive the most invitations, and (c) rejections are concentrated in the hands of a few.

Keywords: Social network analysis · ERGM · Team formation

1 Introduction

Teams are the basic unit of work in most organizations. Trends toward empowerment and autonomy are giving people more say in the formation of their teams. Whereas most traditional teams are appointed by a manager and relatively static, modern dynamic work environments encourage members to self-assemble their own teams, searching for and inviting new teammates as opposed to being passively staffed into them. And while some teams are embedded in organizations and formally staffed, many others are formed voluntarily by individuals for specific projects [29]. In this case, team members have broad autonomy to make decisions about team membership [12]. Guided by information technologies and access to rich information, self-assembled teams originate from complex networks that exhibit patterns of interdependence between individuals and their social environment, where current members and potential collaborators interact to enhance team performance.

Despite the pervasive interest in the potential of self-assembled teams, little is known about the decision processes individuals use as they self-organize into teams. Related literature on team composition explores the link between the personal characteristics represented in a team and its cohesion, processes, and performance [9,14]. Far less is known about the organizing rules that come before and ultimately determine team composition [15,25].

To fill this gap, we examine through a network perspective the factors most likely to explain the selection, acceptance, and rejection of team members. We seek to discover the assembly mechanisms by focusing on individuals' attributes and their network relations, the structural signatures that most reliably determine the invitations and responses of members as they self-assemble into teams. Our findings are from an observational field study of 780 students from 14 project courses across five U.S. universities. Participants formed teams using *MyDreamTeam*¹, an online team formation system. Our data includes survey results reporting the individuals' skills, traits, and social networks, as well as digital traces of their searches, invitations, and responses to invitations. Using inferential network analysis, we model team invitations that were sent, accepted, rejected, and ignored as four distinct networks. A directed tie represents an invitation or response between two individuals. We use Exponential Random Graph Models (ERGMs) to analyze these networks by studying how their structural signatures, including participants' traits and social networks, influence participants' teammate choices.

Our contributions are twofold. First, this work highlights the dependencies of self-assembly formation mechanisms on individuals' choices, social networks, and team member characteristics. Second, we contribute to the expanding literature on complex social networks and self-assembled team formation with an empirical case study using network analysis. These findings delineate the mechanisms by which structural signatures, including individual traits and social networks, influence individuals' team member selection.

2 Related Literature

Research on team formation has identified several factors that explain what people perceive as important when choosing others with whom they want to work. We categorize these factors at the individual, relational, and network levels.

At the *individual level*, previous studies have shown that people forming teams often look for individuals who have the work experience and relevant skills to perform the team's tasks [8]. Similarly, team members search for others with complementary skills that are missing in their teams [29]. In team member selection, personal characteristics determine to what extent individuals are looking for others who are similar to them. Based on homophily and social identification theories, studies have shown that individuals like to work with others who share a certain level of similarity, since they may be biased against (or uncomfortable with) people who are different from themselves [28]. Previous studies have also

¹ <http://sonic.northwestern.edu/mdt>.

demonstrated that working with similar individuals facilitates team communication and reduces cultural barriers and uncertainty that can arise when working with unfamiliar people [18]. Some studies also find gender influences teammate selection. Women display a greater preference to work with other women over men because of the disparity of influential behaviors between men and women in small groups [22]. Balancing personality traits is also fundamental for team member selection. In a corporate environment, research has shown that teams having members with high agreeableness, extroversion, and conscientiousness are more likely to be cooperative and work more effectively with stakeholders [26].

At the *relational level*, individuals may be more inclined to work with those who are familiar to them than strangers. Familiar relationships vary in their quality and strength. Basic familiarity may involve indirect connections (e.g., friends of a friend, unknown co-workers at the same organization) while more extensive familiarity accrues to those who have directly collaborated in the past [11]. The familiarity principle asserts that people who have had positive interactions in the past are likely to be positively disposed toward one another, since there is an expectation that the person's behavior will be similar in future interactions [15]. Thus, in seeking predictability, people are likely to choose to work with others whom they have already worked with, particularly if the experience was positive.

Finally, the *network level* considers how the social structural context and individuals' network positions affect team member selection. One study found that network structures determine how easy or hard it is for entrepreneurs to recognize opportunities and collaborators; while small-world network structures provide local social clusters facilitating team member selection through similarity and familiarity mechanisms, scale-free network structures provide a single, larger social cluster, and make team formation more instrumental [1]. Highly connected individuals are more likely to be selected as teammates than those who are less connected in their social environment because they have more expansive access to resources via their connections [3]. For example, individuals who are highly connected with other groups are accordingly more likely to have access to novel ideas that originate outside of the team. Another example is the presence of brokers, who are individuals that bridge different, sparsely connected groups. A broker can facilitate resources and collaboration between otherwise disconnected groups. In summary, well-positioned individuals are more likely to be nominated by those who are looking for access to different resources [2].

While most prior research has explored factors that influence team member selection at the individual, relational, or network levels, less is known about how factors at all three levels compositely shape team member selection. Moreover, research has focused on who ends up in a team together, but has yet to explain the micro-dynamics through which invitations are extended by one party, and separately, the micro-dynamics of through which other parties respond to invitations [10]. We study these behaviors separately and identify to network mechanisms arising at all three levels by analyzing tie formation in the network. Our research questions are:

RQ1. Which structural signatures shape team-assembly mechanisms?

RQ2. Which individuals make the most appealing teammates?

RQ3. Which individuals are most likely to be accepted (rejected)?

3 Methodology

To answer these questions, we conducted a field study observing team self-assembly as it unfolded in 14 courses using team project-based learning. Data was collected over the course of 3 years and consists of 9 undergraduate and 5 graduate courses conducted at five U.S. universities. A total of 780 students (214 international students and 566 U.S. nationals; 474 undergraduate and 306 graduate) participated in the study, with an average course size of 55.71 ($SD = 22.76$). The participants' gender ratio was slightly imbalanced, with 420 female students and 360 males. The average age was approximately 25 ($SD = 7.12$), and the average number of fellow students each participant had worked with was 12.92 ($SD = 19.47$). The team task given to the students was a project related to their coursework for the undergraduate courses and a case study analysis and discussion tasks for graduate courses. All students assembled into one team, and they only participated in one course. We asked participants' voluntary consent to analyze their platform usage after the team assembly was completed. We notified them that the individual data collected would not be shared and that their participation in this study would not impact their grades.

3.1 Procedure

Team formation occurred online using a team recommender system called *MyDreamTeam*. On the platform, participants first create profiles, search for others, and send invitations that can be accepted or rejected until teams are formed. Participants completed the following steps:

3.1.1. Initial Survey. Participants were asked to populate a profile on the team formation platform. They provided background information, such as their name, nationality, and gender. The platform enabled participants to display public information in their profiles, such as their background, hobbies, and motivations. Participants were also prompted for more information relevant to potential teammate identification, including their individual attributes and which other participants constituted their previous collaborations, to estimate potentially influential variables (More details in Sect. 4.1). Participants' responses were confidential; they were, however, used by the platform in providing recommended teammates based on search queries.

3.1.2. Search Stage. After completing their profiles and surveys, participants were prompted to fill out a search query to find potential collaborators in response to their preferences. The query prompted participants to provide their preferences for potential teammates on the attributes collected in the initial survey on a 7-point Likert scale, ranging from "Not important at all" (-3), to "Don't care"

(0), and “Yes, for sure.” (+3). The platform used all of the preferences included in the search query and rank-ordered all potential teammates based on their match to the query and displayed them in a list.

3.1.3. Team Formation Stage. After browsing potential collaborators’ profiles from the search results, participants sent invitations to others to join their teams. The recipient had a choice between accepting, rejecting, or ignoring each invitation. For each course, participants had to assemble a team within one week. Otherwise, the instructors assigned them to an existing team.

4 Network Modeling

We use Exponential Random Graph Models (ERGMs) to identify the individual, relational, and network-level variables that best explain the motivations behind team member selection. ERGMs are a type of stochastic model that provides an appropriate analytic methodology to test multi-theoretical multilevel network hypotheses [4, 21]. This statistical model estimates the likelihood of the observed network structures emerging from all possible network configurations generated based on certain hypothesized self-organizing principles. Similar to logistic regressions, ERGM uses Maximum Likelihood Criterion to estimate the network statistics’ coefficients. Positive and significant coefficients indicate that the corresponding independent variable is more likely to influence invitations being extended than by chance. Negative and significant coefficients indicate that the independent variable is less likely to result in an invitation being extended than by chance alone. We use Markov-Chain Monte Carlo (MCMC) to identify maximum likelihood estimates (MLE) for parameter values. MCMC simulates thousands of random networks fitting the model’s quantifiable properties, rather than attempting to count the impossibly large number of possible network’s edges permutations. Once the ERGM and its coefficients are estimated, we test whether the observed network is likely to be observed within the distribution of simulated networks. Analyses were carried out using the *ergm* package from *statnet* [13].

We modeled four separate networks (Fig. 1). We first define the *sent invitation network* by gathering all the invitations sent by participants, not differentiating between accepted, rejected, or ignored invitations. We then create the *accepted network*, *rejected network*, and *ignored network* by filtering out the invitations according to senders’ responses. To study the overall effect among all these cases, we create an adjacency matrix for each of the 14 courses and then combined them with a block diagonal matrix specification. This method ensures structural zeros disallow the possibility of participants from different courses forming ties [17]. For the accepted, rejected, and ignored networks, we estimated ERGMs using additional structural zeros for ties where invitations were not sent. In other words, we estimated the statistical coefficients conditioned on the original invitations sent. This ensured that acceptances or rejections could not exist between two participants when no invitation was originally sent. A network was created from

each matrix, composed of nodes (representing participants) with attributes (representing their individual traits) explained in Sect. 4.1 and directed ties pointing from the sender node of an invitation to its receiver node. After populating each of the network's nodes with their respective attribute values, it was passed into an ERGM, yielding log-odds estimations for the likelihood of ties forming between any two nodes as a function of the estimates [27].

To measure the fit of the estimated ERGM to the observed data, we use the simulation-based Goodness of Fit (GoF) test from the *ergm* package. We sample one network out of every 1,000, spread across 10 million iterations, and compare the characteristics of generated networks to the statistics of the observed networks.

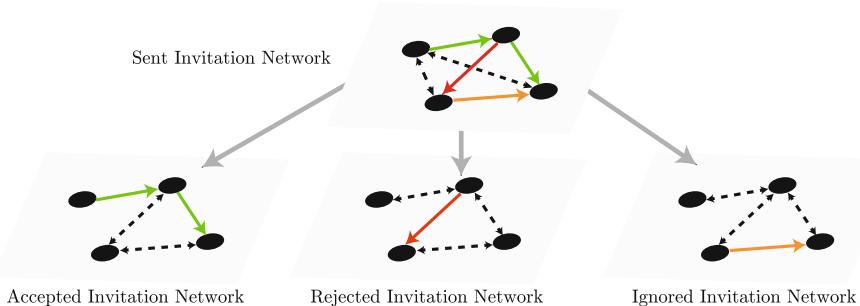


Fig. 1. Four networks to study self-assembly: sent, accepted, rejected, and ignored invitations networks. The latter three are created by filtering out the invitations according to senders' responses. Dashed lines represent potential invitations (directed ties) that could emerge.

4.1 Variables and Measures

Individual Level: The following attributes were included as nodal covariates for both senders (out-links) and receivers (in-links).

Competence. To assess participants' level of competence at skills relevant to their course, we asked them to self-report their degree of proficiency in six areas that were most useful for completing the project. Skill level was self-reported on a 5-point Likert scale ranging from “Not at all skilled” to “Extremely skilled” [20]. We then averaged these six scores per individual to get an overall competence score for each person.

Leadership Experience. We measured individuals' prior leadership experience using the 8-item Adolescent Leadership Activities Scale [19]. As the items showed acceptable reliability (Cronbach's alpha α equal to 0.76), they were averaged into one leadership experience score.

Psychological Collectivism. Defined as the propensity participants had for group activity as opposed to individual work, we assessed five facets of psychological

collectivism from [16]: preference for in-groups, reliance on in-groups, concern for in-groups, acceptance of in-group norms, and prioritization of in-group goals. This 15-item measure demonstrated acceptable reliability ($\alpha = 0.84$), and so a single psychological collectivism score was computed per individual.

Social Skills. We measured participants' social skills using the 7-item Political Skill Inventory [7] to capture four dimensions: social astuteness, interpersonal influence, networking ability, and apparent sincerity. Observing acceptable reliability ($\alpha = 0.76$), we computed this score for each individual.

Creativity. We determined participants' creative self-efficacy using [23]. The 3-item scale measured participants' belief in their own ability to complete creative goals. Observing acceptable reliability ($\alpha = 0.70$), we computed a creativity score for each person by averaging the items.

Personality. We used the mini-IPIP scales [5] which assessed the Five-Factor Model attributes of agreeableness ($\alpha = 0.72$), conscientiousness ($\alpha = 0.69$), extroversion ($\alpha = 0.75$), neuroticism ($\alpha = 0.61$), and openness ($\alpha = 0.73$). Participants responded to 20-items (four per trait), and the items were then averaged for each trait. We note that reliability was borderline on Neuroticism and Conscientiousness, but as these are established measures, we left them as intended.

Search Results Prevalence. To control whether the frequency with which individuals were recommended on the platform influenced their received invitations, we measured the number of times a given participant was ranked in the top-10 choices recommended to any other participant by the search platform.

Relational Level: In the initial survey, we gave participants a full list of the people in their course and asked them which people they had previously collaborated with, to represent the prior collaboration network. We consolidated each participant's responses by assigning a relationship between two participants if at least one participant reported a connection to the other.

Network Level: Finally, we computed metrics at the network level with the following terms:

Popularity. We measured the likelihood that a participant will receive a disproportionate number of invitations compared to others. We included the geometrically weighted indegree term (i.e., the weighted count of invitations they received), which models participants' indegree distribution and estimates how concentrated are the received invitations in certain participants. A significantly positive estimate implies a less centralized network that has more middle-degree participants and invitations are homogeneously distributed, whereas a significantly negative estimate indicates a skewed network that has high- and low-indegree participants and invitations are concentrated only in some of them. A negative estimate reflects the presence of hubs which received a lot more invitations. Popularity was modeled using the `gwidegree` term in the `ergm` package.

Activity. We measured the likelihood that a participant will send out a disproportionate number of invitations compared to others. We included the geometrically weighted outdegree term (i.e., the weighted count of invitations they sent), which models participants' outdegree distribution and estimates how concentrated are the sent invitations in certain users. Similar to the previous term, a significantly positive coefficient implies a less centralized network and a homogeneous distribution of senders, whereas a significantly negative coefficient indicates that certain senders are responsible for most of the invitations. Activity was modeled using the `gwodegree` term in the *ergm* package.

Indirect Teammate Selection. This statistic measures the extent to which participants who did not send invitations to others are potentially connected through third-participants. This term was measured by calculating the geometrically weighted frequency of dyadic shared partners, representing the likelihood of participants inviting people who in turn also invite a shared third person. This statistic was modeled using the `gwdesp` term in the *ergm* package.

5 Results

In total, participants sent 2,639 invitations, accepted 1,022 invitations, rejected 232 invitations, and ignored 1,385 invitations. ERGM results are displayed in Table 1. The GoF test determined that the observed networks' statistics were well explained by the ERGM models, lying within 95% of the confidence interval. We then analyze the results.

Individual Level: We separate the ERGMs individual-level terms for readability into two parts: indegree terms (pertaining to the receiver's attributes) and outdegree terms (pertaining to the sender's attributes).

Receiver Terms. The recipients of invitations were more likely to be female than male ($\beta = -0.140$), displayed high levels of competence ($\beta = 0.506$) and leadership experience ($\beta = 0.284$), and appeared frequently in search results ($\beta = 0.003$). Those who accepted invitations tended to be younger compared to the class ($\beta = -0.025$), displayed higher leadership experience ($\beta = 0.270$), and appeared in fewer top-10 search result listings ($\beta = -0.016$). Participants who rejected invitations displayed lower leadership experience ($\beta = -0.351$), higher creativity ($\beta = 0.332$), higher agreeableness ($\beta = 0.437$), lower neuroticism ($\beta = -0.337$), and appeared less frequently in search results ($\beta = -0.006$). Those who ignored invitations tended to appear more often in search results ($\beta = 0.007$), but had lower agreeableness ($\beta = -0.284$).

Sender Terms. Those sending invitations were more likely to have higher levels of competence ($\beta = 0.198$), leadership experience ($\beta = 0.176$), psychological collectivism ($\beta = 0.171$), agreeableness ($\beta = 0.148$), and openness ($\beta = 0.111$), as well as appearing slightly more frequently in search results ($\beta = 0.002$). In contrast, they displayed lower extraversion scores ($\beta = -0.087$). Those whose invitations

were accepted frequently had higher leadership experience ($\beta = 0.389$), lower psychological collectivism ($\beta = -0.333$), and appeared marginally less often in search results ($\beta = -0.007$). The rejected invitations' senders tended to displayed higher creativity ($\beta = 0.257$) but low competence ($\beta = -0.5$). Ignored invitations frequently came from participants with low scores on leadership experience ($\beta = -0.205$).

Table 1. ERGM Maximum Likelihood estimates. Standard error in parentheses. Dependent variable: teammate invitation network.

Network level term	Sent	Accepted	Rejected	Ignored
<i>Receiver effects</i>				
Age	0.004 (0.002)	-0.025 (0.008)**	0.010 (0.007)	0.000 (0.005)
Gender (M)	-0.140 (0.045)**	0.157 (0.119)	-0.157 (0.142)	0.029 (0.086)
Competence	0.506 (0.058)***	0.248 (0.162)	-0.185 (0.203)	-0.03 (0.113)
Leadership Experience	0.284 (0.044)***	0.270 (0.107)*	-0.351 (0.113)**	0.004 (0.076)
Psychological Collectivism	0.068 (0.052)	0.048 (0.144)	0.219 (0.171)	-0.167 (0.107)
Social Skills	0.019 (0.042)	0.052 (0.101)	-0.079 (0.124)	0.001 (0.073)
Creativity	0.065 (0.035)†	-0.078 (0.087)	0.332 (0.119)**	-0.064 (0.064)
Search Results	0.003 (0.000)***	-0.016 (0.002)***	-0.006 (0.002)*	0.007 (0.001)***
Prevalence				
Agreeableness	-0.065 (0.062)	0.289 (0.158)†	0.437 (0.179)*	-0.284 (0.117)*
Conscientiousness	-0.045 (0.053)	0.038 (0.131)	0.022 (0.157)	-0.042 (0.097)
Extroversion	0.063 (0.056)	0.154 (0.139)	-0.283 (0.177)	-0.065 (0.103)
Neuroticism	0.055 (0.052)	-0.169 (0.125)	-0.337 (0.144)*	0.164 (0.094)†
Openness	0.096 (0.053)†	-0.002 (0.134)	0.212 (0.158)	-0.084 (0.101)
<i>Sender effects</i>				
Age	0.001 (0.002)	-0.011 (0.007)	0.013 (0.008)	0.004 (0.005)
Gender (M)	-0.028 (0.038)	0.145 (0.119)	0.082 (0.149)	-0.087 (0.088)
Competence	0.198 (0.053)***	0.310† (0.167)	-0.500 (0.232)*	0.123 (0.121)
Leadership Experience	0.176 (0.033)***	0.389 (0.105)***	-0.106 (0.114)	-0.205 (0.074)**
Psychological Collectivism	0.171 (0.045)***	-0.333 (0.137)*	-0.356 (0.183)†	0.205 (0.109)†
Social Skills	0.006 (0.032)	0.160 (0.098)	0.062 (0.130)	-0.033 (0.075)
Creativity	0.043 (0.027)	-0.077 (0.083)	0.257 (0.121)*	-0.082 (0.065)
Search Results	0.002 (0.001)**	-0.007 (0.002)**	0.001 (0.002)	0.002 (0.001)
Prevalence				
Agreeableness	0.148 (0.049)**	-0.179 (0.153)	0.247 (0.196)	-0.017 (0.110)
Conscientiousness	-0.014 (0.044)	0.120 (0.132)	0.166 (0.174)	-0.045 (0.099)
Extroversion	-0.087 (0.042)*	0.069 (0.136)	-0.009 (0.165)	-0.082 (0.104)
Neuroticism	-0.021 (0.042)	0.032 (0.126)	0.196 (0.162)	-0.028 (0.094)
Openness	0.111 (0.044)*	0.050 (0.126)	-0.249 (0.169)	-0.112 (0.097)
<i>Relational-level effects</i>				
Previous Collaboration	0.275 (0.084)**	0.201 (0.179)	-0.628 (0.304)*	0.016 (0.163)
<i>Network-level effects</i>				
Edges	-8.218 (0.616)***	-4.068 (1.557)**	-2.064 (1.684)**	3.699 (1.129)**
Popularity	1.037 (0.204)***	1.590 (0.153)***	-1.509 (0.219)***	-1.491 (0.140)***
Activity	-1.665 (0.126)***	1.900 (0.153)***	-0.739 (0.212)***	-1.573 (0.136)***
Indirect Teammate Selection	-0.128 (0.008)***	-0.269 (0.047)***	0.280 (0.047)***	-0.091 (0.018)***
AIC	17,023	2,591	1,248	2,811
BIC	17,307	2,780	1,437	3,000

Significance codes: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*), $p < 0.1$ (†).

