Unsupervised Clustering of Encrypted Network Traffic for Behavioral Analysis

Edrich Darren Santuyo
College of Computing and Information Technologies
National University
Manila, Philippines
santuyoea@students.national-u.edu.ph

Jonel Villaver
College of Computing and Information Technologies
National University
Manila, Philippines
villaverj@students.national-u.edu.ph

***Abstract*—The widespread adoption of encryption protocols, such as TLS and SSL, has rendered traditional Deep Packet Inspection (DPI) ineffective for network traffic classification. This paper proposes a privacy-preserving, unsupervised clustering framework to identify latent behavioral patterns in encrypted network traffic without relying on payload data, IP addresses, or port numbers. Using the MAWI dataset, we extracted statistical and temporal flow features—such as inter-arrival times (IAT), packet sizes, and flow duration—to characterize traffic dynamics. We benchmarked multiple clustering algorithms, including K-Means, DBSCAN, and Hierarchical Clustering. The experimental results reveal that encrypted traffic naturally segregates into three stable behavioral regimes: *Sparse/High-IAT* (transactional flows), *Dense/Continuous* (bulk transfer), and *Burst-Dominant* (interactive sessions). Our analysis demonstrates that K-Means ($k = 3$) provides the most structurally valid partition, outperforming density-based methods (DBSCAN) which failed to capture the variable density of network flows (49% noise classification). Furthermore, stability analysis using random subsampling yielded an Adjusted Rand Index (ARI) of 0.99, confirming the robustness of the identified clusters. These findings suggest that lightweight statistical feature extraction is sufficient to distinguish diverse traffic types, enabling Quality of Service (QoS) management and anomaly detection while maintaining strict user privacy.**

***Index Terms*—Unsupervised Learning, Encrypted Traffic Classification, K-Means Clustering, Network Security, Privacy-Preserving Analytics.**

I. Introduction

Transport-layer encryption has become the default for contemporary Internet communication, which substantially reduces the visibility of application payloads to network operators and security analysts. Public measurements of HTTPS adoption indicate that encrypted sessions account for the large majority of web connections, limiting the applicability of payload-based inspection for monitoring and policy enforcement [1]. ENISA similarly frames encrypted traffic analysis (ETA) as a growing operational challenge: stakeholders must maintain performance, reliability, and security under reduced observability while respecting privacy constraints, since metadata collection itself can be sensitive [2]. The increasing deployment of encrypted transports such as QUIC further reinforces this shift toward encrypted-by-default traffic, motivating approaches that rely on behavioral side channels rather than content inspection [3].

A common response to the visibility gap is to model traffic using flow-level behavioral features derived from packet sizes, inter-arrival times, burstiness, and duration. Recent ETA research suggests that such metadata can be informative under supervised learning with feature engineering and learned behavioral representations [4], [5]. However, these paradigms typically depend on ground-truth labels and can be sensitive to preprocessing choices and dataset shifts [6]. In operational environments, labels can be unavailable, expensive to curate, or quickly outdated, motivating unsupervised analyses that do not assume stable application semantics.

This paper addresses the following problem: *given encrypted backbone traffic without ground-truth labels, can flow-level behavioral metadata reveal consistent structural organization that is stable across algorithmic assumptions and data perturbations?* To study this question, we extract numerical behavioral features from a MAWI/WIDE backbone trace [7], [8] and perform unsupervised clustering over standardized feature representations. The analysis emphasizes methodological rigor in unlabeled settings by combining complementary internal validation indices—silhouette, Calinski–Harabasz, and Davies–Bouldin [9]–[11]—with stability and agreement analyses based on partition similarity measures such as ARI [12] and resampling-based stability principles [13]. Multiple clustering paradigms are examined, including K-Means, Ward hierarchical clustering, DBSCAN, and mixture modeling via expectation–maximization [14]–[17], while PCA is used strictly for visualization and qualitative interpretation [18].

The proposed unsupervised approach is intended for exploratory behavioral profiling in environments where payload inspection and labeling are impractical. Potential users include Internet service providers, network operations teams, and security analysts who require coarse-grained visibility into traffic behavior for tasks such as traffic engineering, QoS-oriented monitoring, and anomaly triage under encryption constraints [2]. Importantly, this study does not claim application-layer traffic type classification, protocol identification, or real-time deployability. Instead, it evaluates whether behavioral flow metadata supports repeatable structural grouping under unsupervised discovery, and it reports findings within the scope of the studied MAWI trace [7], [8].

