IEEE Symposium Series on Computational Intelligence SSCI 2021

Paper ID: 1202

Exploring Unsupervised Learning Methods for Automated Protocol Analysis

Arijit Dasgupta[1], Yi-Xue Yan[2], Clarence Ong[1], Jenn-Yue Teo Bugsy[3], Chia-Wei Lim Andrew[3]

NUS National University of Singapore [1]

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE [2]

DSO NATIONAL LABORATORIES [3]

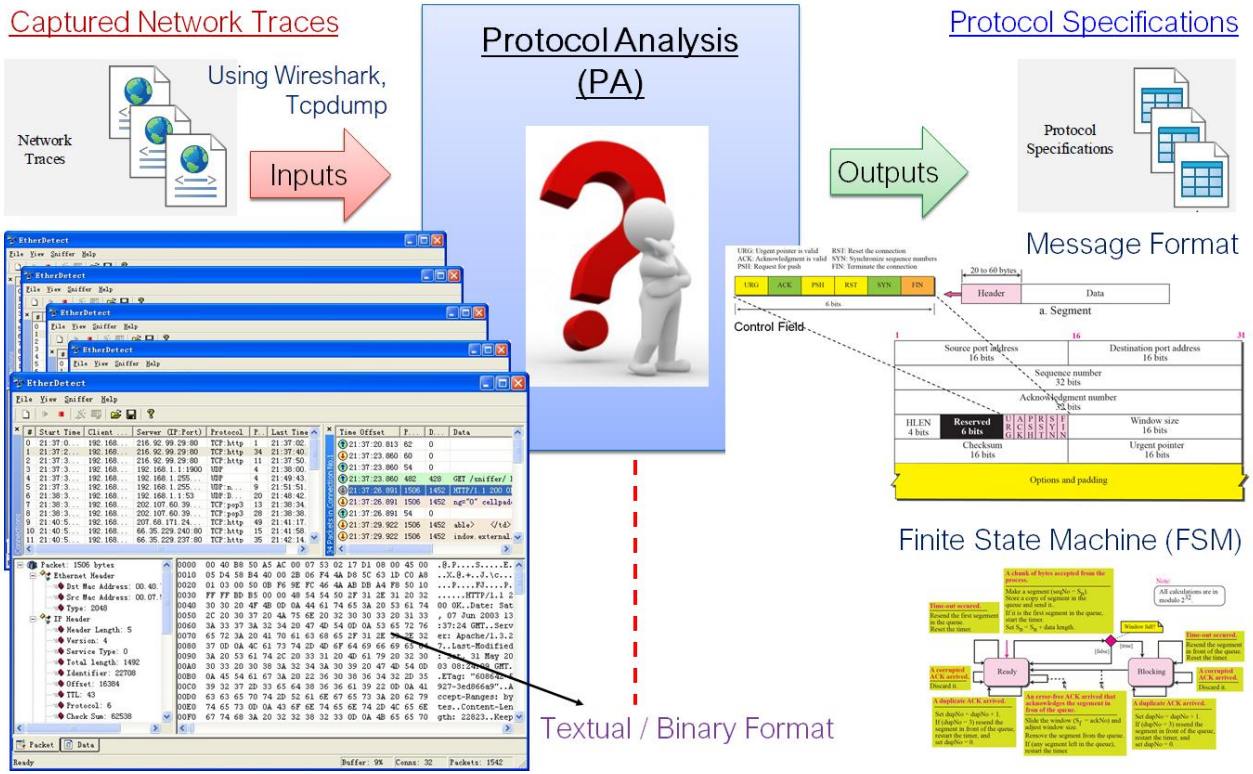
IEEE

Agenda

- **Introduction**
  - Background
  - Motivation
  - Contributions
- **Related Works**
- **Proposed APA Framework**
  - Architecture
  - Data Pre-processing
  - Feature Extraction
  - Unsupervised Clustering
- **Experiments**
  - Datasets
  - Experimental Setup
  - Performance Metrics
- **Results & Discussion**
  - Comparing Tokenisation methods
  - Testing hyperparameter automation techniques
  - Overall performance comparison
- **Conclusion & Future Works**

Introduction: Background

- What is Protocol Analysis (PA) via Static Traffic Analysis (STA)?
- Derive detailed specifications of unknown protocols.
- Use Cases: Network Resource management, IoT Interoperability, Protocol Security Audit, Simulation and Conformance Testing, etc.



Introduction: Motivation

Need for Automated Protocol Analysis (APA) via STA

- **Traditionally PA is done manually by experts and is very time consuming.**
  - ➢ Taking months or even years for complex unknown protocols.
  - ➢ Additional challenge to recruit train and retain PA experts.

- **Need for APA was first raised since 2007. (**Discoverer [1])
  - ➢ Proposed approaches inspired by diverse disciplines such as Bioinformatics, Natural Language Processing (NLP) and Artificial Intelligence (AI).
  - ➢ Significantly improved the efficiency of analysis process and reduce the reliance on human experts.
  - ➢ NETZOB [2] most comprehensive open-source APA framework to date. (Our baseline for comparisons)

Introduction: Key Contributions

Assume no prior knowledge and explore various unsupervised methods with our key contributions as follow:

- **Developed a comprehensive APA framework.**
  - ➢ For evaluation of various combinations of state-of-the-art unsupervised feature extraction & clustering methods.