Relational Level: We found that participants were significantly more likely to send invitations to those who they had collaborated with before ($\beta = 0.275$). In the same vein, participants were highly unlikely to reject invitations from their former collaborators ($\beta = -0.628$). However, prior collaboration was not significantly related to the likelihood of accepting or ignoring an invitation.

Network Level: All structural terms were significant. The popularity term was positive for the sent ($\beta = 1.037$) and accepted networks ($\beta = 1.590$). This means that the invitations received and accepted by participants were relatively evenly distributed. In contrast, there was an uneven pattern in the other networks where a few individuals rejected ($\beta = -1.509$) and ignored ($\beta = -1.491$) most invitations. Participants' activity demonstrated a similar pattern, with a wide distribution of participants sending accepted invitations ($\beta = 1.900$) but a concentrated number of individuals having their invitations rejected ($\beta = -0.739$) or ignored ($\beta = -1.573$). The negative term in the sent network shows the concentration of senders among a few participants ($\beta = -1.665$). Finally, the indirect team selection was negative in all but the rejected networks, meaning participants often rejected the same people as their potential teammates ($\beta = 0.280$).

6 Discussion and Conclusion

In response to our first research question about network structural signatures, we found an egalitarian structure of participation where participants were eager to invite and respond to others (i.e., most participants engaged in the different self-assembly team mechanisms). In contrast to prior work [15, 24], we did not find that participants' invitations were concentrated on any specific participants of this study, though their rejection was unevenly distributed. This interaction structure can be explained by the team formation system design: all participants were able to reach out to others directly, without intermediates, but had to evaluate which invitations to reject. This may not directly mirror real contexts, where potential connections are often socially distanced by more than one degree of separation, and the role of brokers or popular individuals can be substantial. Digital platforms enable more egalitarian network structures in which all users are able to search for and find others [6].

To answer which individuals make the most appealing teammates (RQ2) and which factors determine acceptance and rejection (RQ3), we analyzed structural signatures at the individual and relational level. Our results show that these factors relied on the sender and receiver roles. Certain factors are relevant in all self-assembly mechanisms, whereas others are only important at certain moments of the team selection interaction. As previous studies similarly have demonstrated, a receiver's competence was influential in every stage of the team formation process. Further, those whose invitations were rejected were often the least skilled participants. Experienced leaders also garnered more frequent and positive reception. Surprisingly, participants' acceptances were not highly influenced by their frequent appearance in the search system. Our analysis found

that highly competent participants were more likely to send invitations and less likely to be rejected. Strong leaders were also more likely to be accepted and not ignored. People with high psychological collectivism, who intrinsically prioritize teamwork, were found to send more invitations but were accepted less often, likely because of the high numbers of invitations they sent. As was expected, agreeable and open people were more likely to send invitations, but unexpectedly, extroverted people were less likely to send invitations.

Our results must be interpreted cautiously because of the following limitations. First, we did not assess background variables such as ethnicity, nationality, or religion that may have affected results. More case studies in other environments could assess the generalizability of the team assembly factors identified in this study. Second, we relied on users' self-reported skills, which may not be accurate. Future studies may consider peer-evaluations as a way to confirm others' expertise. Third, as would be true of any platform, many design features likely to affect behavior. Lastly, we did not control students' interactions during the lectures and outside of them, meaning some teams may have formed because students agreed to do so offline, but that was not possible to measure.

In summary, we explored through a network perspective the phenomenon of choosing, accepting, and rejecting team members at the genesis of the process of self-assembling into teams. By conducting a study of team formation among 780 participants who formed 160 teams, we discovered which factors most strongly influenced their decisions. We found network structures, previous relationships, and individual attributes all influenced preferences thereby shaping who ultimately ends up working with whom.

References

1. Aldrich, H.E., Kim, P.H.: Small worlds, infinite possibilities? How social networks affect entrepreneurial team formation and search. *Strat. Entrep. J.* **1**(1–2), 147–165 (2007). <https://onlinelibrary.wiley.com/doi/abs/10.1002/sej.8>
2. Burt, R.S.: Structural holes and good ideas. *Am. J. Sociol.* **110**(2), 349–399 (2004). <https://doi.org/10.1086/421787>
3. Burt, R.S., et al.: Brokerage and Closure: An Introduction to Social Capital. Oxford University Press, Oxford (2005)
4. Contractor, N.S., Wasserman, S., Faust, K.: Testing multitheoretical, multilevel hypotheses about organizational networks: an analytic framework and empirical example. *Acad. Manag. Rev.* **31**(3), 681–703 (2006). <https://doi.org/10.5465/amr.2006.21318925>
5. Donnellan, M.B., Oswald, F.L., Baird, B.M., Lucas, R.E.: The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychol. Assess.* **18**(2), 192 (2006)
6. Faraj, S., Johnson, S.L.: Network exchange patterns in online communities. *Organ. Sci.* **22**(6), 1464–1480 (2011)
7. Ferris, G.R., Treadway, D.C., Kolodinsky, R.W., Hochwarter, W.A., Kacmar, C.J., Douglas, C., Frink, D.D.: Development and validation of the political skill inventory. *J. Manag.* **31**(1), 126–152 (2005)

8. Gilley, J.W., Morris, M.L., Waite, A.M., Coates, T., Veliquette, A.: Integrated theoretical model for building effective teams. *Adv. Dev. Hum. Resour.* **12**(1), 7–28 (2010). <https://doi.org/10.1177/1523422310365309>
9. Gómez-Zará, D., Andreoli, S., DeChurch, L., Contractor, N.: Discovering collaborators online: assembling interdisciplinary teams online at an Argentinian university. *Cuadernos.info* (44), 21–41 (2019). <https://doi.org/10.7764/cdi.44.1575>
10. Gómez-Zará, D., Paras, M., Twyman, M., Lane, J.N., DeChurch, L.A., Contractor, N.S.: Who would you like to work with? In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, pp. 659:1–659:15. ACM, New York (2019). <http://doi.acm.org/10.1145/3290605.3300889>
11. Granovetter, M.S.: The strength of weak ties. *Am. J. Sociol.* **78**(6), 1360–1380 (1973). <http://www.jstor.org/stable/2776392>
12. Hahn, J., Moon, J.Y., Zhang, C.: Emergence of new project teams from open source software developer networks: impact of prior collaboration ties. *Inf. Syst. Res.* **19**(3), 369–391 (2008). <https://pubsonline.informs.org/doi/abs/10.1287/isre.1080.0192>
13. Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., Krivitsky, P.N., Morris, M.: ERGM: Fit, Simulate and Diagnose Exponential-Family Models for Networks. The Statnet Project (<http://www.statnet.org>) (2018). <https://CRAN.R-project.org/package=ergm>. R package version 3.9.4
14. Harris, A.M., Gómez-Zará, D., DeChurch, L.A., Contractor, N.S.: Joining together online: the trajectory of CSCW scholarship on group formation. *Proc. ACM Hum. Comput. Interact.* **3**(CSCW), 148:1–149:27 (2019). <https://doi.org/10.1145/3359250>
15. Hinds, P.J., Carley, K.M., Krackhardt, D., Wholey, D.: Choosing work group members: balancing similarity, competence, and familiarity. *Organ. Behav. Hum. Decis. Process.* **81**(2), 226–251 (2000). <http://www.sciencedirect.com/science/article/pii/S0749597899928753>
16. Jackson, C.L., Colquitt, J.A., Wesson, M.J., Zapata-Phelan, C.P.: Psychological collectivism: a measurement validation and linkage to group member performance. *J. Appl. Psychol.* **91**(4), 884 (2006)
17. Leifeld, P., Cranmer, S.J., Desmarais, B.A.: Temporal exponential random graph models with btergm: estimation and bootstrap confidence intervals. *J. Stat. Softw.* **83**(6) (2018). <http://eprints.gla.ac.uk/139203/>
18. Lim, B.C., Klein, K.J.: Team mental models and team performance: a field study of the effects of team mental model similarity and accuracy. *J. Organ. Behav.* **27**(4), 403–418 (2006). <https://onlinelibrary.wiley.com/doi/abs/10.1002/job.387>
19. Mumford, M.D., Baughman, W.A., Threlfall, K.V., Uhlman, C.E., Costanza, D.P.: Personality, adaptability, and performance: performance on well-defined problem solving tasks. *Hum. Perform.* **6**(3), 241–285 (1993)
20. Osterman, P.: Skill, training, and work organization in american establishments. *Ind. Relat. J. Econ. Soc.* **34**(2), 125–146 (1995)
21. Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph (p^*) models for social networks. *Soc. Netw.* **29**(2), 173–191 (2007). Special Section: Advances in Exponential Random Graph (p^*) Models
22. Rudman, L.A., Goodwin, S.A.: Gender differences in automatic in-group bias: Why do women like women more than men like men? *J. Pers. Soc. Psychol.* **87**(4), 494 (2004)
23. Tierney, P., Farmer, S.M.: Creative self-efficacy: its potential antecedents and relationship to creative performance. *Acad. Manag. J.* **45**(6), 1137–1148 (2002)

24. Wagner, C.S., Leydesdorff, L.: Network structure, self-organization, and the growth of international collaboration in science. *Res. Policy* **34**(10), 1608–1618 (2005)
25. Wang, J., Hicks, D.: Scientific teams: self-assembly, fluidness, and interdependence. *J. Inf.* **9**(1), 197–207 (2015). <http://www.sciencedirect.com/science/article/pii/S1751157714001187>
26. Yilmaz, M., O'Connor, R.V., Colomo-Palacios, R., Clarke, P.: An examination of personality traits and how they impact on software development teams. *Inf. Softw. Technol.* **86**, 101–122 (2017). <http://www.sciencedirect.com/science/article/pii/S095058491730040X>
27. Yon, G.G.V., de la Haye, K.: Exponential random graph models for little networks. arXiv preprint [arXiv:1904.10406](https://arxiv.org/abs/1904.10406) (2019)
28. Zellmer-Bruhn, M.E., Maloney, M.M., Bhappu, A.D., Salvador, R.B.: When and how do differences matter? An exploration of perceived similarity in teams. *Organ. Behav. Hum. Decis. Process.* **107**(1), 41–59 (2008). <http://www.sciencedirect.com/science/article/pii/S0749597808000113>
29. Zhu, M., Huang, Y., Contractor, N.S.: Motivations for self-assembling into project teams. *Soci. Netw.* **35**(2), 251–264 (2013). <http://www.sciencedirect.com/science/article/pii/S0378873313000166>, special Issue on Advances in Two-mode Social Networks



Predicting Movies' Box Office Result - A Large Scale Study Across Hollywood and Bollywood

Risko Ruus and Rajesh Sharma^(✉)

Institute of Computer Science, University of Tartu, Tartu, Estonia
risko.ruus,rajesh.sharma@ut.ee

Abstract. Predicting movie sales figures has been a topic of interest for research for decades since every year there are dozens of movies which surprise investors either in a good or bad way depending on how well the film performs at the box office compared to the initial expectations. There have been past studies reporting mixed results on using movie critics reviews as one of the sources of information for predicting the movie box office outcomes. Similarly using social media as a predictor of movie success has been a popular research topic. We analyze the Hollywood and Bollywood movies from three years, which belong to two different geo as well as cultural locations. We used Twitter for collecting the wisdom of the crowd features (4.3 billion tweets, 1.41 TB in compressed size) and used movie critics review scores from movie review aggregator sites *Metacritic* and *SahiNahi* for Hollywood and Bollywood movies respectively. In addition, we also used metadata about movies such as budget, runtime, etc. for the prediction task. Using three different machine learning algorithms, we investigated this problem as a regression problem to predict the movie opening weekend revenues. Compared to past studies which have performed their analysis on much smaller datasets, we performed our study on a total of 533 movies. In addition to r^2 , we measured the quality of our models using MAPE and we find out that a model (Random Forest) based on all the three features (Metadata, Critics, Twitter) gives the best results in our analysis.

Keywords: Box office forecasting · Machine learning · Twitter

1 Introduction

Hundreds of movies are released every year in the world. However, not every movie turns out to be a commercial success. For example, only three or four major movies out of every ten major Hollywood movies produced are profitable [1]. Forecasting the box office results has been a big concern for the movie industry as early box office predictions help to make vital decisions regarding marketing budget allocation and distribution. Equally important is determining the best screen allocation for a movie in each country since empty seats mean bad

business for movie studios and cinemas alike. However, past studies have shown it is difficult to predict the tastes of moviegoers [2–4] and subsequently forecasting the box office results has been a big concern for the movie industry.

Litman was the first to study multivariate regression models [5, 6] for predicting the box office outcome of movies. Predictor variables considered in such research include the number of theaters the movie is scheduled to be released in, parental rating and the budget of the film. Many researchers consider predicting commercial movie success as a classification problem. For example in [7–9] movies are classified into different categories usually ranging from a flop to a blockbuster. These segments are created by using the movie production budget as an estimated figure for calculating how much a movie should make to earn its production costs back. The problem with this approach is that while recently many studios have started to reveal their film production budgets, the money spent on marketing is not disclosed and can influence the actual profitability of the movie significantly. Also as mentioned in [10], star actors are often paid a percentage of the movie profits and their salaries might not be included in the movie production budget figures making the movie budget deceptively low. For these reasons we have followed the example of studies such as [11–13] and consider predicting commercial movie success as a regression problem and predict the amount of money a movie is expected to earn after its opening weekend.

Most of the previous studies involving predicting movie success ahead of its release have worked by exploring either social media platforms such as Twitter [14], Wikipedia [15], Facebook [16], Google search queries [17] or have only analyzed movie expert's reviews [3, 18–20].

Social media content can be thought of as a very large collection of collective wisdom. When asking the right questions from such data, it is possible to make predictions about future outcomes and the question we will be asking is about predicting the box office outcome of upcoming movie releases [14]. In comparison, movie critics reviews refer to the views expressed by a smaller group of domain experts. In this work, we followed a holistic approach and used both social media platform (Twitter in our case) and movie critics for predicting the box office outcome of the movies. The models can be used by stakeholders, including distributors and movie theatre operators to make improved financial decisions when promoting the film at the *critical period*¹ of its release.

This work is an empirical study, which involves collecting all the necessary data for building prediction models for the Hollywood and Bollywood movies released between April 2015 and April 2018. For model building, we used three different types of features. The first we call as *Metadata* which includes general movie information e.g. budget and opening theatre count. The second set of features are called *Twitter* features which uses the hourly tweet rate from two weeks before the film's release and the sentiment score of the movie tweets. The

¹ We use the same definition for the critical period as [14]. It is defined to be between a week before the movie is released until two weeks from its release date. This is usually the time when most of the promotional budget is being spent on various forms of advertising.

third set of features, *Critics*, takes input from movie expert reviews. We evaluate prediction results using Linear Regression, Random Forest and XGBoost machine learning algorithms.

This paper has following contributions:

1. **Wisdom of the crowd and experts:** Our empirical study shows that people's collective wisdom (gathered from Twitter) when combined with the critics' review and metadata about movies can help to predict movie opening weekend box office results better when using these sources of information separately.
2. **Large scale study:** To the best of our knowledge this research is made on the largest amount of Hollywood and Bollywood movies.
3. **Hollywood & Bollywood:** The work offers a unique cross-cultural comparison of box office predictions for Hollywood and Bollywood - the two of the world's biggest movie markets.

Rest of the paper is organized as follows. In Sect. 2 we give a brief overview of previous related research regarding predicting movie box office results. Section 3 focus on describing the data collection process for predicting the final results. An overview of our empirical results is in Sect. 4. Finally, Sect. 5 describes our overall contribution and proposes some directions for future research.

2 Related Works

In this section, we provide an overview of the past research done on predicting the success of movies. We look at works which have used either social media or critics movie reviews as a source for predicting box office revenue.

2.1 From Social Web Platforms

Before the rise of the internet most of the dependent variables used for predicting movie box office outcome, have been based on movie metadata e.g., its genre, parental rating and actors which as reported by [21] can explain approximately 60% of the variances.

With the rise of dedicated communities for movie lovers, blogs and various web services, researchers have been looking for additional sources of information, which could help predict the movie economical success even better. For example, [12] were able to predict box office revenue from 600,000 blog entries with a relative error of 26.21%. Authors of [22] have compared the predictive power of tweet sentiment analysis and online movie review sites such as *imdb* and *Rotten Tomatoes*² and find that Twitter users are more positive in their reviews compared to the dedicated review site's ratings.

Some studies like [8] have compared the prediction sources of different web resources and social networks, namely IMDb, Twitter, and YouTube. They find

² <https://rottentomatoes.com>.

that the popularity of the leading actress estimated by the followers count the actress has on Twitter is a strong predictor, but the sentiment score from movie trailer comments does not help to determine the financial success of a movie.

In a novel study, [14] have shown that data from Twitter, in particular, the average hourly tweet rate and sentiment analysis of the tweets can be used to predict movie box office outcomes using a simple linear regression model $r^2(t) = 0.98$ at the release night of the movie). However [15] does point out in Fig. 5 of their work that the paper of [14] achieves such a high score because most of the 24 movies considered are commercial successes, which the model is capable of predicting better than movies with low or moderate success.

In their work on 312 movies [15] show that movie box office performance can be estimated from the activity levels of Wikipedia articles about the movie before its release. Similarly to Wikipedia activity levels, Facebook official movie fan page activity is used as a prediction feature in [16]. Predictions from social media can be made not only about movie's financial success as [23] were able to rate movies very close to their IMDb star rating using tweets from Twitter and comments from YouTube. For predicting Academy Award nominations and movie box office results, [24] show successful results using movie comments from IMDb users as a possible source of information. A whitepaper from Google [17] on 99 movies released in 2012 shows that Google search volume explains 70% of the variance in the opening weekend box office performance of the film.

Research involving predicting movie profitability is not only limited to Hollywood releases. For example, Korean researchers in [25] have studied their local market on a dataset of 212 domestic movies using metadata and features from multiple social media networks. Similarly predicting movie box office success on the Chinese domestic market has been researched by [26] using 57 movies with 5 million tweets collected from the Sina Weibo microblog³. The only previous study on predicting the box office results of Bollywood movies that uses features from social media is done only on 14 movies by [27].

2.2 From Expert Movie Reviews

Predicting movie box office outcome using critic reviews as a source has attracted less attention from researchers compared to using social media platforms. The authors of [3] look at expert reviews and find confirmation to the common belief that positive reviews help box office performance and bad reviews have a negative impact on the sales. Some research has also done on the textual data of critic's movie reviews like [11] who use movie earnings text analysis on pre-release reviews and metadata features available before movie's release for predicting the opening weekend box office results.

In comparison to above, in his study on movies released in 2003 in the U.S. [19] finds that Metacritic.com scores do not have a strong relationship with the gross earnings of the films. Rotten Tomatoes ratings are used by [28] to find the critic scores to have a positive and significant effect on the movie box office

³ <https://www.weibo.com>.

revenue although it is much smaller when compared to independent variables like the number of opening screens and the budget of the movie.

The aggregate movie critic score impact on movie box office revenue is studied by [18], and they find it to have a small positive effect. However, they do report that the impact is more influential on the total gross revenue of the movie and weaker for predicting the opening weekend earnings. However, authors of [29] find in similar to [3] and in opposite to [18] in their study focusing on individual movie critics, that critics act as more influencers rather than predictors. For a Bollywood movies study, authors of [20] look at both the online user-generated and the expert reviews from daily newspapers and find that volume and valence from both sources have had a positive effect on the financial success of movies.

3 Dataset Description

In this section, we describe various sources of the datasets being used for analysis.

3.1 Movie Selection

We considered Hollywood and Bollywood movies released between April 10th, 2015 and April 6th, 2018. For the sake of consistency, we focused only on the films that are released on Fridays. For Hollywood movies, we only included movies, which had a wide release from its first release day that is a film which runs in 600 or more cinemas [30]. If a movie had a limited release initially, but later went into a wide release then we did not include that in our work. For Bollywood movies, since we did not find any definition for a wide release, thus, we did not apply any such selection criteria for them.

3.2 Tweets Collection

Our tweets had two main sources. The first one being the Twitter itself. Following the approach of recent papers like [27, 31] we used the unique hashtags to match a tweet to a movie. This approach has the benefit of being able to find tweets about a movie with a non-unique title like *Sisters* when people have marked them with a hashtag such as #SistersMovie⁴. When inspecting the official Twitter pages of such films, we found that the movie studios often pick the main hashtag for the movie and use it consistently in their marketing campaigns. When such tweets reach their audience, then they tend to use the same hashtag in their own tweets. In our work, we also decided to identify tweets by the hashtags that were used most often to refer to the movie the tweet was about.

For historical tweets, monthly dumps of the *Spritzer* version of the Twitter Streaming API by Archive Team⁵ were used as a second source of tweets. Authors in [32] studied the Spritzer version of the Twitter stream on a number of datasets

⁴ <https://twitter.com/sistersmovie>.

⁵ <https://archive.org/details/twitterstream>.

to see if there is any sampling bias in the stream. They found these dumps to be suitable for conducting research experiments and the sampling ratio measured on their datasets was on an average of 0.95%. The total size of our tweet set downloaded from archive.org was 1.41 TB in compressed format containing 4.3 billion tweets.

After gathering and validating the tweets, we had to find the Hollywood and Bollywood movies released during these years and look up the right hashtags for each film from the web. For finding the relevant Hollywood and Bollywood movie release dates we used the Box Office Mojo and Box Office India websites and collected the movies which release date fitted into our historical tweet set timeline. Finding hashtags for the films was again a manual process of looking at the official Twitter pages of the movie and searching for the most popular hashtags people had been using when tweeting about the film. If a tweet did not contain any hashtags or did not contain hashtags about films, then we skipped processing it. Further, if the tweet included any movie hashtags we were interested in, then the number of movies the tweet was about was calculated. If the tweet had hashtags for multiple distinct films, then we discarded the tweet since we could not determine, which movie the tweet was mostly about. Finally, the tweet referring to a single film was stored and assigned to the movie. A total of 281,322 tweets mentioning hashtags for Hollywood and Bollywood movies were extracted from the dataset containing all the tweets.

3.3 From Expert Review Aggregator Sites

Critics' movie reviews are usually published a few days before or on the public release date of the movie, which leaves enough time to influence the movie-goers decision whether to go and see the film or not. Similar to previous work done in studies [19, 33, 34], we decided to use movie review aggregator scores and review counts as an input variable for predicting the box office outcome. For Hollywood movies, we collected movie review scores from the critic score aggregator website *Metacritic*⁶ and for Bollywood, we gathered the review scores from the movie info portal *SahiNahi*⁷. The main reason we picked these review sites was that compared to many competitor review sites we investigated, these two had scores available for the most movies in our dataset. Also as mentioned before, *Metacritic* had been used in a number of past studies. Although we did not find any articles, which had used *SahiNahi* scores as an input variable for box office score predictions, but at the same time we did not find any other Bollywood movie critic aggregate site scores having been used either.

3.4 From Movie Revenue Information Sites

General movie information e.g. runtime, genre and the box office results for Hollywood movies was collected from *Box Office Mojo*⁸ website which is often used

⁶ <https://www.metacritic.com>.

⁷ <https://www.sahinahi.com/>.

⁸ <https://www.boxofficemojo.com/>.

as a source of financial movie information in similar studies to ours [14, 15]. In the case of Bollywood, we collected the data from movie information portal *Box Office India*⁹. Since for Bollywood movies the parental rating information was not available from *Box Office India*, we gathered the information from Times of India daily news website¹⁰ which includes movie reviews for most of the Bollywood movies. For us, the most interesting data points were the number of theatres the movie was released in, the opening weekend gross domestic income and the budget of the movie.

3.5 Data Cleaning

Unfortunately we did not end up having all the features for every movie we collected available. For example, for some Hollywood and Bollywood movies, the budget info had not been disclosed. Because we use the budget as one of the predictor variables then movies with no budget information were discarded from further study. Also for a few movies like *The Bounce Back*, the Metascore was not available because there were not enough critic reviews about the movie available for Metacritic to generate an aggregated score. After the cleaning was applied, there were 347 Hollywood and 186 Bollywood movies in our study for a combined total of 533 movies.

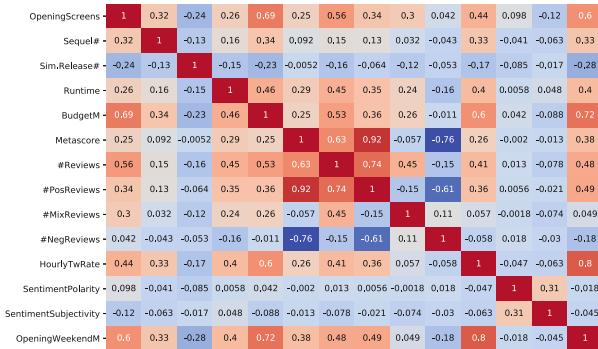


Fig. 1. Feature correlations for the Hollywood dataset

3.6 Exploratory Data Analysis

The heatmaps on Figs. 1 and 2 show numeric feature correlation information, which can give us strong hints for understanding which variables could be important for predicting the opening weekend box office. In case of Hollywood on Fig. 1 the top three positively correlated features are the number of tweets (0.80), budget (0.72) and the number of theaters (0.60), which all indicate quite strong

⁹ <https://boxofficeindia.com/>.

¹⁰ <https://timesofindia.indiatimes.com/entertainment/hindi/movie-reviews>.

correlations. We expect these features to be also useful for regression models for predicting the movie revenue. The top three negatively correlated features are the number of releases on the same weekend (-0.28), the number of negative reviews (-0.18) and tweet sentiment subjectivity (-0.045). The negative correlation here does not necessarily mean that a feature will not be useful for making box office predictions. On the opposite, the moderate negative correlation of releases on the same weekend variable hints at the expected outcome that more movies opening at the same weekend compete for the same general population to go and see their film and the more movies there are to choose from the less they make on average compared to films that have none or few competitors. It can also hint that sometimes smaller movies try not to compete with big blockbuster movie releases and will release on a different weekend to avoid the strong competition from the hit movies. The weak correlation with the negative review count also shows that the more negative reviews the film has, the less money it is likely to make.

OpeningScreens	1	-0.4	0.64	0.84	0.24	0.38	0.3	0.22	0.61	-0.064	0.061	0.87
Sim,Release#	-0.4	1	-0.21	-0.36	-0.037	-0.047	-0.012	-0.058	-0.23	0.047	0.033	-0.41
Runtime	0.64	-0.21	1	0.63	0.28	0.35	0.34	0.11	0.44	-0.0092	0.058	0.58
BudgetM	0.84	-0.36	0.63	1	0.32	0.36	0.35	0.12	0.72	-0.069	0.07	0.87
CriticRating	-0.24	-0.037	0.28	0.32	1	0.41	0.73	-0.29	0.28	0.15	0.2	0.39
#Reviews	0.38	-0.047	0.35	0.36	0.41	1	0.78	0.59	0.31	0.068	0.1	0.34
#PosReviews	0.3	-0.012	0.34	0.35	0.73	0.78	1	-0.054	0.38	0.15	0.2	0.43
#NegReviews	0.22	-0.058	0.11	0.12	-0.29	0.59	-0.054	1	0.005	-0.083	-0.094	-0.02
HourlyTwRate	-0.61	-0.23	0.44	0.72	0.28	0.31	0.38	0.005	1	-0.034	0.16	0.68
SentimentPolarity	-0.064	0.047	-0.0092	-0.069	0.15	0.068	0.15	-0.083	-0.034	1	0.52	-0.037
SentimentSubjectivity	0.061	0.033	0.058	0.07	0.2	0.1	0.2	-0.094	0.16	0.52	1	0.086
OpeningWeekendM	0.87	-0.41	0.58	0.87	0.39	0.34	0.43	-0.02	0.68	-0.037	0.088	1

Fig. 2. Feature correlations for the Bollywood dataset

4 Prediction Analysis

The aim of this work is not to identify the best model based on r^2 but rather based on MAPE. We first report model performance using only the Metadata features and then we create also models where we combine Metadata with Critics and Twitter features. We run the experiment using Linear Regression, Random Forest and XGBoost algorithms and use k-fold cross-validation for calculating the average model performance using $k=10$ folds. We use the default hyperparameters and do not perform any parameter tuning in this experiment to optimize for better scores.

4.1 Hollywood

Table 1 (Columns 2 and 3) shows the performance metrics for different models on the Hollywood movies dataset. We can see that the models with the most

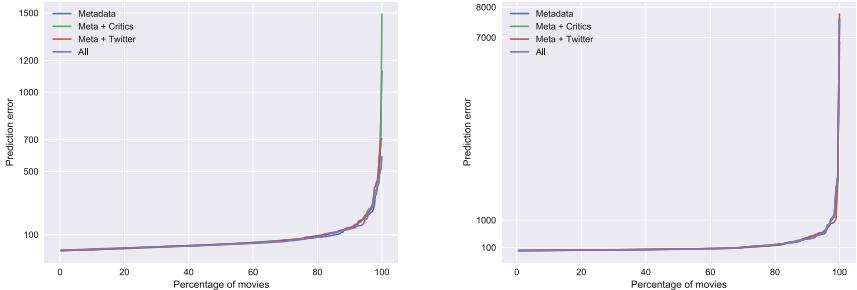
features report also the best performance. On average the model using Random Forest algorithm and all the available features predicts with roughly 64% of error. To understand this score and explore it further we gathered the predicted results from each fold and then plotted all the movie predictions by their absolute prediction error. Figure 3a illustrates this experiment and shows that more than 80% of movies have an error of less than 100%, but there are a few outliers, which the model predicted with a very large error. Interestingly, the models with critics features ($0.767 r^2$, 69% MAPE) perform worse compared to the Twitter counterpart ($0.739 r^2$, 64% MAPE) in case of the Random Forest algorithm. After inspecting the few movies, which had large errors between the predicted and actual values, we noticed that most of such movies were independent films such as *The Bronze (2015)*, which made \$400k during its opening weekend. Our models predicted it would make much more since it was the least earning movie in the dataset, but the dataset did not include many similar movies to learn how to predict revenues so low.

4.2 Bollywood

Similar to Hollywood we can see for Bollywood from Table 1 that Random Forest algorithm achieved the best (80% MAPE) performance with all features included, but interestingly the best r^2 score (0.863) was reported by a Linear Regression model where Metadata features were used together with Critics features. However, this model with best r^2 achieved the worst performance (184% MAPE). This illustrates that when comparing model performance, picking the

Table 1. Model performance for Hollywood and Bollywood movies. Within a column, boldface shows the best result for a metric.

Model		Hollywood		Bollywood	
		r^2	MAPE	r^2	MAPE
Metadata	Random Forest	0.700	71.524	0.759	108.268
	XGBoost	0.702	73.625	0.831	115.203
	Linear Regression	0.464	167.757	0.795	183.336
Metadata \cup Critics	Random Forest	0.767	69.183	0.793	91.012
	XGBoost	0.709	75.782	0.862	90.506
	Linear Regression	0.527	164.487	0.863	184.479
Metadata \cup Twitter	Random Forest	0.739	64.081	0.803	89.338
	XGBoost	0.724	65.499	0.849	86.507
	Linear Regression	0.686	111.766	0.783	177.653
All	Random Forest	0.777	64.007	0.777	80.011
	XGBoost	0.748	65.435	0.862	86.34
	Linear Regression	0.715	111.745	0.855	184.292



(a) Hollywood opening weekend box office absolute prediction errors using Random Forest algorithm

(b) Bollywood opening weekend box office absolute prediction errors using Linear Regression

Fig. 3. Prediction errors

best model based only by r^2 score might not lead to the best performing model in practice. Figure 3b shows the prediction errors using this model resulted in. There are outliers with prediction errors nearly 8000% such as the movie *Uvaa* which affects the MAPE value a lot for all four models with different feature combinations. Since we included also some less popular movies, but their overall distribution in the dataset was not very high, our models do not predict low box office income accurately and tend to overestimate the predictions.

4.3 Hollywood vs. Bollywood

Since Hollywood and Bollywood movie markets are quite different we cannot compare the prediction errors using metrics like MAE and RMSE, but the r^2 and MAPE values are still comparable. In our experiment on the Bollywood dataset, the r^2 values are higher, which indicates that more variance in the predicted opening weekend revenue is explained by the dependent variables we used for predicting. However, the reported MAPE values are larger for Bollywood than for Hollywood models. This can be explained by a few hard-to-predict outlier Bollywood movies, which have significantly larger errors than the outliers in case of Hollywood. In the case of Bollywood dataset, outliers have also a bigger total impact since there were twice as many movies for Hollywood in our study. The difference between MAPE errors for Critics and Twitter models is larger for Hollywood movies than it is for Bollywood. For Bollywood movies, we are looking only at tweets in English and did not consider tweets in Hindi. This means we are capturing a larger sample of Tweets for Hollywood movies which benefits the performance of Hollywood Twitter-based models compared to Bollywood.

5 Conclusion and Future Work

Movie sales prediction has been an interest to many researchers as they often carry huge investments. In this work, we investigated the movie sales prediction problem from two different perspectives. Firstly by analyzing the reviews given by movie critics. Secondly, we focus on the wisdom of the crowd, collected using social media platform, Twitter.

Our reported r^2 scores are not as high as some of the earlier works have reported, but it is worth noting that most of such papers use statistical Ordinary Least Squares method without cross-validation on a small set of movies with similar budgets and coming from major studios. In addition to reporting r^2 scores, we report MAPE values. The results of our prediction analysis show that adding more features will generally improve model performance. Similarly, in both Hollywood and Bollywood dataset experiments, roughly 80% of movies our models were able to predict with 100% of error or less.

The number of movies with their related tweets was the highest we have seen studied so far, but future work should include more movies to build even more effective machine learning models. In our current work, we used the movie critic aggregator scores as a general sentiment polarity score for the movies. For future work we propose to extract different aspect-level sentiment information from movie reviews similar to [35]. Separate aspect-level sentiment scores e.g. for acting, directing, music could all be used as features for the prediction model.

Acknowledgments. This work has been supported in part by the SoBigData project (under grant agreement no. 654024).