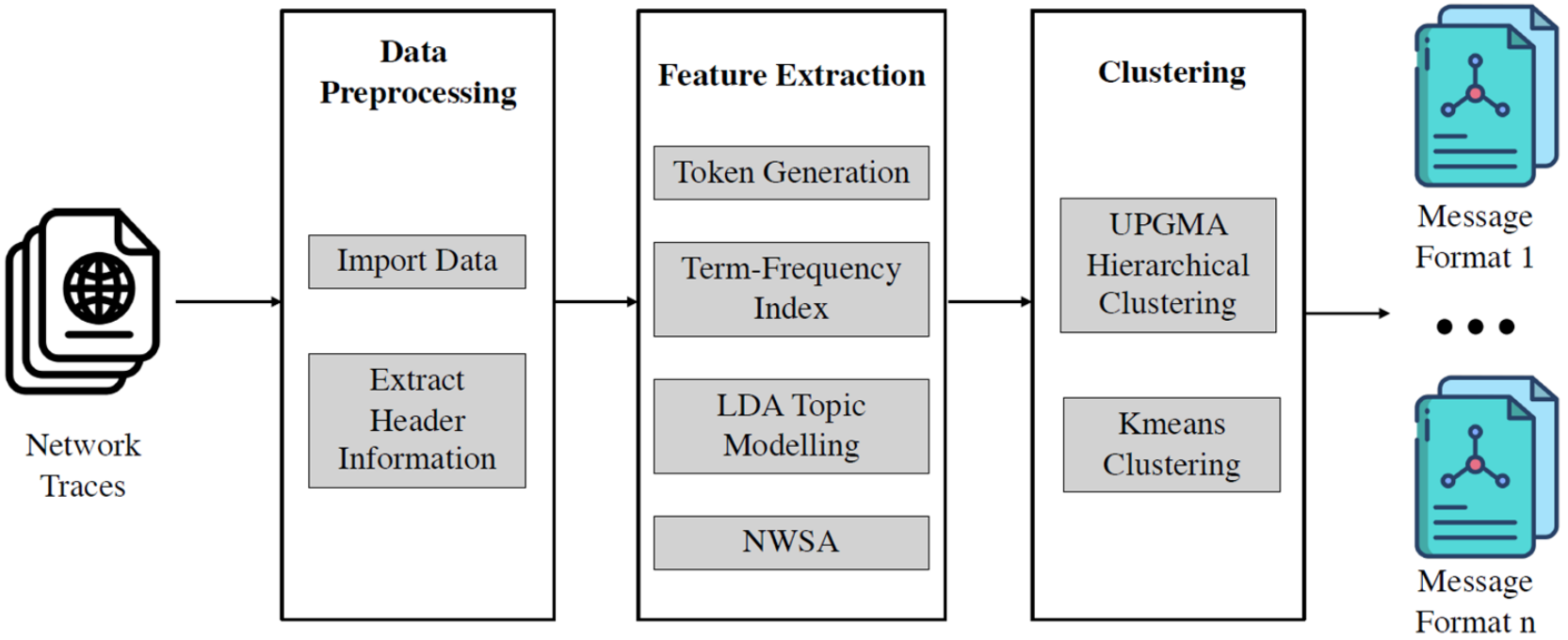
- **Proposed novel methods and insights.**
  - ➢ Automated techniques for model optimization for APA.
  - ➢ Insights into techniques for automatic field-based tokenization.

- **Comprehensive experimentation and proposed hybrid approach.**
  - ➢ Unsupervised automated features extraction and unknown protocol message clustering for APA.
  - ➢ Improved hybrid approach over state-of-the-art open-source APA tools (e.g. NETZOB and other related works).

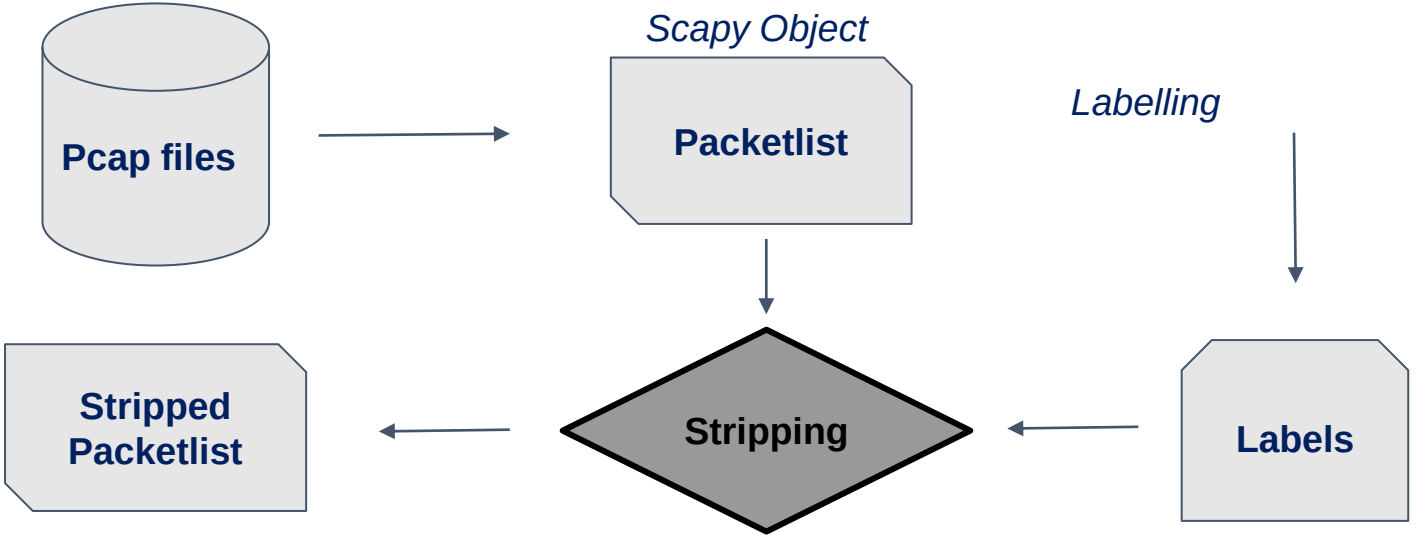
Related Works

- Existing work typically focus on inferring message format types form packets of a single unknown protocol, using the entire packet for feature extraction
- However, packet streams can belong to multiple unknown protocols and it is desirable to extract header information from packets automatically
- Discoverer [1] used tokenisation, recursive and merging clusters
- NETZOB [2] uses sequence alignment to infer message formats and cluster protocols
- These techniques require expert knowledge and a range of assumptions that may not apply for a suite of completely unknown protocols
- Existing techniques like Latent Dirichlet Allocation [3], N-grams and frameworks like NEMESYS [4] have also been used for field-based tokenisation

Proposed APA Framework: Overview



Proposed APA Framework: Data Preprocessing



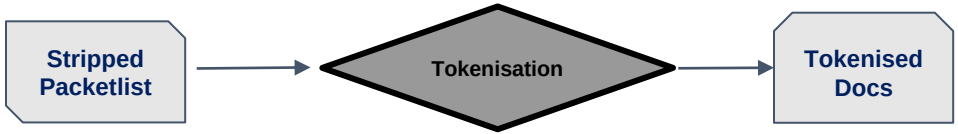
Proposed APA Framework: Tokenisation

Proposed Methods

- **N-grams (3-grams)**

A	B	C	D	E	F	G

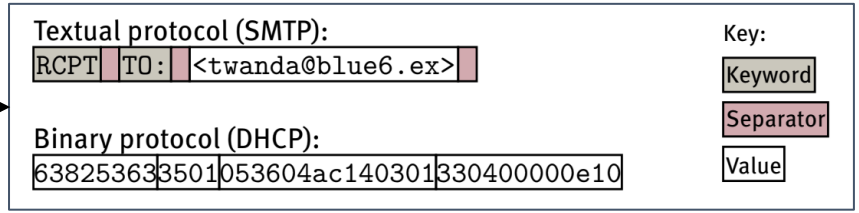
*Only the First X Bytes used (header length)**



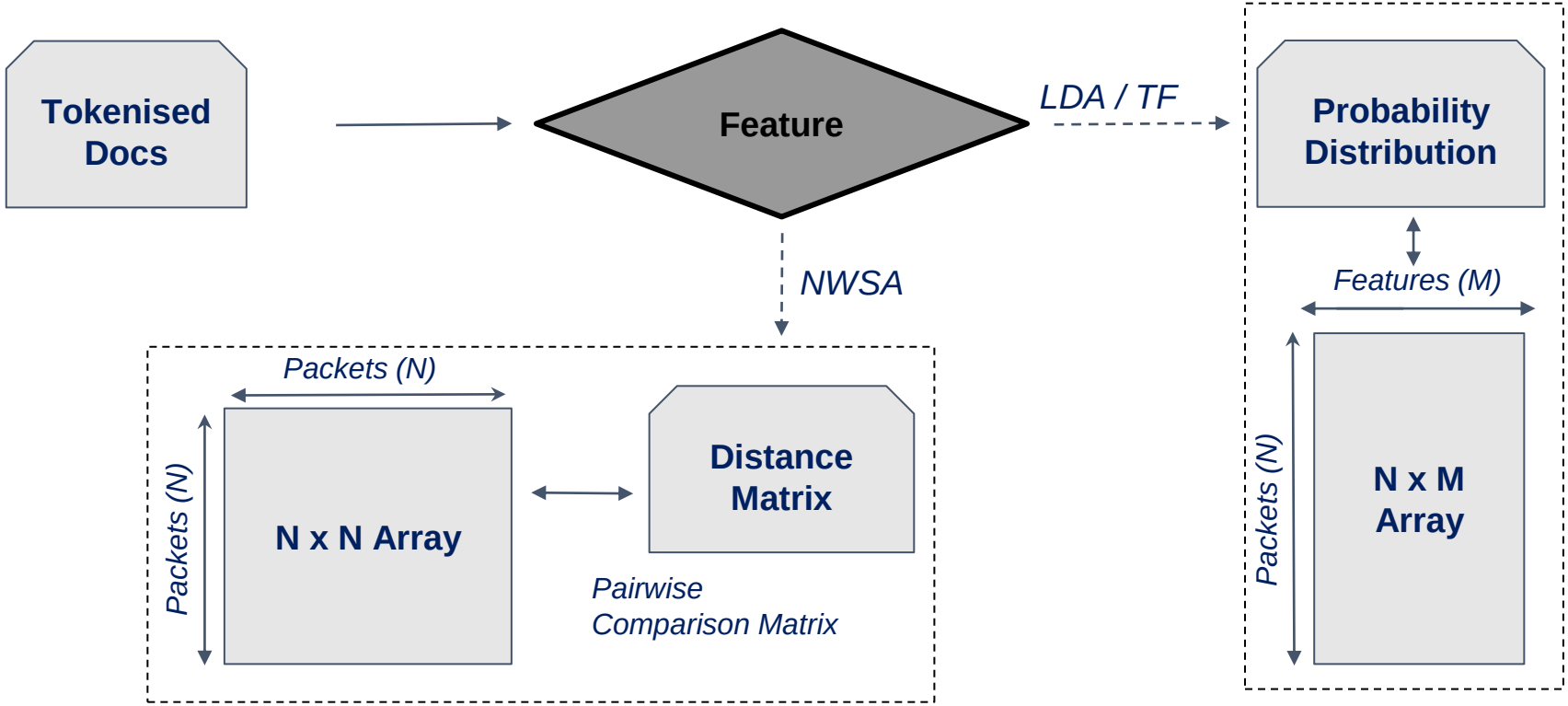
- **NEMESYS [4] (no a priori info required)**
  - ➢ Tokenization based on **field boundary** inference