References

1. Vogel, H.L.: Entertainment Industry Economics: A Guide for Financial Analysis. Cambridge University Press, Cambridge (2014)
2. Sawhney, M.S., Eliashberg, J.: A parsimonious model for forecasting gross box-office revenues of motion pictures. *Mark. Sci.* **15**(2), 113–131 (1996)
3. Basuroy, S., Chatterjee, S., Abraham Ravid, S.: How critical are critical reviews? The box office effects of film critics, star power, and budgets. *J. Mark.* **67**(4), 103–117 (2003)
4. Liu, Y.: Word of mouth for movies: its dynamics and impact on box office revenue. *J. Mark.* **70**(3), 74–89 (2006)
5. Litman, B.R.: Predicting success of theatrical movies: an empirical study. *J. Popul. Cult.* **16**(4), 159–175 (1983)
6. Litman, B.R., Kohl, L.S.: Predicting financial success of motion pictures: the '80s experience. *J. Media Econ.* **2**(2), 35–50 (1989)
7. Sharda, R., Delen, D.: Predicting box-office success of motion pictures with neural networks. *Expert Syst. Appl.* **30**(2), 243–254 (2006)
8. Apala, K.R., Jose, M., Motnam, S., Chan, C.-C., Liszka, K.J., de Gregorio, F.: Prediction of movies box office performance using social media. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1209–1214. IEEE (2013)

9. Quader, N., Gani, M.O., Chaki, D., Ali, M.H.: A machine learning approach to predict movie box-office success. In: 2017 20th International Conference of Computer and Information Technology (ICCIT), pp. 1–7. IEEE (2017)
10. Simonoff, J.S., Sparrow, I.R.: Predicting movie grosses: winners and losers, block-busters and sleepers. *Chance* **13**(3), 15–24 (2000)
11. Joshi, M., Das, D., Gimpel, K., Smith, N.A.: Movie reviews and revenues: an experiment in text regression. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 293–296. Association for Computational Linguistics (2010)
12. Abel, F., Diaz-Aviles, E., Henze, N., Krause, D., Siehndel, P.: Analyzing the blogosphere for predicting the success of music and movie products. In: 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 276–280. IEEE (2010)
13. Hon, L.Y.: Expert versus audience's opinions at the movies: evidence from the North-American box office. *Mark. Bull.* **25**, 1–22 (2014)
14. Asur, S., Huberman, B.A.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 492–499. IEEE (2010)
15. Mestyán, M., Yasseri, T., Kertész, J.: Early prediction of movie box office success based on wikipedia activity big data. *PloS One* **8**(8), e71226 (2013)
16. Tang, W.-H., Yeh, M.-Y., Lee, A.J.T.: Information diffusion among users on Facebook fan pages over time: its impact on movie box office. In: 2014 International Conference on Data Science and Advanced Analytics (DSAA), pp. 340–346. IEEE (2014)
17. Panaligan, R., Chen, A.: Quantifying movie magic with Google search. Google Whitepaper—Industry Perspectives+ User Insights (2013)
18. Eliashberg, J., Shugan, S.M.: Film critics: influencers or predictors? *J. Mark.* **61**, 68–78 (1997)
19. King, T.: Does film criticism affect box office earnings? Evidence from movies released in the us in 2003. *J. Cult. Econ.* **31**(3), 171–186 (2007)
20. Niraj, R., Singh, J.: Impact of user-generated and professional critics reviews on Bollywood movie success. *Australas. Mark. J. (AMJ)* **23**(3), 179–187 (2015)
21. Chang, B.-H., Ki, E.-J.: Devising a practical model for predicting theatrical movie success: focusing on the experience good property. *J. Media Econ.* **18**(4), 247–269 (2005)
22. Wong, F.M.F., Sen, S., Chiang, M.: Why watching movie tweets won't tell the whole story? In: Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks, pp. 61–66. ACM (2012)
23. Oghina, A., Breuss, M., Tsagkias, M., de Rijke, M.: Predicting IMDb movie ratings using social media. In: European Conference on Information Retrieval, pp. 503–507. Springer (2012)
24. Krauss, J., Nann, S., Simon, D., Gloor, P.A., Fischbach, K.: Predicting movie success and academy awards through sentiment and social network analysis. In: ECIS, pp. 2026–2037 (2008)
25. Kim, T., Hong, J., Kang, P.: Box office forecasting using machine learning algorithms based on SNS data. *Int. J. Forecast.* **31**(2), 364–390 (2015)
26. Liu, T., Ding, X., Chen, Y., Chen, H., Guo, M.: Predicting movie box-office revenues by exploiting large-scale social media content. *Multimedia Tools Appl.* **75**(3), 1509–1528 (2016)

27. Gaikar, D.D., Marakarkandy, B., Dasgupta, C.: Using Twitter data to predict the performance of Bollywood movies. *Ind. Manag. Data Syst.* **115**(9), 1604–1621 (2015)
28. Brewer, S.M., Kelley, J.M., Jozefowicz, J.J.: A blueprint for success in the US film industry. *Appl. Econ.* **41**(5), 589–606 (2009)
29. Boatwright, P., Basuroy, S., Kamakura, W.: Reviewing the reviewers: the impact of individual film critics on box office performance. *Quant. Mark. Econ.* **5**(4), 401–425 (2007)
30. BoxOfficeMojo: Bob Office Tracking By Time. <http://www.boxofficemojo.com/about/boxoffice.htm>. Accessed 15 Apr 2018
31. Shim, S., Pourhomayoun, M.: Predicting movie market revenue using social media data. In: 2017 IEEE International Conference on Information Reuse and Integration (IRI), pages 478–484. IEEE (2017)
32. Wang, Y., Callan, J., Zheng, B.: Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *ACM Trans. Web (TWEB)* **9**(3), 13 (2015)
33. Gopinath, S., Chintagunta, P.K., Venkataraman, S.: Blogs, advertising, and local-market movie box office performance. *Manag. Sci.* **59**(12), 2635–2654 (2013)
34. Hennig-Thurau, T., Houston, M.B., Walsh, G.: Determinants of motion picture box office and profitability: an interrelationship approach. *Rev. Manag. Sci.* **1**(1), 65–92 (2007)
35. Piryani, R., Gupta, V., Singh, V.K.: Movie prism: a novel system for aspect level sentiment profiling of movies. *J. Intell. Fuzzy Syst.* **32**(5), 3297–3311 (2017)



Using Machine Learning to Predict Links and Improve Steiner Tree Solutions to Team Formation Problems

Peter Keane^{1(✉)}, Faisal Ghaffar², and David Malone¹

¹ Maynooth University, Maynooth, Ireland

peter.keane.2014@mumail.ie, David.Malone@mu.ie

² Innovation Exchange, IBM Ireland, Dublin, Ireland

faisalgh@ie.ibm.com

Abstract. The team formation problem has existed for many years in various guises. One important problem in the team formation problem is to produce small teams that have a required set of skills. We propose a framework that incorporates machine learning to predict unobserved links between collaborators, alongside improved Steiner tree problems to form small teams to cover given tasks. Our framework not only considers size of the team but also how likely are team members are going to collaborate with each other. The results show that this model consistently returns smaller collaborative teams.

Keywords: Team formation · Link prediction · Steiner tree

1 Introduction

Online team formation and patent collaboration within enterprises are vast and widely studied fields [6, 13, 15]. Team formation is the problem of identifying a set of individuals with skills that are required by a task for its completion. One can image a setting where a stream of incoming tasks and a system in automatic fashion is finding out a number of individuals who can complete those tasks from a pool of interconnected community. Online social networks (e.g., Facebook, LinkedIn, Enterprise networking tools etc.) in our personal and professional lives provide us opportunity to connect with each other and work together on common tasks. At the same time it is quite challenging to find a team of experts. The problem has been considered as NP-hard [6]. However, there have been approximate solutions to the problem and many of these rely on social or collaboration networks among individuals. The proposed solutions leverage techniques from graph theory, utilising topological features of social and business networks, combined with graph theory to examine the existing structures of teams, interactions, and collaborations within the enterprise [8, 11, 12, 16]. Others works have been depended on machine learning, to see what individual's characteristics likely to contribute in forming a link between colleagues and suggest people with these features collaborate [1, 7].

Most of the online team formation work [6, 13] focus on reducing the communicative cost among potential team members in a network. The *communicative cost* in a network is considered as a measure of how effectively team members can collaborate with each other [6]. There is a correlation between team size and cost, both communicative and financial [4, 9]. The problem that we focus on in this paper is producing small, ideally minimal teams, that have the required skills to complete a task. In producing a small team, we aim to keep costs down.

In this paper, we propose a framework which incorporates both individual's attributes as well as topological features from individual's network into machine learning link prediction task to improve the enhanced Steiner algorithm for team formation proposed in [6]. Our aim is to use machine learning to produce an augmented graph, containing both real and predicted links, before feeding it into the Enhanced Steiner Tree Algorithm [6], in order to return a minimal (weight) team which covers the necessary skill set. In doing so, we are aiming to combine previous work both on link prediction using machine learning [1] and work carried out using minimal spanning tree solutions [6]. We hope that by combining the two approaches we will overcome possible shortcomings of same. The Enhanced Steiner Tree algorithm provides good solutions to the team formation problem, but one possible shortcoming we point to is that it may sometimes neglect to consider isolated nodes or nodes that sit on unconnected components of the network.

We propose that, sometimes, it may be better to recommend collaboration between two inventors who aren't closely connected. There may be a valuable inventor who has all the required skills, and has parameters that are conducive to a good collaborative relationship with at least one member of the potential team. It may be more prudent to recommend this person to the team instead of bringing in many more inventors to cover the skills that this one inventor has. This will help in terms of finding small teams, and keep cost down.

To evaluate our scheme, we will show how we implemented it on a collaboration graph consisting of IBM Patent inventors. Our data is drawn from the US Patent Office (USPTO) data set. We evaluate the scheme in terms of the team sizes produced.

The paper is structured as follows. In Sect. 2 we introduce notation and discuss other work carried out in the area of team formation. Section 3 will give a high level outline of the model. A detailed explanation will be provided in Sect. 4 along with the method for evaluating our model. Finally, Sect. 5 will show our results. The conclusions are presented in Sect. 6.

2 Preliminaries

2.1 Mathematical Definitions and Notations

Some definitions will be laid out in this section. Full details of these definitions can be found in “Modern Graph Theory [2]” by Bollobás.

For the purpose of this paper, $G = (V, E)$, will denote a weighted graph, with a node/vertex set V and edge set E . In our case, the graph will be a

collaboration graph where the nodes will represent patent inventors and the edges collaborations. The weights, w_e , are assigned to each edge and represent the strength of the collaborations, where the higher the weight, the less strong the collaboration is. We will need to work with an induced subgraph of G based on the skills required for the task. This is a subgraph $G'(V', E')$ where $V' \subset V$ and $E' \subset E$ are the edges with both nodes in V' .

Steiner Tree problems, named after Jakob Steiner, are a combinatorial optimisation problem. They typically marry the problems of finding minimal spanning trees with the shortest path problem. For example, in [5] it is used to find “*a shortest network which spans a given set of points*.”

As input to the machine learning part of our scheme, we will use a number of simple graph theory quantities, such as clustering coefficient [14].

2.2 Related Work

We are motivated by forming small collaborative teams. In [9] the authors observe that smaller teams lead to lower communicative and co-ordination costs. They also explicitly state that “*Most of the software development cost is related to the programmers’ salaries*”. To this end, it seems prudent to recommend the smallest possible teams to cover necessary skills required for given tasks.

One of the primary algorithms we will rely on in this paper is the Enhanced Steiner Tree algorithm [6]. We also use their problem definition for the *Team Formation Problem*. That is to say, given a task T , individuals V with skills, and a Graph G representing the network between those individuals, we aim to return a minimal team which contains individuals that not only cover the skills required for T , but can also work effectively together.

We will also draw on work on bonding and bridging measurements as social capital [10] and collaborative ratio measurement [15]. These metrics are as inputs to our machine learning scheme in Sect. 4.3.

3 Proposed Framework

The steps involved in our proposed model are outlined in the high level block diagram shown in Fig. 1.

The scheme begins with *Raw Data*, where all of the raw patent data downloaded and summarised. We identified the number of inventors, to assess our graph size, and also how many different patent classes there, as these will represent our skills and determine the induced subgraphs.

At the *Pre-Processing* stage, we selected the subset of patents that we were interested in working with, the IBM patents, using the assignee information.

We next generate the *Skills & Inventor Network*. We consider the distribution of skills, (i.e. patent classes) across inventors, and created a cross-domain graph where each node represents an inventor, and an edge between them represents collaboration between inventors from different domains. We can also extract a subgraph where certain inventors have expertise in certain skills.

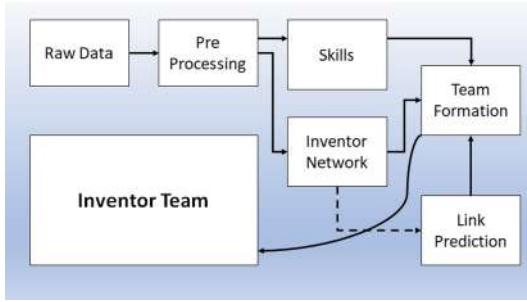


Fig. 1. Block diagram of proposed model

Link Prediction is the stage where machine learning is used to predict links based on certain parameters. At this point, the original inventor network is augmented with the false positives from the machine learning model.

Finally, the *Enhanced Steiner Tree* [6] is the algorithm that we use to form a small team which possesses the necessary skills to cover the skill requirement.

4 Scheme Implementation

In this section, we will lay out the steps involved in producing the graphs used for analysis, and the CSVs used for machine learning purposes.

4.1 The USPTO Database

Data was downloaded in TSV format from the USPTO PatentsView website¹, specifically the *patent*, *ipc*, *patent_inventor* tables and tables relating to company assignments.

Using the company assignee data for IBM, all the IBM patents were extracted. They were cross referenced with the other tables and Python's pandas package was used to merge the relevant data to produce a final working table which consisted of the following headings: *patent_id*, *inventor_id*, *full_class_id*. The full *full_class_id* used is the International Patent Classification².

4.2 Graph Creation

The main graph is our collaboration graph $G = (V, E)$, where V contains the inventors of IBM patents, and E contains any collaboration between the inventors on a patent. The extracted patent collaboration graph has 39,199 nodes and 158,279 edges with average degree of 8 and clustering coefficient of 0.4665. The weights assigned to each edge correspond to the strength of the collaboration

¹ <http://www.patentsview.org/download/>.

² <https://www.wipo.int/classifications/ipc/en/>.

Result: A Graph $G(V, E)$ where V is a patent holder, and E is collaboration between patent holders

Initialisation: Group the dataframe by unique patent ids, resulting in a Group with a title of patent id, and within the group are all of the inventor ids associated with that patent;

for Each Group do

```

if Patent ID Group Has Only One Unique Inventor then
    Add node with Inventor ID;
else
    for Each Pair of Inventors do
        if Pair is Not Connected then
            Create nodes for inventors and add edge of weight 1 between
            Inventors;
        else
            Add 1 to current edge weight;
    Invert all weight values;

```

Algorithm 1. Creating Collaboration Network

and it is calculated as inverse of the number of co-inventions to indicate the cost between two inventors.

This graph was created from the table generated by Algorithm 1. Python's pandas package was used for grouping the data. As noted above, patent classes are used to represent the skills required on a team. A very small scale example of such a graph is shown in Fig. 2, where the weight is the cost of traversing the edge.

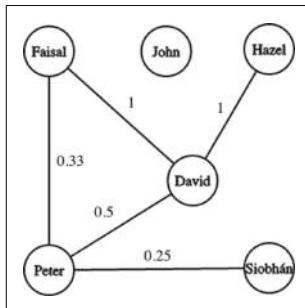


Fig. 2. How a graph created by Algorithm 1 would look

The initial graph has 39,199 nodes and 158,279 edges. As such it would be computationally expensive to run many measurements on it. Consequently, the graph was reduced by selecting specific patent class ids and inducing the subgraph on all nodes/inventors holding patents in those classes.

One thing to note is that our graph includes all patents from the relevant database, which runs from 1976–2011. Some links are temporal, and a future

user can just limit the graph they create to currently active inventors by using a date of last patents as a cutoff point.

4.3 Machine Learning Model Training

From analysis of the collaboration graph G another dataset was created for the purpose of training a machine learning model. The aim is to produce a set of features that might predict if two inventors might collaborate, even if that is not reflected in the patent data set.

From here the machine learning dataset was created with the following headings: vi , vq , $bonding_vi$, $bridging_vi$, $bonding_vq$, $bridging_vq$, vi_collab_ratio , vq_collab_ratio , $vi_cluster$, $vq_cluster$, $vi_patents$, $vq_patents$, $common_neighbors$, $expert_jaccard$, $resource_allocation$, $connected$. We will explain each of these in turn. For each pair of inventors, vi and vq refer to each inventor of the pair. Some of the measurements relate to each inventor individually, and these are denoted by having vi or vq in their name.

The bonding and bridging capital [10] of each inventor was calculated by getting the communities of the graph, using the best partition function in networkx's community module, and seeing how many of the neighbors of each inventor are in the same community. The bonding capital is given by:

$$\text{Bonding Capital} = \frac{|\text{Neighbors in same community}|}{|\text{Total number of neighbors}|}. \quad (1)$$

Conversely, the bridging capital is given by:

$$\text{Bridging Capital} = \frac{|\text{Neighbors \textbf{not} in same community}|}{|\text{total number of neighbors}|}. \quad (2)$$

The collaborative ratio [15] of each inventor is given by:

$$\text{Collaborative Ratio} = \frac{|\text{Collaborative patents of inventor}|}{|\text{Patents of inventor}|}, \quad (3)$$

where a *collaborative patent* is any patent which has more than one inventor.

The values $vi_cluster$ and $vq_cluster$ are simply the clustering co-efficient of each inventor. Likewise, $vi_patents$ and $vq_patents$ are simply the total number of patents held by each inventor and $common_neighbors$ is self explanatory.

The Jaccard Index of two sets is given by:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}. \quad (4)$$

To find the $expert_jaccard$ between two inventors, we take the Jaccard Index of the set of patent classes in which each inventor held expertise. We define expertise by an inventor being in the top quartile of inventors in a patent class, as measured by number of patents held in that class.

Resource allocation is the *resource allocation index* of the pair of nodes as given by Python's networkx module. Finally, the connected value is 1 if there is an edge between v_i and v_q in the collaborator graph, and 0 if there is no edge.

We now have a dataset for a machine learning algorithm which will attempt to predict if two nodes are connected based on the features above.

4.4 Machine Learning Model

We now use Python's xgboost module to create a model, with the *connected* column providing the binary classification target for prediction.

As there is no link between most pairs of inventors, the connection column exhibits substantial skew that might impact predictions. To counteract this, SMOTE oversampling [3] was employed to ensure a more accurate model was trained. Following the training of this model, a second graph split between random domains was created as before, and a dataset of the same format was created. The model was tested on this new, unseen, dataset and the predicted *connected* column was compared against the actual one. From this, the accuracy of the model was measured.

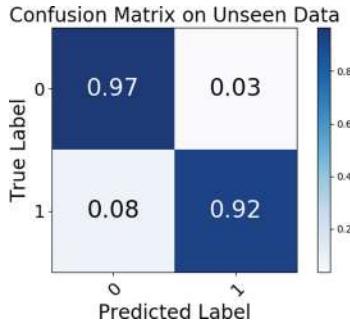


Fig. 3. Confusion matrix of machine learning model on unseen data

Figure 3 shows that the model provides good accuracy across all classifications. The model is predicting ones, or there being an edge between inventors, with a 92% accuracy. Conversely, it is incorrectly predicting that there is a zero, or no edge between inventors, 8% of the time. The main area of interest for us is the top right quadrant. Our model predicts that there is no edge between inventors, 97% of the time, and predicting incorrectly that there is an edge between inventors just 3% of the time. This 3% of *false positives* will be what we use to create additional links on our graph.

4.5 Creating New Links

Having built a model to predict links, we are now ready to introduce the graph that is augmented with extra edges that are predicted by that model. This is to

Result: An Augmented Graph with Machine Learning False Positive Edges added

Initialisation: Gather all pairs of nodes between which there was an edge predicted by the machine learning model but no edge in the collaboration graph;

for *Each Pair of Nodes do*

| Add edge between them;

Algorithm 2. Creating Machine Learning Augmented Graph

overcome a limitation of using the Enhanced Steiner Tree Algorithm [6] directly on the collaboration graph, in that it will not consider collaborations that sit in unconnected components.

As such, we propose that artificial links are created on the graph between pairs of inventors for which our machine learning model incorrectly predicted there was a link, or the 3% as seen in the top right quadrant of Fig. 3. The logic here being that the machine learning model predicted, based on the parameters discussed in Sect. 4.3, that these two inventors had attributes that were considered to be conducive to a collaborative relationship. This procedure is shown in Algorithm 2.

The graph was augmented by creating an edge between every pair of inventors for which the model predicted an edge, but there didn't already exist such an edge. We also need to provide a weight for the edge. This was set as $1 - P$ where P was the probability of an edge existing as predicted by the model. This weight is applied to all edges, including those included in the original collaboration graph.

This provides us with our augmented graph on which we will run the Enhanced Steiner Tree algorithm to identify a team.

4.6 Testing

Given the relation between team size and cost, the main criterion for testing was the team size necessary for the given skills.

This raised the obvious point that, having added in edges predicted by the machine learning model, the cardinality of the team would almost always be smaller, as the graph is now better connected. In fact, even if we added edges at random, we would expect teams to be smaller. We take advantage of this intuition, and use teams generated from a randomly augmented graph as a comparison. To generate this randomly augmented graph, we add edges at random between pairs of nodes until the number of edges of this randomly augmented graph equals the number of edges of the machine learning augmented graph.

In the evaluation we will compare the average team size required for different skillsets (i.e. sets of required skills). Since there are many skillsets, we compare the average team size as a function of the number of skills.