Textual protocol (SMTP):
RCPT TO: <twanda@blue6.ex>

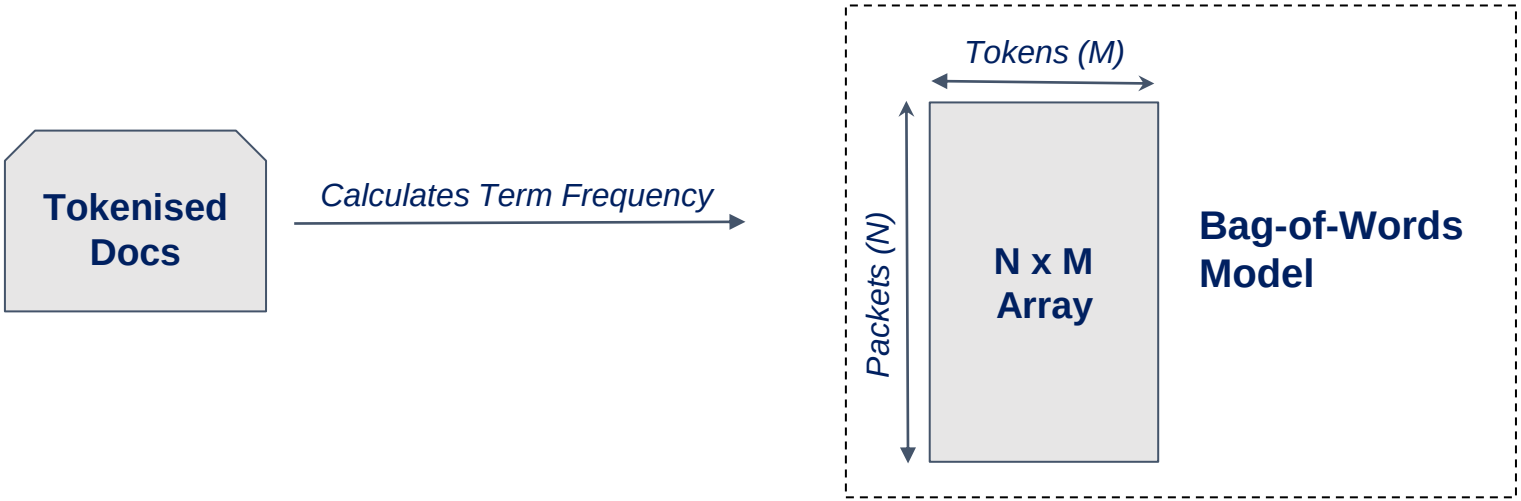
Binary protocol (DHCP):
63825363501053604ac140301330400000e10



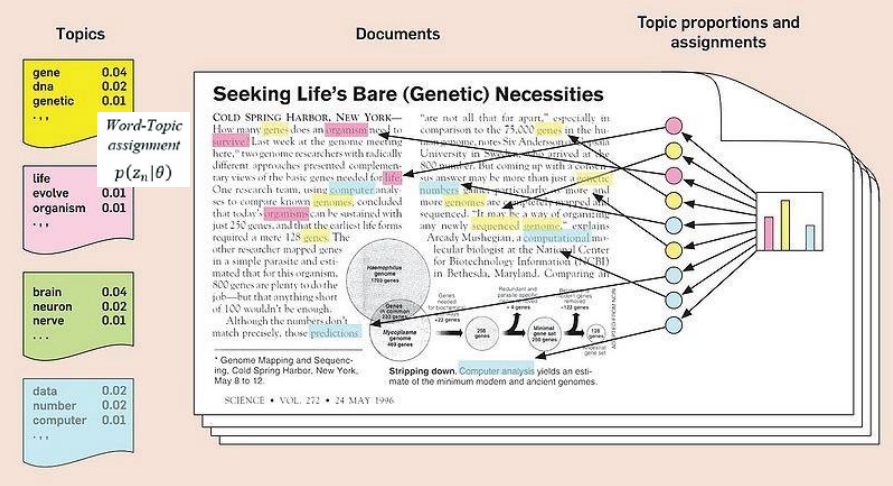
Proposed APA Framework: Feature Extraction (1/6)

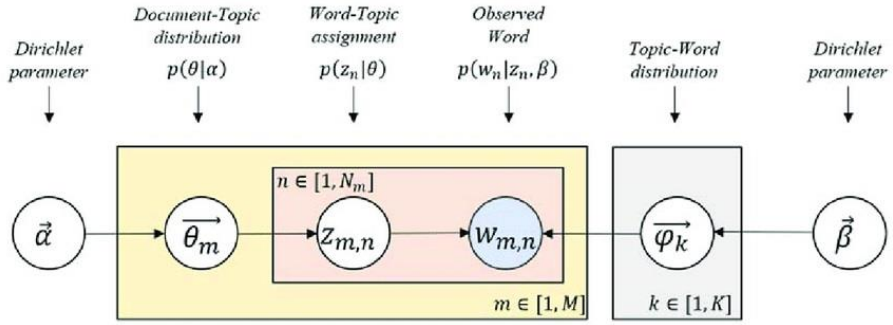


Proposed APA Framework: Feature Extraction (2/6)



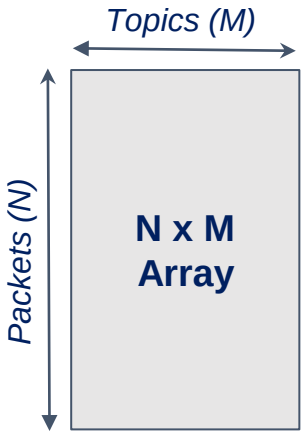
Proposed APA Framework: Feature Extraction (3/6)





Latent Dirichlet Allocation (LDA)
Used in NLP (Topic Modelling)

Topic : Words observed together frequently and coherently.



Proposed APA Framework: Feature Extraction (4/6)

Needleman-Wunsch Sequence Alignment (NWSA)

Example:

Mismatch (-1), Match (+1) and Gap (0)

Pairwise comparison between sequences to derive Distance Matrix used for subsequent Clustering step.

Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Proposed APA Framework: Feature Extraction (5/6)

Key hyperparameters for tuning

- LDA Topic Size and Protocol Header Length are key parameters that will affect clustering performance (very sensitive)
- No way to tune manually when Protocol Messages are unknown or unlabelled
- No obvious (unsupervised) metric(s) to tune these parameters
- Could only be tuned with ground truths

Proposed APA Framework: Feature Extraction (6/6)

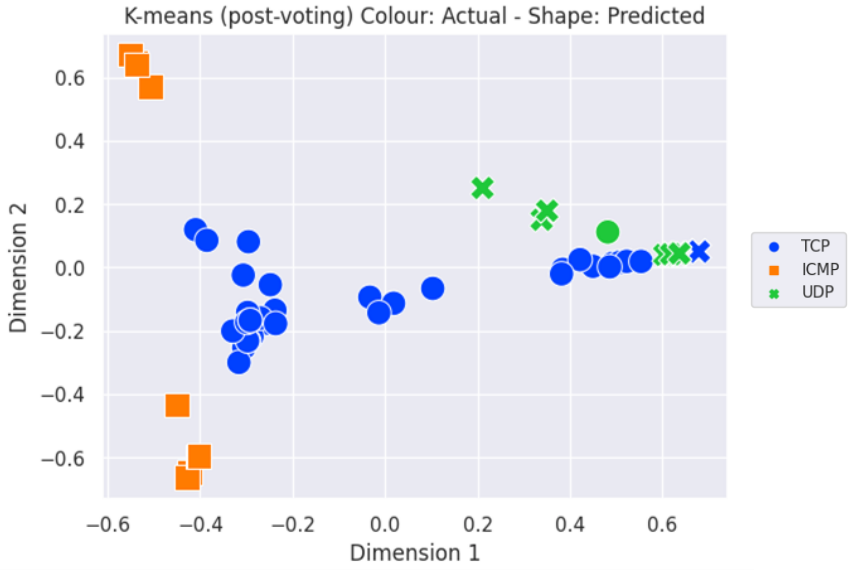
Proposed Solution: Exclusivity and Semantic Coherence

- **Estimating Optimum LDA Topic Size**
  - ➢ Identifying the Topic Size that generated the highest **Mean Exclusivity** and **Mean Semantic Coherence**
  - ➢ Intuitively, the **best topic size** should generate topics that are **most different** and showcased the **least coherence** from one another.

- **Estimating Optimum Protocol Header Length**
  - ➢ Estimation of an appropriate header length by iterating through and observing for a **distinct "optimum topic size"**.
  - ➢ Intuitively, the more **isolated** the **furthest point** is from the other points, the more likely the appropriate header length is being used in that iteration.

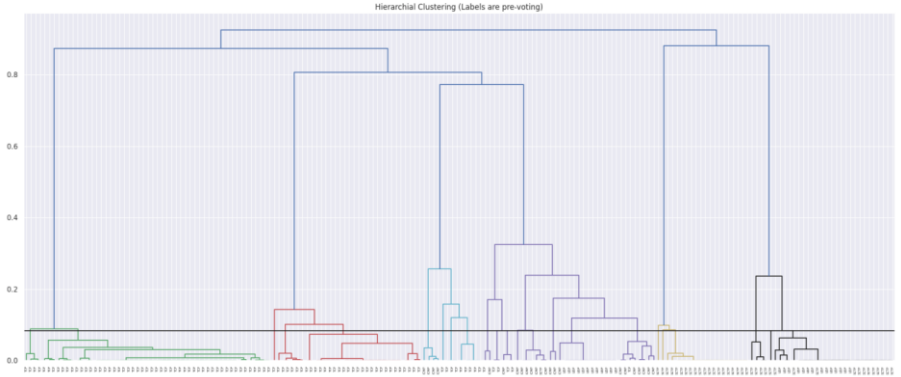
Proposed APA Framework: Clustering (1/3)

Cluster Method #1: K-means



*Only compatible with LDA & TF**

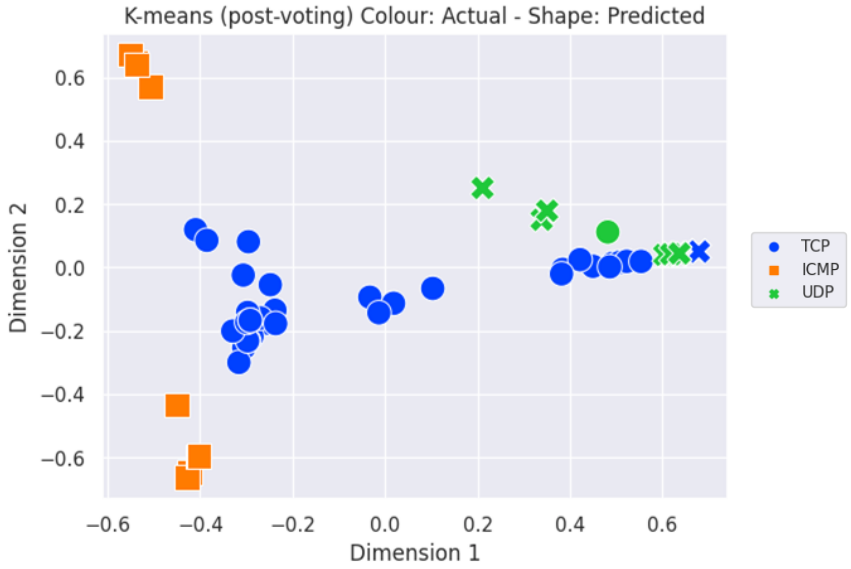
Cluster Method #2: Unweighted pair group method with arithmetic mean (UPGMA) Clustering