Specifically, a task T is a set of patent classes that could be interpreted as the skillset required for the task at hand. For this task, the Enhanced Steiner

Table 1. Average cardinality of teams for task of size 10

Task	G	G random-augmented	G ML-augmented
T_1	13.89	7.97	6.88
T_2	9.69	7.92	6.41
T_3	7.86	5.36	4.56
T_4	5.76	5.69	3.83

Tree algorithm [6] can be run on (1) the original graph (2) the machine learning augmented graph, and (3) the randomly augmented graph.

5 Results

This evaluation was initially performed by selecting 10 patent classes at random. The algorithm was run 100 times, and the average team size recorded. This was repeated for four different random tasks, T , of size 10.

Table 1 shows the results for these four initial tests performed on differing tasks T , all of size 10. Our augmented graph returns a smaller team than the original collaboration graph, as expected, but also consistently returns a smaller team than the graph with randomly created links, even though the random graph and augmented graph both have the same number of edges.

Following that, a random T was selected as before, but with a size of 8. The same tests were again run, in this case 50 times, and the average number of members in each team recorded. T was increased by 1 random element after each iteration until the size of T was 24 patent classes.

The results displayed in Fig. 4 show the outcome of the tests for differing numbers of patent classes in task T . The plot shows the number of elements in a task T on the x-axis, and the average number team members returned having run the Enhanced Steiner Tree algorithm 50 times on all three graphs.

The average cardinality of the team returned for the augmented graph is lower than that of the original collaboration graph, intuitively and as we discussed in Sect. 4.6. More interestingly, the average cardinality of the team returned when the Enhanced Steiner Tree algorithm is run on the machine learning augmented graph is consistently lower than the graph which was augmented with random links, even though the random graph has the same number of edges as the machine learning augmented graph.

This result is replicated for nine other random tasks T as shown in Fig. 5. In this table, each graph follows the same legend and has the same axes as the graph shown in Fig. 4, although some of the y axis scales differ. This shows that the machine learning augmented graph consistently returns a smaller team, than that returned from the randomly augmented graph and the regular graph, which covers the skillset across differing tasks of differing sizes.

While the team size is an important factor, one might also want to consider the communicative cost. One crude way to represent this would be the

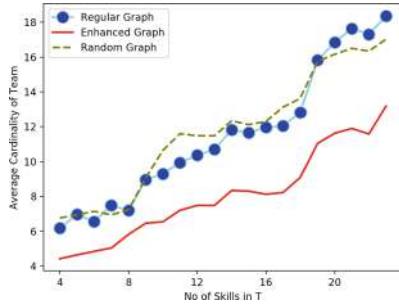


Fig. 4. Average cardinality of team V number of skills in task, T

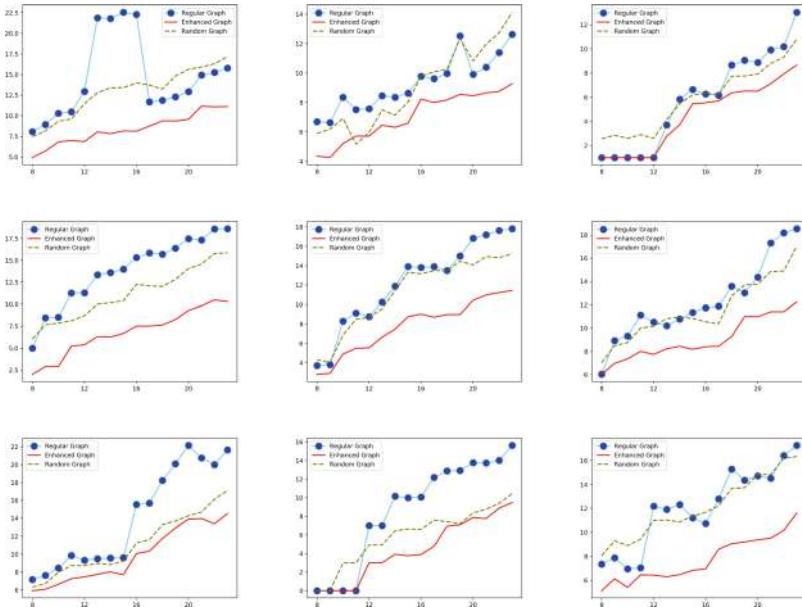


Fig. 5. Average cardinality of teams for differing task sizes

sum of weights of the links used by the team. The results of this are shown in Fig. 6. Interestingly, though the average cardinality of the team suggested by the machine learning graph is lower, the sum of weights tends to be higher at larger numbers of skills. The physical or financial communicative cost is, of course, determined by the circumstances. Consequently, it may be interesting to explore other, more meaningful measurements of the communicative cost.

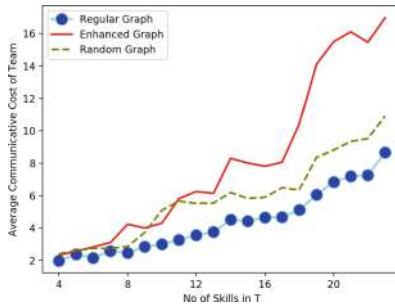


Fig. 6. Communicative cost based on machine learning probabilities

6 Conclusion

We aimed to provide a method to find small teams that cover necessary skillsets using a combination of graph-based techniques and machine learning. We built upon the Enhanced Steiner Tree algorithm, which provides a good solution to the team formation problem, but can sometimes neglect to include inventors which may not be well connected in the network. We use machine learning to predict links to connect these inventors based on potential relationships.

Our results show that when we add these predicted links we consistently get a smaller team which covers the skills required for a given task T , compared both to the original collaboration graph and a randomly augmented graph. This suggests our augmented graph produces compact teams with good collaborative potential.

A substantial area for future research is the communicative cost of the teams. There may be a more direct method available to users for measuring potential communicative cost. This could be traded off against the cost of additional team members in an expanded model. Another area to be considered is how to measure the efficacy of a team, pre-existing or newly formed by this method. This could give a whole new dimension on the team formation problem.

Future work may also consist of including the temporal aspects of the links in the team finding algorithm, instead of the graph fed into the algorithm. We may also explore using bipartite networks, where one set of nodes is the expertise, and the other is the inventors. It is worth noting, that the enhance line of the Enhanced Steiner Tree algorithm does create additional nodes based on skills, and connects them to inventors who possess those skills, but it is not a true bipartite graph, as the inventors are still connected directly.

Acknowledgement. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) and is co-funded under the European Regional Development Fund under Grant Number 13/RC/2077.

References

1. Hasan, M.A., Chaoji, V., Salem, S., Zaki, M.: Link prediction using supervised learning. In: SDM 2006: Workshop on Link Analysis, Counter-Terrorism and Security (2006)
2. Bollobás, B.: Modern Graph Theory, vol. 184. Springer, New York (2013)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Philip Kegelmeyer, W.: Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
4. Gorla, N., Lam, Y.W.: Who should work with whom?: building effective software project teams. *Commun. ACM* **47**(6), 79–82 (2004)
5. Hwang, F.K., Richards, D.S., Winter, P.: The Steiner Tree Problem, vol. 53. Elsevier, Amsterdam (1992)
6. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 467–476. ACM (2009)
7. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. *Phys. A Stat. Mech. Appl.* **390**(6), 1150–1170 (2011)
8. Newman, M.E.J.: Clustering and preferential attachment in growing networks. *Phys. Rev. E* **64**(2), 025102 (2001)
9. Pendharkar, P.C., Rodger, J.A.: The relationship between software development team size and software development cost. *Commun. ACM* **52**(1), 141–144 (2009)
10. Sharma, R., McAreavey, K., Hong, J., Ghaffar, F.: Individual-level social capital in weighted and attributed social networks. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1032–1037. IEEE (2018)
11. Spadon, G., de Carvalho, A.C.P.L.F., Rodrigues-Jr, J.F., Alves, L.G.A.: Reconstructing commutes network using machine learning and urban indicators. *Sci. Rep.* **9**(1), 1–13 (2019)
12. Tang, J., Wu, S., Sun, J., Su, H.: Cross-domain collaboration recommendation. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1285–1293. ACM (2012)
13. Wang, X., Zhao, Z., Ng, W.: A comparative study of team formation in social networks. In: International Conference on Database Systems for Advanced Applications, pp. 389–404. Springer (2015)
14. Watts, D.J., Strogatz, S.H.: Collective dynamics of ‘small-world’ networks. *Nature* **393**(6684), 440 (1998)
15. Wu, S., Sun, J., Tang, J.: Patent partner recommendation in enterprise social networks. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 43–52. ACM (2013)
16. Zhang, J., Lv, Y., Yu, P.: Enterprise social link recommendation. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management, pp. 841–850. ACM (2015)



Scientometrics for Success and Influence in the Microsoft Academic Graph

George Panagopoulos¹(✉), Christos Xypolopoulos¹, Konstantinos Skianis¹,
Christos Giatsidis¹, Jie Tang², and Michalis Vazirgiannis^{1,3}

¹ Ecole Polytechnique, Palaiseau, France

{george.panagopoulos,cxypolop,konstantinos.skianis,
mvazirg}@polytechnique.edu, xristoskamad@gmail.com

² Tsinghua University, Beijing, China

jietang@tsinghua.edu.cn

³ Athens University of Economics and Business, Athens, Greece

Abstract. Measuring and evaluating an author's impact has been a withstanding challenge in the academic world with profound effects on society. Apart from its practical usage for academic evaluation, it enhances transparency and reinforces scientific excellence. In this demo paper we present our efforts to address this problem capitalizing on the field-based citations and the author oriented citation network extracted from the Microsoft Academic Graph, to our knowledge the largest network of its kind. We separate impact into two dimensions: success and influence over the network, and provide two novel scientometrics to quantify some of their aspects: (i) the distribution of the h-index for specific scientific fields and a search engine to visualize an authors' position in it as well as the top percentile she belongs to, (ii) recomputing our previously introduced D-core influence metric on this huge network and presenting authority/integration of the authors in the form of D-core frontiers. In addition we present interesting insights on the most dense scientific domains and the most influential authors. We believe the proposed analytics highlight under-examined aspects in the area of scientific evaluation and pave the way for more involved scientometrics.

Keywords: Scientometrics · Influence · Large-scale network analysis

1 Introduction

In a time when trained scientists far exceed the demand for academic personnel [16], tools to evaluate scientific impact are more timely than ever. However, given the steep growth of scientific literature in terms of size and variety during recent years, the paper and author evaluation criteria need to evolve to account for these aspects. More specifically, academic evaluators like tenure committees or funding agencies often fail to take everything into consideration because there is so much work published that is overwhelming [9]. Hence they rely on metrics

like the number of citations or journal impact factor that though useful, do not suffice to capture the whole spectrum of a scientist's success [1].

In this demo paper, we demonstrate a set of analytics that provides novel insights on a scientist's impact based on the Microsoft Academic Graph (MAG), the largest openly available corpus of bibliographic data. More specifically, we utilize information regarding papers, their authors and the fields they belong to, to extract field-based citations and the *Author Oriented Citation (AOCI)* network (to our knowledge the largest existing of its kind). We then derive two interactive visualizations that aspire to convey the author's success and influence over the academic network:

- We extract a field-based h-index (*f-h-index*) for each author using the citations she receives in her papers that belong to specific fields. Afterwards we form *f-h-index* distributions and position the author's *f-h-index* in them together with the top percentile she belongs to, for the distributions of the top three fields she most frequently publishes in.
- We compute the *D-core decomposition* of the AOCI graph, a measure of influence in directed networks [5] adopted in the Aminer scientific search tool. Subsequently, we use the sub-graphs induced by the decomposition to form the D-core matrix, a rectangular heat map that displays the outmost cores (i.e. the densest citation graphs) that an author belongs to, indicating her aggregate influence in terms of authority (incoming citations) and community integration (in terms of outgoing citations).

Both visualizations along with the corresponding author search engines are deployed online¹. The first challenge towards computing the above was ensuring the quality of the dataset. Although MAG has been preprocessed extensively, we found out that more than 30% of the unique names have been assigned multiple IDs. To dive deeper, we performed an exploratory analysis on the number of papers for these IDs and show that more than 80% have been assigned to one paper only. This means that their h-index is 1 and they will have a minuscule contribution to the influence estimation, thus we proceed to remove them from the analysis. The second challenge adheres to the issue of scalability. Creating AOCI in a typical manner is not feasible since it includes a join operation between a table of 320Gb with a table of 30Gb. We overcome this by breaking the author-cites-paper table in multiple batches with non-overlapping authors and calculating the ego-networks of the authors in each batch in parallel. Subsequently, we filter the network based on edge weights to keep only the most consistent citing relationships and run the D-core decomposition. As a means of anecdotal evaluation, we examine the densest D-core subgraph and provide a list of the most influential authors as well as a wordcloud with the most frequent scientific fields in it. The vast majority of the retrieved fields are related to particle physics, a field well known for its massively dense collaboration and citation patterns [3]. The overall data pipeline employed can be seen at Fig. 1. The code to reproduce the analysis can be found online².

¹ <http://graphdegeneracy.org/scientometrics/>.

² <https://bit.ly/2mboMBY>.

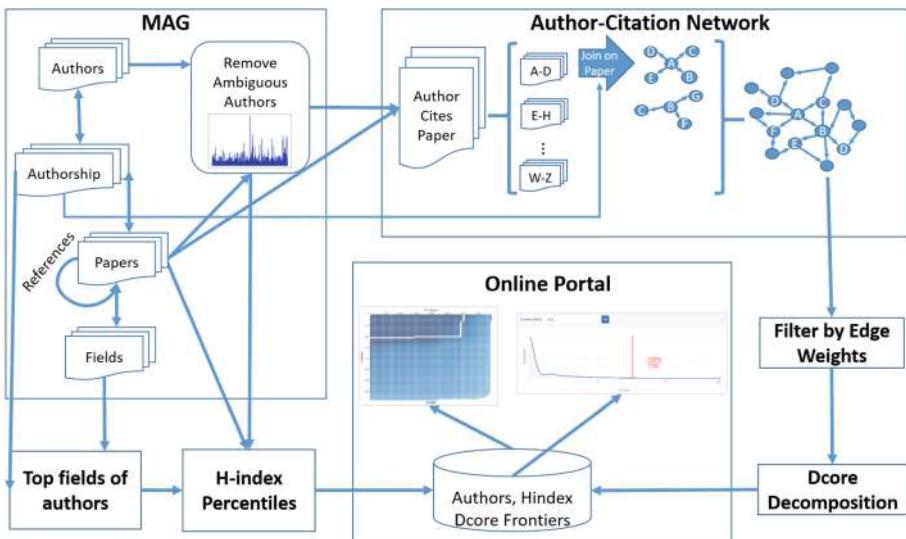


Fig. 1. Schematic representation of the data pipeline.

2 Data Preprocessing

MAG is the biggest openly available bibliographical corpus, including more than 250 million authors and 219 million papers [13]. We employed a snapshot version³ provided by Microsoft Academic Services in Azure storage, but before that, we have examined thoroughly other options such as DBLP, Aminer, and the openly available MAG through the Open Academic Graph [14]. The reason we decided to go with the official version of MAG is: (i) MAG h-index estimates were closer to other services like Google Scholar and Scopus, (ii) open data were missing conference names, scientific fields, etc. That said, MAG required significant preprocessing before performing the analysis.

As a first step, we removed names that do not include English characters, which diminishes the initial 83 to about 63 million. Afterwards we deal with the issue of multiple IDs referring to the same name. This problem refers to the name ambiguities created by noisy registers in the dataset, such as “A. H. Tang”, “Arthur Tang”, “A. Tang” etc. which refer to the same person but are interpreted as different [18]. In our case, a name is ambiguous if it has more than 1 IDs, which is quite common. To be more specific, out of the 63 million names, roughly one third (23.427.411) has more than one IDs, amounting to 176.631.328 ambiguous ID assignments. One can see in Fig. 2 that the number of possible IDs for the same ambiguous name can reach more than 27.000. Roughly 3.500 names had more than 1.000 IDs each, revealing the scale of the problem.

More than 80% (142.851.013) of these ambiguous IDs share a peculiar characteristic: they have only one paper. For example, the name *Kristin Person* that

³ Created at 9/2/2019.

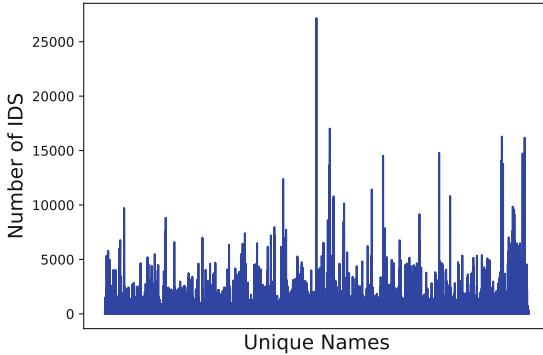


Fig. 2. Number of ids for the ambiguous names in MAG.

has 27169 author IDs from Fig. 2, has 27126 IDs with only one paper each. We can hypothesize that these IDs were created specifically for each paper because they could not be assigned to an already existing name. These IDs are not useful for our purpose since by default they will have an h-index of 1 and will create very weak edges in AOCI. These 142 million IDs are assigned to 22.503.619 ambiguous names, which is more than 95% of all ambiguous names. Since we lack an efficient way to perform name disambiguation at this scale, we remove them from the analysis, ending up with 40 million IDs.

3 H-index Distribution

H-index is considered one of the most successful scientometrics [6], employed uniformly from academic institutions and bibliographical platforms such as Google Scholar and Aminer. In spite of its broad usage though, it is known to suffer from certain limitations [2], one notably being that as is just a number, it can be easily misinterpreted if presented out of context. In other words, a scientist's success might be best depicted by comparing her h-index with the rest of academia. Another important disadvantage is the significant different rates of citations exhibited throughout different principles [17, 19], which renders interdisciplinary comparisons impossible. More specifically, in several occasions we aim to compare scientists from aberrant but collaborating fields, like biology and computer science, where that latter has a significantly lower publication rate than the former. In our dataset, biology has the higher average h-index with an average of 7.1 and compute science is seventh with 5.9 because h-index is biased towards scientific fields with increased publication volume. To this end, we propose a set of plots that present an author's position in her most relevant fields, by visualizing the h-index from citations she receives in papers that belong to each field (*f-h-index*), relative to the distribution of all authors' *f-h-index*.

To retrieve a field-specific h-index, we first assigned fields to each paper based on the subfield related to it, which was assigned with an associated *confidence*

Table 1. The 14 fields assigned to papers in Microsoft Academic Graph

Biology	Medicine
Mathematics	Computer science
Engineering	Chemistry
Physics	Business
Visual arts	Media
Language	Geography
Agriculture	Food

metric by Microsoft Academic Services [12]. The fields can be seen in Table 1. In case the sub-field contained no categories, we assigned to it the categories of its related sub-field. Thus, each paper p has a number of fields f assigned to it with a certain confidence $c_{p,f}$. To retrieve the most prevalent fields for an author a with P_a papers, we computed her *relevance* to a specific field as:

$$r(a) = \sum_{p \in P_a} c_{p,f} \quad (1)$$

We keep the top three fields for each author based on this ranking. Subsequently, we separate the papers she wrote in each of these fields, and derive a list of their citations. Note here that a paper might belong to more than one field, hence its citations are included in the lists of all these fields. The *f-h-index* is computed through the typical h-index formula [6], by sorting the author's papers on a given field based on their number of citations and taking the maximum position h where the citations are at least h . We also derive the general h-index of all authors, as an indication of an author's overall performance. The relationship between an author's top three *f-h-index* and her h-index is not straight forward because of the presence of papers in multiple fields and our disregard for the fields she publishes more rare to i.e. they are not in the top three. Apart from the aforementioned preprocessing, we aggregate authors based on their name and keep the record with the biggest h-index, which results in a set of 23.202.638 names. In order to make the distributions more legible we also removed all authors with an h-index of 1, which diminishes the number to 10.050.770. Finally, the distribution is in logarithmic scale to alleviate the severe skewness that characterizes such distributions. The search is based on the last name and the results are sorted in descending fashion based on the overall h-index.