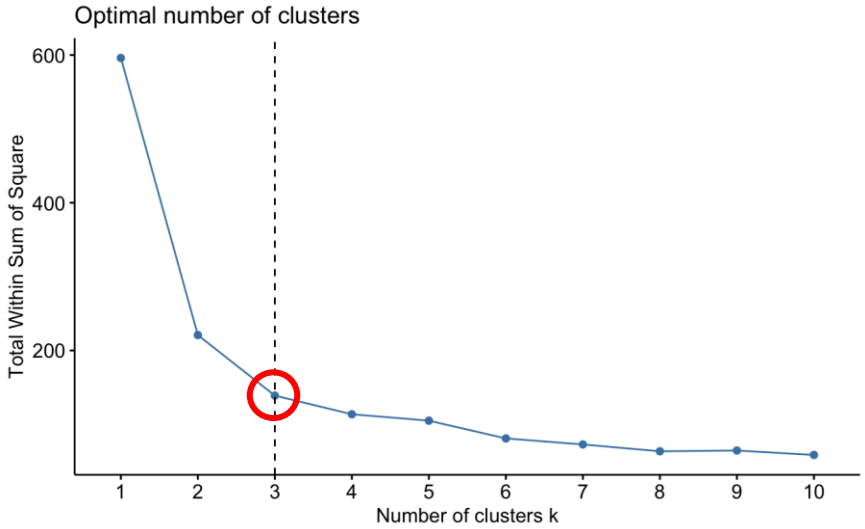
*Fully compatible with NWSA. LDA & TF outputs must be converted using a distance metric pairwise comparison**

Proposed APA Framework: Clustering (2/3)

Determining Number of Clusters



*Only compatible with LDA & TF**



Elbow Method used to determine the optimal number of clusters (conventional)

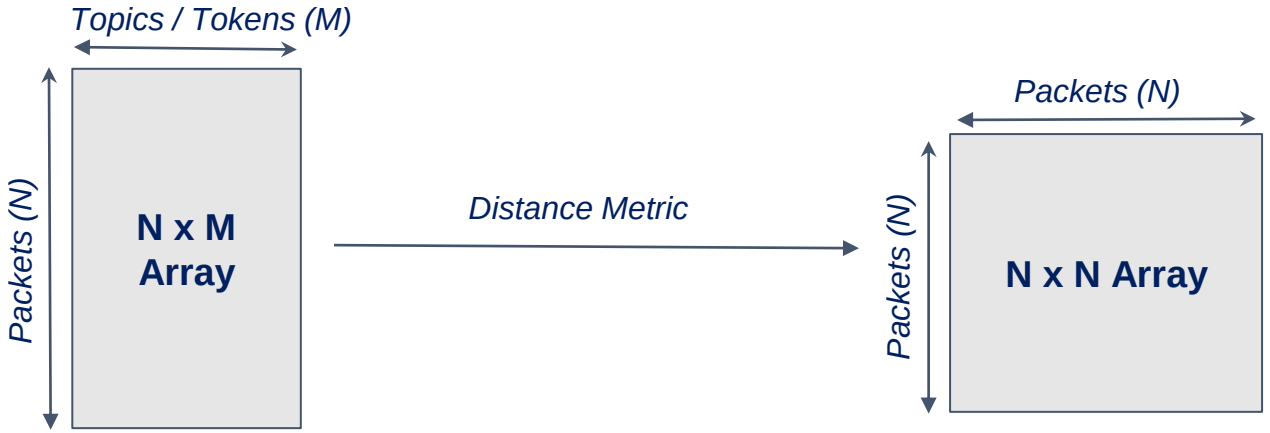
Proposed APA Framework: Clustering (3/3)

UPGMA Compatibility Preparation

Cosine Similarity Distance Metric

- Takes into account the **angle** formed **between** any 2 **vectors** representing their respective protocol messages.
- Vectors that are generally pointing in the **same direction** are regarded to be **similar**, while those in **opposite directions** are regarded as **dissimilar**.

LDA / TF Output



Experiments: Dataset

- **9** datasets of **200** packets each
- Each protocol is assumed to be unknown
- **5** Protocol-level clustering datasets:- *TCP, SCTP, ICMP, HTTP & DNS* (fine-grain)
- **4** OSI-level clustering datasets:- *Link, Transport, App (Text) & App (Binary)*
- Intentional dataset design choices :-
  - Few packets (200) per dataset to give the added challenge of limited data
  - Imbalanced datasets to replicate real-world conditions

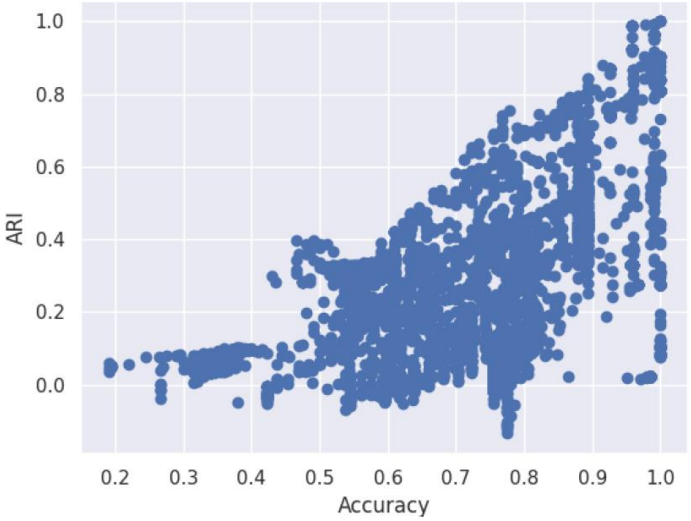
Dataset	Protocols/Types
Link Layer Protocols	Point to Point Protocol (PPP), Link Layer Discovery Protocol (LLDP), IEEE 802.11, Ethernet
Transport Layer Protocols	Internet Control Message Protocol (ICMP), Transmission Control Protocol (TCP), User Datagram Protocol (UDP), Stream Control Transmission Protocol (SCTP)
Application Layer Protocols (Text)	Domain Name Server (DNS), Routing Information Protocol (RIP), Transport Layer Security (TLS)
Application Layer Protocols (Binary)	Trivial File Transfer Protocol (TFTP), Hypertext Transfer Protocol (HTTP), Simple Mail Transfer Protocol (SMTP)
TCP Message Types	7 TCP message types e.g. ACK, PSH ACK, SYN, RST, FIN ACK
SCTP Chunk Types	16 SCTP chunk types e.g. INIT, COOKIE ECHO, DATA, HEARTBEAT, ASCONF, ACK
ICMP Types	4 ICMP types e.g. Reply, Request, Destination Unreachable, TTL Exceeded
HTTP Methods	3 HTTP methods e.g. 200 OK, GET, POST
DNS message types	4 DNS message types e.g. Query, Response Refused, Response No Error, Response No Such Name