An example of use case can be seen in Figs. 3 and 4. Both researchers have the same overall h-index, but the percentiles they belong to in their specific fields are different. More specifically, the one is close to the top 0.3 in computer science and 0.8 in business which is his actual field based on the normalized relevance score, while for the other, the highest percentile is close to 1 and over 1.45 for his actual field. Though indisputably both authors are successful, dr. Joachim Sachs can be considered more impactful in his fields, using this comparison.

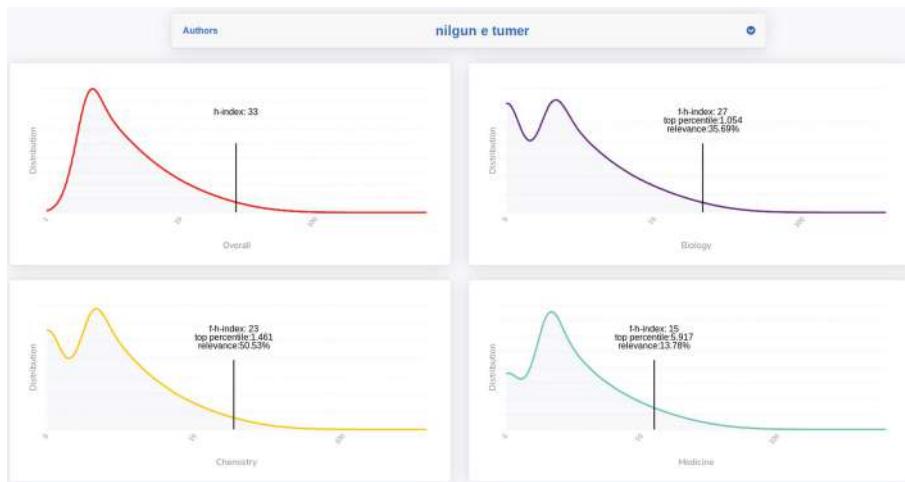


Fig. 3. Field h-index percentiles for a researcher in chemistry, biology and medicine.

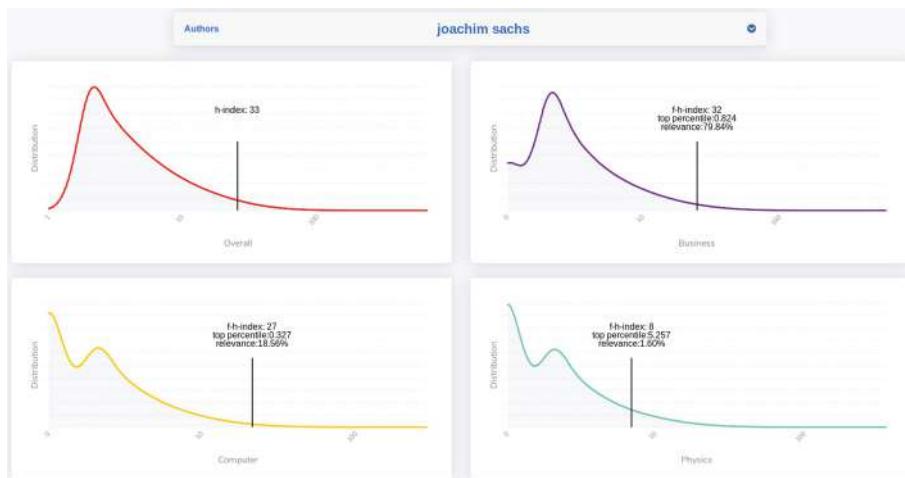


Fig. 4. Field h-index percentiles for a researcher in business, computer science and physics.

4 Author Influence

The influence exerted by scientists to the academic world can be measured from multiple different perspectives. One could argue that co-authorship is a form of influence [11]. However, scientists can have an exceptional number of co-authors without performing remarkable work, thus citation could be considered an indication of influence. The semantic scholar defines an author's influence on another author based on how many times the latter cites the former [15]. Similar

works have analyzed the impact of an institution given its citations to and from other institutions [8]. That said, simply relying on the number of citations does not suffice, as authors may get cited a lot but from a limited number of people, which constraints their actual visibility. Consequently, we deem more appropriate to measures influence based on the scientist's placement in the aforementioned AOCI, which can capture both, the number of in and out citations as well as the number of people that cite and get cited by an author. Especially the out citations is a metric that shows activity and integration in the community.

One of the most popular node influence metrics in other applications is the max core it resides in [7]. k-core decomposition identifies subgraphs that are very densely connected, overlooking highly connected nodes that lie in the periphery of the network, which is the pitfall of simpler methods such as the degree. However, the citation is a strictly directed relationship, meaning that there is a significant difference between citing and getting cited. In order to sustain this dual nature of in and out citations, we employ the D-core decomposition [5], a well-known adaptation of k-core that captures both the authority (via the in citations) and the integration (via the out citations) of an author's directed influence and has been adopted by the popular bibliographic exploration system Aminer.org⁴.

4.1 Author Oriented Citation Network

The sole extraction of AOCI from the MAG was not straight forward due to the scale of the dataset and its dense nature. More specifically, given a certain paper with c citations, each cited paper with an average of a authors, the paper's authors will acquire $c * a$ edges, without taking author overlapping into account. To form the graph we used the paper-references table and the paper-author table from MAG⁵. We first created a table that consists of the authors and the papers they cite, by a simple join on the paper ID. Subsequently, we sort based on the author ID and break the table in 20 batches with no author shared. We can then join each batch with the paper-author table using the reference ID as key to the paper ID to compute author-cites-author recordings which can be turned in a weighted edge-list grouping by the authors and counting their co-occurrence. Essentially this is the out-egonetwork of each author in that batch, and we can create the whole network by computing them as shown at Fig. 1. The resulting edge weights are number of times an author cited another author. This creates a network of 67.614.736 unique author ids with over 17 billion edges. By removing the aforementioned ambiguous IDs and the edges with weight less than 10, we end up with a directed network of 8.960.233 nodes and 599.586.916 edges. Removing edges with equal or less then 10 citations allows us to keep only the nodes that consistently cite each other. In other words, apart from reducing the size of the network in a manageable scale, it also allows us to

⁴ <https://aminer.org/>.

⁵ All the queries were performed in PySpark in a cluster of 32 nodes, 16 GB ram each, Intel(R) Xeon(R) CPU E5-2407 v2 @ 2.40 GHz.

perform a preliminary community detection, as the consistent citing patterns between authors represent the circulation and development of certain scientific ideas.

4.2 D-core Decomposition

Given a directed network $D = (V, E)$ with a set V of vertices and a set E of directed edges between them, following the literature [5], we define a directed core $DC_{k,l}$ as the subgraph where each node $v \in V(DC_{k,l})$, has at least $\deg_v^{in}(DC_{k,l}) \geq k$ and $\deg_v^{out}(DC_{k,l}) \geq l$. It should be noted that like the original k -core decomposition [7], this is not one-time filtering but rather a recursive process where the nodes that are not in accordance with the aforementioned criteria are removed iteratively. This means that even if a node initially complies with the requirements, it may eventually be removed after removing the rest of the nodes that do not comply if any of them are its neighbors. This facilitates the retrieval of the densest possible subgraphs in the network, also called the network “cores”, where each node cites and gets cited with sufficient consistency. To compute the decomposition we first define the maximum K and L , which are the last bounds where the D-core is non-empty. We iterate over all $k' \in [0, K]$ and perform the decomposition for $k = k'$ and all $l \in [0, L_{k'}]$, where $DC_{k',L_{k'}+1}$ is empty. This provides us with a list of $L_{k'}$ subgraphs for each k' . Since this computation is independent throughout different k' , we can parallelize this process by running a different range in $[0, K]$ to a different CPU. We used the retrieved D-cores to create a D-core matrix, a heatmap with the size of the D-cores in a gradient motif. We can then visualize a scientist’s authority/integration aptitude based on the outmost D-cores she belongs to, as indicated in the example of Fig. 5. We remove authors with the same name to facilitate search, based on who has the highest D-cores, and we are left with 6,252,393 names. Since the maximum incore and outcore found are 7800 and 7900 respectively, creating an interactive heatmap of such scale is neither practical nor feasible. To this end, we break the plot in two versions. The “macro” version shows the densest cores and uses a step of 100, meaning the D-cores shown are for $mod(k, 100) = 0$ and $mod(l, 100) = 0$. This includes only 175,952 authors, while the rest lie in the D-cores for $k, l \in [0, 100]$, in which case we employ the “micro” version where the D-core matrix has a step of 1.

4.3 D-core Analytics

The aforementioned plot can be used to derive the most influential users in the network, which reside in the optimal collaboration core (OCD-core) based on $OCI = (k + l)/2$ [5], in our case $k = 7600$ and $l = 7480$. The top authors in OCD-core in terms of degree, which resembles the number of people they cite and getting cited by, along with the h-index and their fields of study as found by searching in the web, are shown in Table 2. We notice that degree and h-index are not totally correlated - implying that the density of the citation network

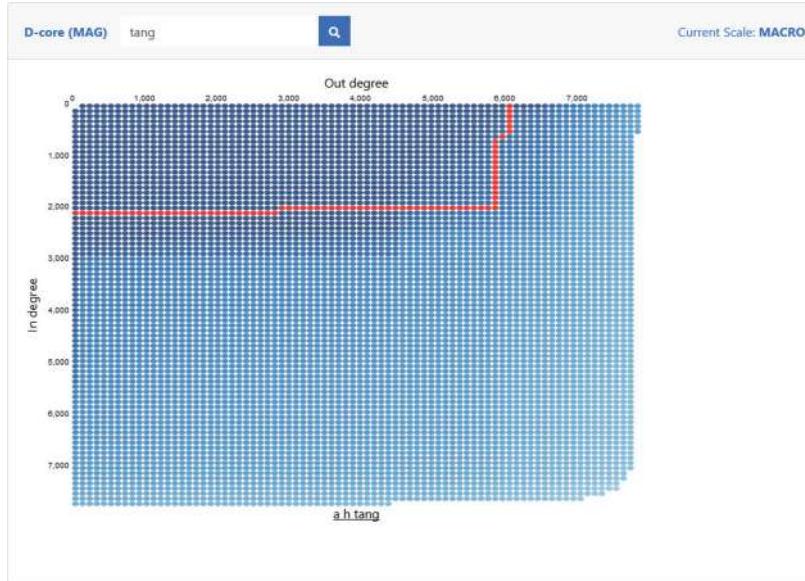


Fig. 5. D-core matrix visualization.

captures different influence aspects, as indicated in the recent rooted citation influence metric [4].

We can observe that the majority of the top authors are physicists. In general, OCD-core includes authors that cite and getting cited by each other in a massive rate, which could be characteristic of some fields. To evaluate further this hypothesis, we employ the one-word fields assigned to papers in MAG, keeping only assignments with confidence >0.5 . We retrieve the fields of the papers of the authors in OCD-core and aggregated the number of authors in each field. Figure 6 shows the word-cloud indicating the number of authors of the top 100 fields. As one can see, most of the words belong to particle physics and related fields. This happens because physics laboratories throughout the world collaborate with each other to validate their experimental findings. Thus the publications are comprised of tens or even hundreds of authors who cite similar older publications, resulting in a very dense network with an increased rate of citations.

5 Future Work

We presented an online system with interactive visualizations that capitalize on the author oriented citation network and field-based citations from MAG to capture different dimensions of a scientist's impact. We claim that this effort contributes to talent management and ranking within but also out of the academic domain. We plan to extend this portal further with the addition of several novel

Table 2. Top 10 authors in terms of degree in OCD-core.

Name	Degree	H-index	Field of study
Hsi Shu Chen	80300	88	Human Space Systems
Andrea Bocci	73476	98	Particle Physics
Swagato Banerjee	72728	86	Particle Physics
Barbara Abbott	72243	84	Developmental Biology
Pauk Kyberd	70650	83	Particle Physics
Martin Weber	69633	77	Particle Physics
George Azuelos	69372	80	Materials Science
Wolfram D. Zeuner	69266	79	Particle Physics
Thomas Hebbeker	69247	88	Particle Physics
Daria Bisello	69211	86	Particle Physics

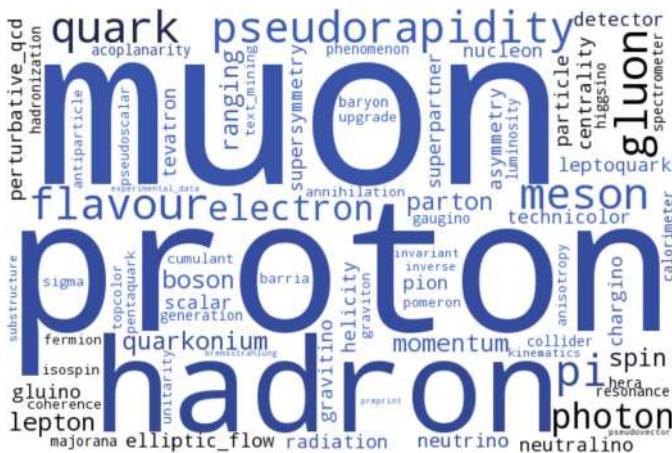


Fig. 6. The most frequent fields appearing in the OCD-core.

analytics that address other well-known scientometric problems such as rising star prediction [10] and self-citation ratios. Moreover, as the author to author citing patterns are inherently weighted (reflecting the quantified impact of an author to another) - we plan to compute the D-core taking the edge weights into account, aiming at more informative and valid influence rankings.

References

1. The number that's devouring science. <https://www.chronicle.com/article/the-number-thats-devouring/26481>. Accessed 19 May 2019
 2. Bornmann, L., Daniel, H.D.: What do we know about the h index? *J. Am. Soc. Inform. Sci. Technol.* **58**(9), 1381–1385 (2007)

3. Chompalov, I., Genuth, J., Shrum, W.: The organization of scientific collaborations. *Res. Policy* **31**(5), 749–767 (2002)
4. Giatsidis, C., Nikolentzos, G., Zhang, C., Tang, J., Vazirgiannis, M.: Rooted citation graphs density metrics for research papers influence evaluation. *J. Informetr.* **13**(2), 757–768 (2019)
5. Giatsidis, C., Thilikos, D.M., Vazirgiannis, M.: D-cores: measuring collaboration of directed graphs based on degeneracy. *Knowl. Inf. Syst.* **35**(2), 311–343 (2013)
6. Hirsch, J.E.: An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci.* **102**(46), 16569–16572 (2005)
7. Malliaros, F., Giatsidis, C., Papadopoulos, A., Vazirgiannis, M.: The core decomposition of networks: theory, algorithms and applications (2019)
8. Massucci, F.A., Docampo, D.: Measuring the academic reputation through citation networks via pagerank. *J. Informetr.* **13**(1), 185–201 (2019)
9. Mohammed, B.: Scientometrics 2.0: toward new metrics of scholarly impact on the social web. *First Monday* **15**(7) (2015)
10. Panagopoulos, G., Tsatsaronis, G., Varlamis, I.: Detecting rising stars in dynamic collaborative networks. *J. Informetr.* **11**(1), 198–222 (2017)
11. Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., Schweitzer, F.: Predicting scientific success based on coauthorship networks. *EPJ Data Sci.* **3**(1), 9 (2014)
12. Shen, Z., Ma, H., Wang, K.: A web-scale system for scientific knowledge exploration. arXiv preprint [arXiv:1805.12216](https://arxiv.org/abs/1805.12216) (2018)
13. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.J.P., Wang, K.: An overview of Microsoft Academic Service (MAS) and applications. In: International Conference on World Wide Web (The WebConf) (2015)
14. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: ArnetMiner: extraction and mining of academic social networks. In: Knowledge Discovery and Data Mining (KDD), pp. 990–998 (2008)
15. Valenzuela, M., Ha, V., Etzioni, O.: Identifying meaningful citations. In: Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
16. Waaijer, C.J., Teelken, C., Wouters, P.F., van der Weijden, I.C.: Competition in science: links between publication pressure, grant pressure and the academic job market. *High. Educ. Policy* **31**(2), 225–243 (2018)
17. Waltman, L., Van Eck, N.J.: The inconsistency of the h-index. *J. Am. Soc. Inform. Sci. Technol.* **63**(2), 406–415 (2012)
18. Zhang, Y., Zhang, F., Yao, P., Tang, J.: Name disambiguation in Aminer: clustering, maintenance, and human in the loop. In: Knowledge Discovery & Data Mining (KDD), pp. 1002–1011 (2018)
19. Zitt, M., Small, H.: Modifying the journal impact factor by fractional citation weighting: the audience factor. *J. Am. Soc. Inform. Sci. Technol.* **59**(11), 1856–1860 (2008)



Testing Influence of Network Structure on Team Performance Using STERGM-Based Controls

Brennan Antone^{1(✉)}, Aryaman Gupta¹, Suzanne Bell², Leslie DeChurch¹,
and Noshir Contractor¹

¹ Northwestern University, Evanston, IL 60201, USA

brennanantone2017@u.northwestern.edu

² DePaul University, Chicago, IL 60614, USA

Abstract. We demonstrate an approach to perform significance testing on the association between two different network-level properties, based on the observation of multiple networks over time. This approach may be applied, for instance, to evaluate how patterns of social relationships within teams are associated with team performance on different tasks. We apply this approach to understand the team processes of crews in long-duration space exploration analogs. Using data collected from crews in NASA analogs, we identify how interpersonal network patterns among crew members relate to performance on various tasks. In our significance testing, we control for complex interdependencies between network ties: structural patterns, such as reciprocity, and temporal patterns in how ties tend to form or dissolve over time. To accomplish this, Separable Temporal Exponential Random Graph Models (STERGMs) are used as a parametric approach for sampling from the null distribution, in order to calculate p-values.

Keywords: Network properties · Team performance · Separable temporal exponential random graph models

1 Introduction

Across many areas of network science, research often asks questions about how different network-level properties or outcomes are related. For instance, when studying networks between team members, researchers may examine whether properties of the whole team's network, such as density or centralization, impact the performance of that team. In this case, each network may have one score for density and one score for team performance, and through the observation of multiple networks, a correlation between density and team performance can be computed. However, it is important to assess the potential spuriousness of these correlations, especially when working with a small sample of networks.

When correlations involve *network statistics*, functions that return a single score computed from a network, proper significance testing may be challenging.

The ties and node attributes that determine the value of the network statistic often are not statistically independent of one another. Complex interdependences may exist between ties and node attributes within the same network: for instance, ties may tend to be reciprocated, or ties may be more likely between nodes of the same gender. We will refer to this type of interdependency as *structural patterns* in how ties form. Additionally, in the event that researchers have collected data at multiple points in time, there may be complex interdependences between ties in repeat observations of the same nodes. For instance, if a tie exists between individuals at one point in time, it is natural to expect that tie to be more likely to exist the next time these individuals are observed. We will refer to this type of interdependency as *temporal patterns* in how ties form.

We argue that, when the value of network statistics being observed may be influenced by structural patterns or temporal patterns in how network form, significance tests of correlations involving these network statistics need to control for these patterns. An existing modeling approach, Separable Temporal Exponential Random Graph Models (STERGM) provides a way to identify both structural and temporal patterns [7]. We demonstrate how, when computing p-values for correlations involving network statistics, STERGMs can be used as a parametric approach for sampling from the null distribution.

We will apply this approach to understand team processes in the crews of long-duration space exploration (LDSE) missions, linking different patterns of social relations between crew members to measures of crew performance on various tasks. This form of significance testing is particularly beneficial to data collected in LDSE analogs such as those operated by NASA (e.g., Human Exploration Research Analog [HERA]), in which relatively few crews may be observed, but at multiple points in time. We demonstrate how crew performance on different tasks benefits from different types of network structure.

2 Motivation

To identify network influences on team performance, our goal is to quantify the possibility that observed correlations may be spurious. We frame our analysis around the null hypothesis that there is, in fact, no correlation between any two variables being tested. In this case, these variables will be some network statistic computed from a team's network and team performance. Based on the data we collected, we can compute an *observed correlation coefficient*, and an *empirical p-value*. The empirical p-value reflects the probability that we would find a correlation coefficient equal to or more extreme than the observed correlation coefficient in the event that there was no correlation between the network statistic and team performance. We will consider what issues structural and temporal patterns may introduce for such an analysis.

2.1 Influence of Structural Patterns

To understand the impact of structural patterns on correlation coefficients, consider a study of four-person teams. If we were to examine the network statistic of

closeness centralization [4], for a directed network there are 27 possible scores a four-person team can have for closeness centralization. Closeness centralization is a deterministic function of the 12 ties in the team. The likelihood of observing each of the 27 levels of closeness centralization may vary depending on structural patterns present, such as reciprocity, closure, or homophily. Under the influence of different structural patterns, different values of network statistics, and therefore correlation coefficients, may be more or less likely to occur by chance alone. Considering this, knowledge about structural patterns should inform how “surprising” it would be to observe a given correlation coefficient in the event that our null hypothesis is true.

2.2 Influence of Temporal Patterns

Further complications may be introduced if a researcher wishes to incorporate multiple observations of a team’s performance over time. Repeat observations may contain new information to help inform conclusions about team performance. To leverage such observations, we would want to control for non-independence (ex. autocorrelations) between the network statistics in repeat observations of a team. This can be accomplished by modeling temporal dependencies between a tie’s current value, that tie’s past or future values, and other ties’ past or future values. These may be simple trends, like the tendency of a tie to continue existing over time, or more complex trends, like the tendency of ties to form if doing so would complete a transitive triad. If there is not a way to control for these types of patterns, using repeated observations of a team would be problematic. Using multiple observations of a team may be critical in research contexts where obtaining data from additional teams may be costly or impossible, but teams are able to be studied for an extended period of time.

2.3 Existing Approaches

The development of null models for significance testing has a long history in network science, because interdependencies between ties must be controlled for when performing significance testing. Two fundamental approaches for significance testing when dealing with such interdependencies are nonparametric approaches and parametric approaches for sampling from the null distribution. *Nonparametric approaches* often rely on a type of permutation test, in which observations are shuffled in some systematic way. For example, Quadratic Assignment Procedure (QAP) and its extension multiple regression QAP (MRQAP) are often used to test correlations between the presence of ties in two or more networks by performing random relabeling of nodes in each permutation while keeping network structure constant [5,6]. Alternatively, approaches based on network rewiring are often used to permute the location of ties or events in a network while maintaining properties of that network’s degree distribution [12,14].

Parametric approaches entail the estimation of a model to sample from the null distribution, the distribution of the test statistic (ex. correlation coefficient) that would be observed in the event that the null hypothesis was true. Whereas

nonparametric approaches all either maintain network structure or specify exact rules for how network structure should be permuted, a parametric approach can estimate, based on data, the types of complex interdependencies that exist and perform appropriate permutations to control for them.

We propose a parametric approach for significance tests involving network-level statistics, in which both structural patterns and temporal patterns are modeled using STERGM. This approach will offer the distinct benefit of allowing observations of networks at multiple points in time to be used, by modeling how networks change between observations when sampling from the null distribution.

3 Approach for Testing the Influence of Network Structure on Team-Level Performance

Let us assume we have observed networks between team members, including different node attributes or other exogenous attributes that may affect network formation. Some of these networks may be collected from the same team at multiple points in time. We will refer to an ordered set of networks we collect from the same team as a *temporal path of networks*. We also assume that we have measured some *performance metric*, which assigns a single score to each network. Finally, we assume we have chosen a *network statistic*, a deterministic function of the ties and node attributes in a network, that we are interested in correlating with our performance metric. Our goal will be to assess the probability our sample might produce a correlation as extreme as the one observed if there was truly no correlation between our network statistic and performance metric.

We begin by calculating the correlation between the network statistic for each network and the corresponding performance metric from our empirical data. This produces the *empirically observed correlation coefficient*. While any form of correlation could be used, we suggest that due to the discrete or non-normally distributed nature of many network statistics it would often be appropriate to use a Spearman rank correlation coefficient [17]. Our approach will aim to estimate the probability that we would observe a correlation coefficient that extreme, if there was truly no correlation between our network statistic and performance metric in the full population. This is the empirical p-value. To conduct such a test, we want to control for structural and temporal patterns shaping how networks form. To accomplish this, we must define a version of the *null model* that accounts for each of these trends that may occur in the observed networks.

We propose using Separable Temporal Exponential Random Graph models (STERGM) as a flexible framework to construct our null models [7]. STERGMs describe how networks are likely to change over time by defining a joint probability distribution for the presence of ties in a series of repeat observations of networks. This distribution is defined by two sets of assumed sufficient statistics: A vector of *formation statistics* $\mathbf{g}^-()$ and their corresponding weights $\boldsymbol{\theta}^-$ describe how likely it is that a subset of ties \mathbf{Y}^- that did not exist in a previous network \mathbf{Y}^t will form. A vector of *persistence statistics* $\mathbf{g}^+()$ and their corresponding weights $\boldsymbol{\theta}^+$ describe how likely it is that the subset of ties \mathbf{Y}^+

that existed in a previous network \mathbf{Y}^t will continue to exist. In both cases, the assumed sufficient statistics are a function of both the ties being predicted and some dyadic or node covariates \mathbf{X} . The joint probability of each subset of ties, for a network at a single point in time, is expressed as:

$$P(\mathbf{Y}^- = \mathbf{y}^-) = \frac{e^{\theta^- \mathbf{g}^-(\mathbf{y}^-, \mathbf{X})}}{\sum_{i \in \mathbf{Y}^-} e^{\theta^- \mathbf{g}^-(i, \mathbf{X})}} \quad (1)$$

$$P(\mathbf{Y}^+ = \mathbf{y}^+) = \frac{e^{\theta^+ \mathbf{g}^+(\mathbf{y}^+, \mathbf{X})}}{\sum_{i \in \mathbf{Y}^+} e^{\theta^+ \mathbf{g}^+(i, \mathbf{X})}} \quad (2)$$

STERGMs are estimated using conditional maximum likelihood estimation as described in Krivitsky & Handcock 2014, in our case applying this technique to estimate a single model that jointly captures trends that occur amongst all of the temporal paths of networks we observed. As part of this estimation, as with any STERGM model, model convergence and goodness of fit should be assessed in order to make sure that the parameter estimation was successful and that the model replicates trends observed in the empirical data.

In comparison to TERGMs or ERGMs, STERGMs offer the benefit of representing complex temporal patterns in either the formation or persistence of ties between observations. Thus, when using multiple observations from the same networks over time, this provides an explicit mechanism for controlling for temporal dependencies between them when sampling from the null distribution.

STERGMs fit to our data are used to sample correlation coefficients from the null distribution. We simulate random networks according to the STERGM using Markov Chain Monte Carlo (MCMC) sampling. To obtain a simulated correlation coefficient, we take each different team in our dataset and simulate that team's temporal path of networks based on the node attributes and exogenous factor values for that team. We then calculate a correlation between the network statistics for all of the simulated networks and the performance metric for each network that we observed in our empirical data. By repeating this, we obtain a sample that approximates the null distribution, the distribution of correlation coefficients we would obtain based on our null model.

We then compare the empirically observed correlation coefficient to our samples from the null distribution. If we let n denote the total sample size of simulated correlation coefficients and k as the count of simulated correlation coefficients that are at least as extreme as our observed correlation coefficient (greater than or equal to if positive, less than or equal to if negative), then we estimate the p-value as the ratio k/n .

4 Application: Relational Indicators of Crew Success in Long-Duration Space Exploration

We will examine how different patterns of social relations between crew members of long-duration space exploration missions are associated with crew performance. Future lunar and Mars missions will entail extended trips, in which a

small crew must work together more effectively and autonomously, since communication delays will grow as the crew travels away from the earth. Thus, effective team processes will be critical to team success [15]. By understanding the effects of social networks on crew performance, space agencies will better able to staff and support crews on these missions by examining their interpersonal relations.

A challenge in research about long-duration space exploration (LDSE) is the limited ability of relevant data. One environment for collecting data is LDSE-analogs, in which participants may complete tasks typical of LDSE while living in an isolated environment for an extended time [9]. These analogs allow researchers to collect high-quality data from only a small number of crews over an extended time. Because of this, there is a need for analysis capable of leveraging repeated observations of a team to test what factors impact crew performance [2].

4.1 Measures

Research Setting. Data was collected from the Human Exploration Research Analog (HERA), an extended simulation of space exploration that is operated by NASA at the Johnson Space Center in Houston, Texas. Participants in HERA missions completed tasks over the course of a 30 or 45 day mission that simulated space exploration, remaining confined in the HERA capsule for the entire duration. Over the course of missions, participants experienced long shifts, sleep deprivation, communication delay with ground control, and emergency simulations designed by NASA to provide a realistic simulation of space exploration.

Respondents. Eight four-person crews completed analog space missions between January 2016 and June 2018. Four crews completed 30 day missions, and four completed 45 day missions. Each crew had a designated commander, a flight engineer, and two mission specialists. Of the 32 respondents, 59.4% were female, the average age was 38.0 years (s.d. = 7.98), and 34.4% had military experience. When asked about their race/ethnicity, 24 respondents selected Caucasian non-Hispanic, two selected Caucasian Hispanic, two selected East Asian, one selected South/Southeast Asian, one selected African American, and two selected “Other”.

Performance Dimensions. Team performance measures the degree to which a team accomplishes its goals. Four dimensions of performance are summarized by McGrath’s Task Circumplex [13]. We used measures derived from four tasks reported in Larson et al., 2019 [10]. The *Generate* task required the crew to develop new ideas. The *Choose* task required them to solve a survival scenario with a known solution. The *Negotiate* task required them to resolve an ethical dilemma incorporating multiple conflicting viewpoints. The *Execute* task was a simulation in which a pilot and co-pilot use a joystick to fly a transit vehicle to collection sites, while the other two crew members use virtual reality goggles to complete Extra Vehicular Activity exploring an asteroid’s surface. For the four 30-day missions, these tasks were administered three times, on mission days 10, 15, and 29.

For the four 45-day missions, the tasks were completed four times, on days 13, 18, 27, and 41. This produced a total of 28 observations of team performance.

Social Relations. Social networks were elicited from the crew via sociometric surveys. We included measures of four relational networks to capture a long-standing distinction in the small group literature on task and social needs. *Task affect* and *task hindrance* capture positive and negative working relationships among crew members. Task affect was measured with the prompt: “With whom do you enjoy working?” Task hindrance was elicited with the prompt: “Who makes tasks difficult to complete?” In addition to assessing manifest social relations, we also included two networks capturing behavioral and motivational aspects of teams: *leadership* and *followership*. Leadership was elicited by asking “To whom do you provide leadership?” Followership relations were assessed by asking: “Who do you rely on for leadership?” These prompts yielded four directed networks, each examined in relation to performance. Performance scores from each task session were matched with the network survey most closely proceeding the task. For the 30-day missions, social networks and performance, respectively were measured on the following pairs of days: days 9 and 10, 14 and 15, and 27 and 29. For the 45-day missions, networks and performance events, respectively, were measured on these pairs of days: 11 and 13, 15 and 18, 26 and 27, and 39 and 41.

Network Statistics. While network *density*, the ratio of observed to possible ties, may influence performance directly, where ties are located relative to one another may also impact performance (Fig. 1). Network theories of teams posit the degree of closure, centralization, and subgrouping among team members are important reflections of the quality of teamwork and the team’s capacity to perform [3]. We selected six network statistics based on these three categories, in addition density, to test their impact on crew performance.

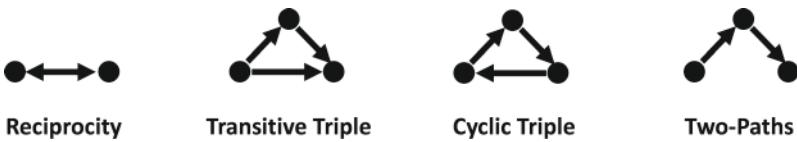


Fig. 1. Basic network structures used in defining network statistics

For the closure category, we measured *normalized transitivity* and *normalized cyclicity*. Normalized transitivity controls for density effects, computing transitivity of a graph as the number of transitive triples divided by the average number of transitive triples in a random graph of the same density. The same approach was used to define a statistic for normalized cyclicity, normalizing by the expected number of cyclic triples.

To examine centralization, we include *closeness centralization*, relative discrepancies in closeness, as defined in Freeman 1978, between team members. Closeness centrality ranks team members based on the proportion of the shortest paths between all team members on which they lie. A team in which each member had a similar closeness centrality would have a lower value of closeness centralization, whereas a team with big differences in closeness centralization would have a high value. Given that our crews include a team member assigned to the role of commander, we also examined centralization using the *relative indegree of commander*, measuring the proportion of all ties directed towards the commander, and the *relative outdegree of commander*, measuring the proportion of all ties directed from the commander towards others.

To examine subgrouping in four-person networks, we consider the amount of two-paths, chains of ties spanning between three crew members, as a way of measuring a tendency against subgrouping. We include a statistic for *normalized two-paths*, using the same approach to normalizing the count of two-paths based on density as used for transitivity and cyclicality.

Controls for Structural and Temporal Patterns. Because teams are observed at multiple points in times, null models need to control for repeated observations of the same team. This is accomplished by including corresponding sufficient statistics in the STERGMs. The first temporal pattern we control for is tie formation, the likelihood that a new tie will form where a tie had not previously existed, by including a *tie likelihood* term in the formation model that counts the number of ties in the network. Similarly, we also control for tie persistence, the likelihood that a tie that has previously existed will continue to exist, by including a tie likelihood term in the persistence model.

Next, we considered the potential effects of extended isolation, in which a crew is forced to work and live together while working long shifts. We included terms for the effects of *elapsed time in isolation* on the likelihood of new ties to form and for effects of elapsed time in isolation on the likelihood of existing ties to persist. We also included a *time between observations* term that examines the effect of the time, in days, since the network data were last collected. This controls for the fact that our data was not collected in uniformly spaced intervals.

Another well documented structural pattern is *reciprocity*. We include measures for the count of reciprocated ties in our STERGMs, to control for tendencies of new ties to be more or less likely to form reciprocated pairs, as well as for existing ties in reciprocated pairs to be more or less likely to persist. Finally, we control for potential homophily effects in our models for formation and persistence terms for *race homophily* and *military experience homophily*.

4.2 Analysis

To develop a null model for our analysis, we estimated one STERGM for each of the four ties, using the temporal networks collected from all teams. STERGMs were fit using Conditional Maximum Likelihood Estimation [7], as implemented

in the tergm package developed for R [8]. To measure the association between a network statistics and performance metric, we used Spearman rank correlation coefficients [17]. P-values were computed for these correlation coefficients using our approach for sampling from the coefficients' null distribution. A total of 250 simulated values of correlation coefficients were utilized.

4.3 Results

Descriptive Results. Intercorrelations between performance scores across the four task dimensions, as well as intercorrelations for the presence of ties in the four networks, are reported in Table 1. In particular, we note that the Spearman rank correlation coefficient between any two of the performance dimensions ranged between -0.62 and 0.38 . Because they are not perfectly correlated, it is critical we separately analyze the associations between network structure and each dimension of team performance. The 28 task affect networks had an average density of 0.83 (s.d. = 0.13). Task hindrance networks had an average density of 0.23 (s.d. = 0.16), leadership networks had an average density of 0.79 (s.d. = 0.17), and followership networks had an average density of 0.63 (s.d. = 0.22).