Experiments: Setup

- **3** objectives
  - Compare NEMESYS [4] and N-grams across 9 datasets
  - Evaluate proposed LDA topic size and protocol header length automation techniques
  - Comparing the APA performance across 5 different combinations of the framework
- **5** Combinations of the APA framework
  - NWSA + UPGMA <N-grams>(methodology used in NETZOB [2])
  - LDA + K-means <N-grams>
  - LDA + UPGMA <N-grams>
  - TF + UPGMA <N-grams>
  - Hybrid approach [TF + UPGMA <N-grams>, LDA (for App Bin protocols), <NEMESYS[4]> (for App Text protocols)]
- Hierarchical hyperparameter tuning was done with grid-search
- Run through all datasets with 10 seeded runs each
- Intel(R) Core(TM) i7-8550U CPU processor @1.80GHz and 16GB RAM, Windows 10
- 42.25s < Run Time < 357.62s with a mean of 138.18s

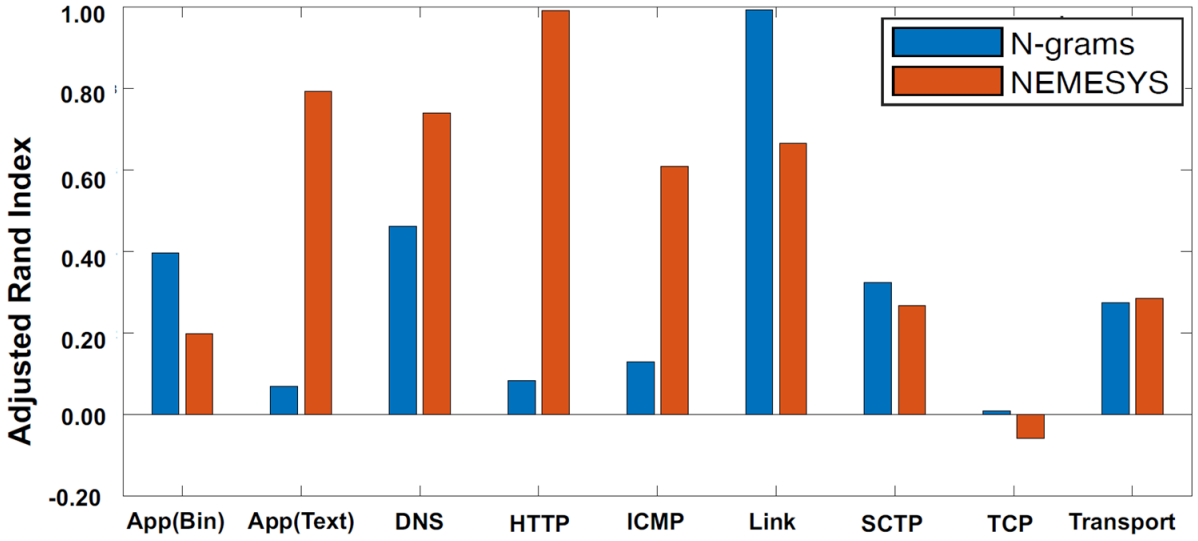
Experiments: Performance Metric

- As this is a purely unsupervised problem, standard classification metrics are not fully compatible
- There is a way to artificially determine the accuracy (discussion detailed in the paper)
- We chose to use the Adjusted Rand Index (ARI), ranges from -1 to 1
- ARI compares how similarly the two clusters are grouped and adjust for chance
- Empirically determined ARI > 0.4 as an indicator of satisfactory performance



> 13,000 runs of the proposed APA framework

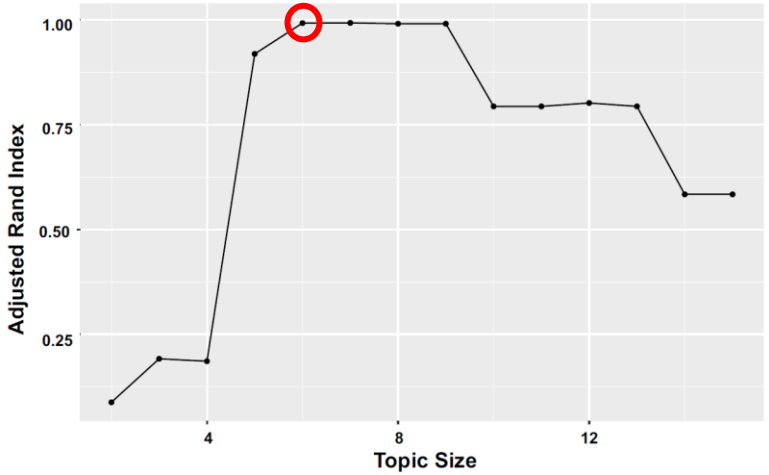
Results & Discussion: Tokenisation Performance