Table 1. Intercorrelations for team performance and network ties

Performance Measure Intercorrelations					Network Tie Intercorrelations			
	Generate	Choose	Negotiate	Execute	Task Affect	Task Hindrance	Leadership	Followership
Generate	-	0.15	-0.23	-0.11	-	-0.49	0.11	0.39
Choose		-	-0.28	-0.62		-	0.06	-0.28
Negotiate			-	0.38	Leadership	-	-	0.17
Execute				-	Followership			-

Table 2. STERGM results for each social relation

	Task Affect		Task Hindrance		Leadership		Followership	
	Log-Odds	Odds Ratio						
Formation Coefficients								
Tie Likelihood	0.78 (1.04)	2.19	-3.48 (0.87) *	0.03	-0.84 (0.82)	0.43	-0.82 (0.71)	0.44
Elapsed Time in Isolation	-0.26 (0.05) *	0.77	-0.10 (0.02) *	0.91	-0.18 (0.03) *	0.83	-0.12 (0.02) *	0.89
Time Between Observations	-0.02 (0.09)	0.98	0.22 (0.08) *	1.24	0.13 (0.07)	1.14	0.13 (0.06) *	1.14
Reciprocity	1.12 (0.78)	3.06	0.24 (0.50)	1.28	0.65 (0.56)	1.91	0.40 (0.44)	1.50
Race Homophily	1.49 (0.60) *	4.43	-0.05 (0.35)	0.95	0.79 (0.42)	2.19	-0.16 (0.33)	0.85
Military Experience Homophily	0.24 (0.53)	1.27	0.55 (0.36)	1.74	0.55 (0.40)	1.74	0.14 (0.33)	1.15
Dissolution Coefficients								
Tie Likelihood	3.35 (1.12) *	28.64	-1.82 (1.10)	0.16	2.98 (1.11) *	19.70	2.12 (0.78) *	8.35
Elapsed Time in Isolation	0.04 (0.07)	1.04	-0.03 (0.09)	0.97	0.09 (0.06)	1.09	0.02 (0.05)	1.02
Time Between Observations	-0.06 (0.09)	0.94	0.08 (0.12)	1.08	-0.19 (0.09) *	0.82	-0.08 (0.08)	0.93
Reciprocity	-0.01 (0.71)	0.99	-0.73 (1.10)	0.48	0.12 (0.60)	1.13	0.43 (0.49)	1.54
Race Homophily	-0.82 (0.58)	0.44	1.69 (0.86) *	5.39	-0.14 (0.50)	0.87	-0.60 (0.45)	0.55
Military Experience Homophily	-0.90 (0.56)	0.41	2.25 (0.93) *	9.48	-0.32 (0.53)	0.73	0.06 (0.44)	1.06
AIC	178.9		89.64		182.8		193.7	
BIC	203.3		105.1		206.8		216.3	

Standard error in parentheses. * p<0.05, * p <0.10

Table 3. Correlation testing results for each social relation

TASK AFFECT NETWORKS				
	Performance Measures			
	Generate	Choose	Negotiate	Execute
Network Properties				
Density	0.30 (0.37)	-0.30 (0.08) *	0.15 (0.00) *	0.45 (0.00) *
Normalized Transitivity	-0.27 (0.29)	0.22 (0.33)	-0.17 (0.05) *	-0.38 (0.08) *
Normalized Cyclicality	-0.20 (0.39)	0.27 (0.30)	0.31 (0.23)	0.47 (0.00) *
Closeness Centralization	-0.33 (0.00) *	0.23 (0.00) *	-0.13 (0.24)	-0.43 (0.00) *
Relative Indegree of Commander	0.28 (0.12)	0.03 (0.00) *	-0.19 (0.41)	-0.04 (0.01) *
Relative Outdegree of Commander	-0.02 (0.02) *	-0.27 (0.50)	0.42 (0.00) *	0.26 (0.08) *
Normalized Two-Paths	-0.25 (0.50)	0.33 (0.06) *	0.17 (0.77)	-0.23 (0.02) *
TASK HINDRANCE NETWORKS				
	Performance Measures			
	Generate	Choose	Negotiate	Execute
Network Properties				
Density	-0.23 (0.18)	0.30 (0.01) *	0.13 (0.24)	-0.34 (0.06) *
Normalized Transitivity	-0.13 (0.34)	0.11 (0.21)	0.00 (0.51)	-0.17 (0.15)
Normalized Cyclicality	-0.28 (0.10)	0.10 (0.20)	0.27 (0.17)	0.02 (0.35)
Closeness Centralization	-0.19 (0.18)	0.08 (0.16)	-0.14 (0.35)	-0.25 (0.03) *
Relative Indegree of Commander	0.01 (0.72)	0.00 (0.92)	0.38 (0.01) *	0.26 (0.09) *
Relative Outdegree of Commander	0.03 (0.70)	0.36 (0.00) *	-0.30 (0.21)	-0.54 (0.00) *
Normalized Two-Paths	0.25 (0.05) *	0.30 (0.03) *	0.07 (0.31)	-0.19 (0.19)
LEADERSHIP NETWORKS				
	Performance Measures			
	Generate	Choose	Negotiate	Execute
Network Properties				
Density	0.35 (0.17)	-0.40 (0.01) *	0.21 (0.00) *	0.35 (0.00) *
Normalized Transitivity	-0.41 (0.04) *	0.28 (0.26)	-0.20 (0.03) *	-0.20 (0.27)
Normalized Cyclicality	0.20 (0.03) *	-0.32 (0.00) *	-0.08 (0.22)	0.44 (0.00) *
Closeness Centralization	-0.43 (0.00) *	0.42 (0.00) *	-0.15 (0.41)	-0.33 (0.00) *
Relative Indegree of Commander	0.36 (0.05) *	-0.33 (0.28)	-0.10 (0.71)	0.20 (0.10)
Relative Outdegree of Commander	-0.05 (0.00) *	0.40 (0.00) *	-0.14 (0.65)	-0.31 (0.00) *
Normalized Two-Paths	0.08 (0.00) *	-0.39 (0.00) *	0.09 (0.82)	0.50 (0.00) *
FOLLOWERSHIP NETWORKS				
	Performance Measures			
	Generate	Choose	Negotiate	Execute
Network Properties				
Density	0.03 (0.76)	-0.40 (0.04) *	0.36 (0.00) *	0.38 (0.00) *
Normalized Transitivity	-0.05 (0.56)	0.14 (0.36)	0.00 (0.69)	-0.13 (0.36)
Normalized Cyclicality	-0.19 (0.22)	-0.37 (0.03) *	0.36 (0.03) *	0.32 (0.05) *
Closeness Centralization	-0.03 (0.04) *	0.18 (0.00) *	-0.03 (0.88)	-0.22 (0.00) *
Relative Indegree of Commander	-0.21 (0.00) *	0.36 (0.00) *	-0.44 (0.02) *	-0.32 (0.00) *
Relative Outdegree of Commander	0.05 (0.89)	-0.08 (0.97)	0.20 (0.00) *	0.18 (0.19)
Normalized Two-Paths	0.04 (0.14)	-0.29 (0.01) *	0.28 (0.25)	0.16 (0.09) *

P-value in parentheses. * p<0.05, * p <0.10

Correlation Significance Testing. Table 2 presents the parameter estimates from separate STERGM models for each network, which were used to perform sampling from the null distribution. Table 3 reports the Spearman rank correlation coefficient between each of the network statistics and performance, alongside p-values for each, calculated using our method to generate a null distribution from 250 simulated correlation coefficients. For brevity, we describe the network relation with the strongest association with each performance dimension.

Naturally, multiple testing problems [1] occur when performing a large quantity of significance tests. Since we intended this as an exploratory analysis, we did not account for multiple testing effects here. For stricter hypothesis testing, p-values should be adjusted using an approach such as Bonferroni correction [1].

For task affect ties, closeness centralization is inversely related to performance on the Generate task, and positively related to performance on the Choose task. Relative outdegree of the commander is positively associated with performance on the Negotiate task, while cyclicalty is positively related to performance on the Execute task. For task hindrance ties, no network statistics were significantly related to performance on the Generate task. However, hindrance density was positively related to Choose and Execute task performance. Finally, the commander’s relative indegree is positively related to Negotiate task performance.

For leadership ties, closeness centralization is inversely related to performance on the Generate task, but positively related to performance on the Choose task. Leadership density is positively related to performance on the Negotiate task, whereas leadership two-paths are positively associated with performance on the Execute task. For followership ties, the relative indegree of the commander is inversely related to performance on the Generate, Negotiate, and Execute tasks, while followership density is positively related to Choose task performance.

5 Discussion

These findings illustrate how STERGM can be used to account for endogeneity due to time when correlating network statistics with an exogenous network level outcome variable. This approach considers the association between network properties and team performance by decomposing the network into individual ties to be modeled. The STERGM-based sampling from the null distribution controls for complex interdependencies between ties (e.g. reciprocity, closure, or tendency of ties to persist over time). We model the network statistic not as a single continuous variable, but as a consequence of a number of discrete ties that have complex interdependencies with one another. STERGM therefore provides a flexible framework to control for various types of interdependencies that have been well-established as occurring across many real-world social networks.

We apply simulation from the null distribution to answer the question: Which network patterns predict team performance? The results suggest elements of closure, centralization, and subgrouping along different relations affect performance. Additionally, we observed multiple cases where a structure that benefited one type of performance undermined another. Further work is needed to discover the underlying social dynamics linking networks to team performance.

Correlations on small sample sizes, such as ours, are difficult to interpret because of the potential for spurious findings. The approach we employ for generating an empirical p-value, which takes into account temporal and other structural dynamics, provides a means for understanding the probability of finding such an effect. In doing so, it serves as a tool to help interpret effects when working with a moderate to small sample of networks.

Though we demonstrate a tool for statistical testing on small network samples, it has a number of limitations. First, the method needs to be compared to existing multilevel techniques which also account for temporal endogeneity [2, 11, 16]. Second, this approach needs to be explored as it applies to smaller or larger samples. How few measurements would merit this approach, and how many measurements could it be usefully applied to? Simulation studies could help explore these questions.

Acknowledgements. This material is based upon work supported by NASA under award numbers NNX15AM32G, NNX15AM26G, and 80NSSC18K0221. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.

References

1. Aickin, M., Gensler, H.: Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am. J. Public Health* **86**(5), 726–728 (1996)
2. Bell, S.T., Fisher, D.M., Brown, S.G., Mann, K.E.: An approach for conducting actionable research with extreme teams. *J. Manage.* **44**(7), 2740–2765 (2018)
3. Crawford, E.R., LePine, J.A.: A configural theory of team processes: accounting for the structure of taskwork and teamwork. *AMRO* **38**(1), 32–48 (2013)
4. Freeman, L.C.: Centrality in social networks conceptual clarification. *Soc. Netw.* **1**(3), 215–239 (1978)
5. Krackhardt, D.: QAP partialling as a test of spuriousness. *Soc. Netw.* **9**(2), 171–186 (1987)
6. Krackhardt, D.: Predicting with networks: nonparametric multiple regression analysis of dyadic data. *Soc. Netw.* **10**(4), 359–381 (1988)
7. Krivitsky, P.N., Handcock, M.S.: A separable model for dynamic networks. *J. R. Stat.* **76**(1), 29–46 (2014)
8. Krivitsky, P.N., Handcock, M.: tergm: Fit, Simulate and Diagnose Models for Network Evolution Based on Exponential-Family Random Graph Models. The Statnet Project (<https://statnet.org>) R package version 3 (0) (2019)
9. Landon, L.B., Slack, K.J., Barrett, J.D.: Teamwork and collaboration in long-duration space missions: going to extremes. *Am. Psychol.* **73**(4), 563 (2018)
10. Larson, L., Wojcik, H., Gokhman, I., DeChurch, L., Bell, S., Contractor, N.: Team performance in space crews: Houston, we have a teamwork problem. *Acta Astronautica* **161**, 108–114 (2019)
11. Lazega, E., Snijders, T.A.: Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications, vol. 12. Springer, Cham (2015)

12. Lungceanu, A., Carter, D.R., DeChurch, L.A., Contractor, N.S.: How team interlock ecosystems shape the assembly of scientific teams: a hypergraph approach. *Commun. Methods Meas.* **12**(2–3), 174–198 (2018)
13. McGrath, J.E.: Groups: Interaction and Performance, vol. 14. Prentice-Hall, Englewood Cliffs (1984)
14. Mukherjee, S., Uzzi, B., Jones, B., Stringer, M.: A new method for identifying recombinations of existing knowledge associated with high-impact innovation. *J. Prod. Innov. Manag.* **33**(2), 224–236 (2016)
15. Salas, E., Tannenbaum, S.I., Kozlowski, S.W., Miller, C.A., Mathieu, J.E., Vessey, W.B.: Teams in space exploration: a new frontier for the science of team effectiveness. *Curr. Dir. Psychol. Sci.* **24**(3), 200–207 (2015)
16. Snijders, T.A.: Multilevel Analysis. Springer, Heidelberg (2011)
17. Spearman, C.: The proof and measurement of association between two things. *Am. J. Psychol.* **15**(1), 72–101 (1904)

Author Index

A

- Aasa, Anto, 415
Abramski, Katherine, 387
Adamic, Lada, 451
Albane, Saadia, 179
Antone, Brennan, 1018
Antunes, Nelson, 203
Araújo, Tanya, 316, 547
Arceo-May, Ezequiel, 403
Asatani, Kimitaka, 709
Ashihara, Kazuki, 28
Aste, Tomaso, 573

B

- Baffier, J.-F., 684
Bagdasar, Ovidiu, 440
Bell, Suzanne, 1018
Bello-Orgaz, Gema, 427
Bentert, Matthias, 494
Bernstein, Abraham, 481
Bhaskar, Kanishka, 762
Blumenstock, Joshua E., 451
Bockholt, Mareike, 81
Boldi, Paolo, 291
Bonato, Anthony, 105
Bonchi, Francesco, 53
Bothorel, Cécile, 507
Bourgoin, Jeremy, 152
Bouthinon, Dominique, 228
Brisson, Laurent, 507
Brunie, Lionel, 599
Bugrim, Andrej, 751

C

- Caamaño, Antonio J., 895
Camacho, David, 427
Canillas, Rémi, 599
Carmona, Chris U., 722
Catanese, Salvatore, 440
Cavallaro, Lucia, 440
Chandra, Anita, 117
Charbey, Raphaël, 507
Chi, Guanghua, 451
Chidean, Mihaela Ioana, 895
Chu, Chenhui, 28
Clark, Ruaridh, 842
Contractor, Noshir, 969, 1018
Cornejo-Bueno, Sara, 895
Courtain, Sylvain, 40

D

- de Bruin, Gerrit Jan, 140
de Jonge, Edwin, 280
De Meo, Pasquale, 440
De Smedt, Johannes, 931
De Weerdt, Jochen, 931
DeChurch, Leslie, 969, 1018
Deeva, Galina, 931
DeLellis, Pietro, 535
Deritei, David, 905
Devezas, José, 3
Diwakar, Shyam, 762
Doi, Shohei, 611
Duarte-Barahona, Raúl, 403

E

- Eades, Peter, 216
 Eikmeier, Nicole, 105
 Eubank, Stephen, 955
 Evtushenko, Anna, 586

F

- Ferro, Alfredo, 255
 Ficara, Annamaria, 440
 Finke, Jorge, 802
 Fiumara, Giacomo, 440
 Fornaia, Andrea, 164
 Friedrich, Hanno, 363
 Froese, Vincent, 469

G

- Gaito, Sabrina, 152
 Galimberti, Edoardo, 53
 Garcia, Susana, 645
 Garibay, Ivan, 633
 Garlatti, Serge, 507
 Gastner, Michael T., 586
 Gerritsen, Charlotte, 337
 Ghaffar, Faisal, 268, 995
 Giannini, Lorenzo, 535
 Giatsidis, Christos, 1007
 Giger, Markus, 152
 Gilliot, Jean-Marie, 507
 Gleich, David F., 105
 Göbel, Maximilian, 547
 Gómez-Zará, Diego, 969
 Grady, Caitlin, 645
 Grislain, Quentin, 152
 Grüter, Rolf, 481
 Gündüz, Semra, 376
 Guo, Tianjian, 203
 Gupta, Aryaman, 1018

H

- Hajjamoosha, Paria, 919
 Hasan, Omar, 599
 Hecking, Tobias, 129
 Helic, Denis, 242, 350
 Hiir, Hendrik, 415
 Himmel, Anne-Sophie, 494
 Hong, Seok-Hee, 216
 Hoppe, H. Ulrich, 129
 Horn, Abigail, 363
 Hossain, Md Ekramul, 774
 Hu, Jingming, 216
 Huerta-Quintanilla, Rodrigo, 403
 Hurley, Neil, 268
 Hutchison, Marc, 387

I

- Ichhaporia, Rustom, 969
 Interdonato, Roberto, 152
 Irfan, Mohammad T., 66

J

- Jabeen, Fakhra, 337
 Jain, Brijnesh, 469
 Jiang, Lin, 883
 Jolad, Shivakumar, 854

K

- Katenka, Natallia, 387
 Keane, Peter, 995
 Khalife, Sammy, 656
 Khan, Arif, 774
 Khedouci, Hamamache, 179
 Kimura, Sonoko, 709
 Kiniwa, Jun, 561
 Kireev, Maxim, 868
 Kivimäki, Ilkka, 40
 Knyazeva, Irina, 868
 Koncar, Philipp, 350
 Koponen, Ismo T., 15
 Korotkov, Alexander, 868
 Kumar, Amish, 789
 Kurizaki, Shuhei, 611

L

- Lebichot, Bertrand, 40
 Leopold, Judith, 751
 Liotta, Antonio, 440
 Louçã, Francisco, 316
 Louçã, Jorge, 325
 Lu, Shaofeng, 883
 Lutzeyer, Johannes F., 191

M

- Macdonald, Malcolm, 842
 Madeddu, Chiara, 53
 Maiti, Abyayananda, 117
 Malik, Rehan, 105
 Mallégol, Antoine, 507
 Malone, David, 995
 Manjalavil, Manju Manohar, 944
 Marathe, Madhav, 955
 Martinez-Jaramillo, Serafin, 722
 Martorana, Emanuele, 255
 Masharipov, Ruslan, 868
 Maust, Joel, 751
 McGeown, William, 842
 McNichols, Logan, 94
 Medina-Kim, Gabriel, 94

Mejia, Alfonso, 645
Micale, Giovanni, 255
Michaud, Jérôme, 305
Migler, Theresa, 94
Miyapuram, Krishna Prasad, 854
Mizuno, Takayuki, 611
Mohammadpour, Paniz, 645
Molter, Hendrik, 519
Mongiovì, Misael, 164
Mortveit, Henning, 955
Mukhopadhyay, Dyutiman, 854
Murata, Tsuyoshi, 815

N

Nagahara, Hajime, 28
Nair, Lakshmi, 762
Nakashima, Yuta, 28
Natera Orozco, Luis Guillermo, 905
Neves, David, 316
Nguyen, Viet Lien, 94
Nichterlein, André, 494
Niedermeier, Rolf, 469, 494, 519
Nikolay, Makarenko, 868
Nikolova, Niia, 842
Nousiainen, Maija, 15
Nunes, Sérgio, 3

O

Okubo, Noriko, 28
Óskarsdóttir, María, 931
Ostertag-Hill, Luca, 66

P

Panagopoulos, George, 1007
Panizo-LLedot, Aíngel, 427
Parmentier, P., 684
Phillips, Andrew C., 66
Pipiras, Vladas, 203
Prieto, Luís, 895
Pulvirenti, Alfredo, 255

Q

Quimbaya, Mauricio, 802

R

Rajtmajer, Sarah, 645
Rakhimberdina, Zarina, 815
Ramadurai, Gitakrishnan, 944
Rapp, Christian, 94
Ravindran, Balaraman, 944
Read, Jesse, 656
Renken, Malte, 469, 519
Renoust, Benjamin, 28, 671, 684
Rocha, Camilo, 802

Romero, Miguel, 802
Rossa, Fabio Della, 535
Ruffieux, Philippe, 507
Ruffo, Giancarlo, 53
Ruprechter, Thorsten, 242
Ruus, Risko, 982

S

Saerens, Marco, 40
Salcedo-Sanz, Sancho, 895
Saluveer, Erki, 415
Sánchez-Restrepo, Harvey, 325
Sandoh, Hiroaki, 561
Santini, Guillaume, 228
Santos, Tiago, 242
Sarasua, Cristina, 481
Sarrat, Laurent, 599
Sasidharakurup, Hemalatha, 762
Scheinert, Steven R., 633
Schlaich, Tim, 363
Sharma, Rajesh, 415, 982
Shvydun, Sergey, 736
Singh, Chakresh Kumar, 854
Skianis, Konstantinos, 1007
Škrlj, Blaž, 671
Slimani, Hachem, 179
Soldano, Henry, 228
State, Bogdan, 451
Sterchi, Martin, 481
Sugawara, Toshiharu, 709
Svyatoslav, Medvedev, 868

T

Takemura, Noriko, 28
Takes, Frank W., 140
Talebzadehhosseini, Seyyedmilad, 633
Tang, Jie, 1007
Taskin, Yassin, 129
Thorburn, Joshua, 427
Torregrosa, Javier, 427
Tramontana, Emiliano, 164
Treur, Jan, 337, 697, 827
Tripathi, Richa, 854
Turiel, Jeremy D., 573

U

Uddin, Shahadat, 774
Ullah, Nimat, 697
Urich, Christian, 919

V

van den Herik, H. Jaap, 140
van der Laan, Jan, 280
Vancso, Anna, 905

Vasarhelyi, Orsolya, [905](#)
Vazirgiannis, Michalis, [656](#), [1007](#)
Veenman, Cor J., [140](#)
Verba, Michael A., [620](#)
Viard, T., [684](#)
Vigna, Sebastiano, [291](#)
Vullikanti, Anil, [955](#)

W

Walden, Andrew T., [191](#)
Wang, Xiaoliang, [883](#)
Wu, Qigang, [883](#)

X

Xue, Fei, [883](#)
Xypolopoulos, Christos, [1007](#)

Y

Yadav, Gitanjali, [789](#)

Z

Zheltyakova, Maya, [868](#)
Zignani, Matteo, [152](#)
Zweig, Katharina A., [81](#)