NEMESYS tokenization is more effective for application layer textual protocols

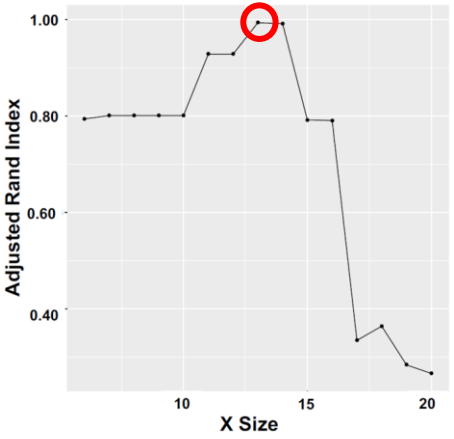
Results & Discussion: Hyperparameter Optimisation

Optimising LDA Topic Size

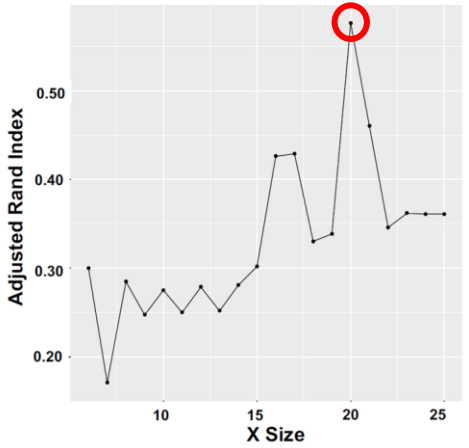


Link layer dataset

Optimising Protocol Header Length



Link layer dataset

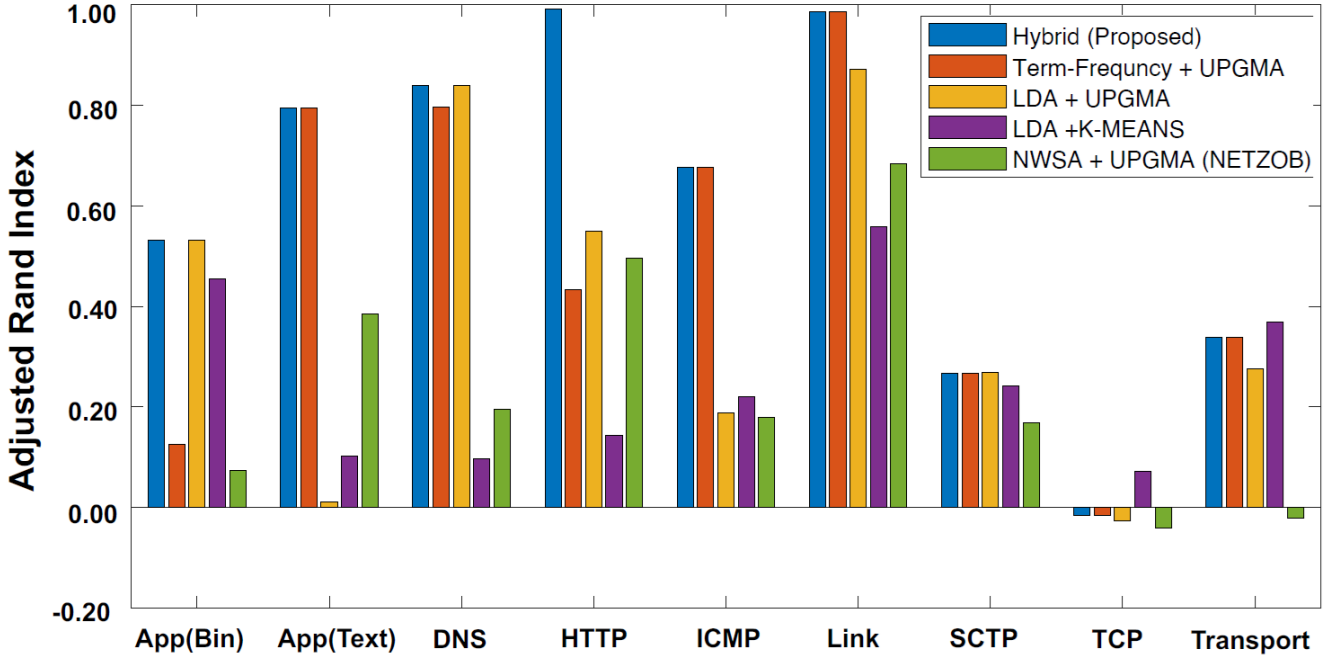


Transport layer dataset

Automation techniques show strong promise and have been validated

IEEE

Results & Discussion: Overall Results



Hybrid approach performed best in 7 of 9 datasets with ARI > 0.4 in 6 of 9 datasets

Conclusion & Future Work

- Proposed a comprehensive APA framework and evaluated various combinations of feature extraction and clustering methods, including those used by NETZOB [2]
- Proposed hybrid approach performed best in 7 of 9 datasets with ARI > 0.4 for 6 of 9 datasets
- This result proves the robustness and generalising ability of our proposed hybrid approach.
- We also validated our proposed automated optimisation methods, for both LDA topic size and extracted protocol header length, that is crucial for practical deployment
- With recent advances in Deep Learning, like Deep Auto-Encoders for automated features extraction, it will be exciting to explore the application of these advanced Machine Learning (ML) methods for unsupervised learning in APA
- Explore more areas of APA with ML

References (selected)

[1] W. Cui, J. Kannan, and H. J. Wang, “Discoverer: Automatic protocol reverse engineering from network traces.” in USENIX Security Symposium, 2007, pp. 1–14.

[2] G. Bossert, F. Guihery, and G. Hiet, “Towards automated protocol ´reverse engineering using semantic information,” in Proceedings of the 9th ACM symposium on Information, computer and communications security, 2014, pp. 51–62.

[3] X. Luo, D. Chen, Y. Wang, and P. Xie, “A type-aware approach to message clustering for protocol reverse engineering,” Sensors, vol. 19, no. 3, p. 716, 2019.

[4] S. Kleber, H. Kopp, and F. Kargl, “{NEMESYS}: Network message syntax reverse engineering by analysis of the intrinsic structure of individual messages,” in 12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18), 2018.