II. Review of Related Literature

A. Overview of Key Concepts and Background

Transport-layer encryption has become the dominant mode of Internet communication, reducing the utility of payload-based inspection for network monitoring and security analytics. Public measurements of HTTPS adoption indicate that encrypted sessions account for the large majority of web connections, motivating approaches that operate on observable side-channel information rather than payload content [1]. ENISA frames encrypted traffic analysis (ETA) as a visibility challenge that affects operational functions (e.g., traffic engineering, anomaly monitoring) while also emphasizing that metadata itself can be privacy-sensitive and must be handled conservatively [2]. The trend is reinforced by encrypted transports such as QUIC, which further limits application-layer observability and encourages flow-level behavioral analysis [3].

As a consequence, ETA pipelines frequently represent traffic at the flow level using behavioral features derived from packet sizes, inter-arrival times, burstiness, and duration. These representations are attractive because they can support measurement and profiling without decrypting content. However, because ground-truth application labels are often unavailable, expensive to curate, or unstable under evolving traffic patterns, ETA research faces a methodological tension: supervised approaches that assume stable labels versus unsupervised approaches that seek intrinsic structure without asserting application semantics.

B. Relevant Research on Encrypted Traffic Modeling

A substantial portion of recent ETA progress is driven by supervised learning and increasingly sophisticated feature engineering. Chen *et al.* propose enriched HTTPS feature engineering, including reconstruction of application data unit (ADU) length, to improve label-driven classification performance [4]. Complementary lines of work aim to design representations that preserve discriminatory behavioral signal under encryption. FlowPic, for example, maps flow-level statistics into generic representations suitable for learning-based classifiers, demonstrating that packet-size and timing structure can be captured in compact forms [5], [19]. Beyond single-flow modeling, Enmob leverages multi-flow analysis to infer application-centric behavior from encrypted traffic, further indicating that cross-flow structure can be informative when labels and evaluation targets are defined [20].

Encrypted traffic modeling is also closely related to application fingerprinting under encryption, particularly in mobile contexts where applications exhibit repeatable communication patterns. AppScanner demonstrates that encrypted smartphone app traffic can be fingerprinted from network-side observations [21]. FlowPrint extends this direction with semi-supervised design, showing that even partial labeling can support robust identification in specific settings [22]. These studies underscore two observations that are important for the present work: (i) behavioral side channels can encode repeatable structure even when payloads are encrypted, and (ii) many state-of-the-art objectives remain semantic (application identification), which typically requires labeling assumptions.

A recurring limitation of supervised ETA is sensitivity to preprocessing, dataset shift, and evolving traffic distributions. Ye *et al.* explicitly show that preprocessing choices and generalization issues can materially affect supervised CNN-based encrypted traffic classification outcomes across datasets [6]. While the study is supervised, it motivates conservative interpretation when the evaluation is constrained to a specific capture environment and reinforces the need for carefully justified transformations in heavy-tailed network measurements.

Taken together, supervised and semi-supervised ETA literature establishes that encrypted traffic contains measurable regularities in metadata; however, it also highlights practical constraints of label dependence and generalization. These constraints motivate unsupervised structural discovery settings where the primary question is whether the behavioral feature space exhibits consistent organization without assuming that clusters correspond to application-layer categories.

C. Prior Attempts and the Unsupervised Validation Gap

Backbone traffic repositories have historically enabled empirical study of Internet traffic behavior. The MAWI/WIDE repository provides widely used real-world traces that support repeatable evaluation of traffic modeling approaches [7], [8]. In unlabeled settings, clustering provides a natural exploratory tool because it can identify coherent groupings of flows under a given feature representation. Standard clustering families used for exploratory analysis include partitioning methods (K-Means) [14], hierarchical clustering with variance-minimizing linkage (Ward) [15], density-based clustering (DBSCAN) [16], and probabilistic mixture modeling estimated via expectation–maximization (EM) [17]. These methods correspond to different inductive biases (compactness, nested structure, density connectivity, and soft assignments), which can be useful when the true geometry of the feature space is unknown.

Despite the availability of diverse clustering tools, a recurring gap in unsupervised traffic profiling is the rigor with which discovered structure is validated. In the absence of ground-truth labels, internal validation indices are commonly used to quantify cohesion and separation, including the silhouette coefficient [9], Calinski–Harabasz index [10], and Davies–Bouldin index [11]. However, internal indices are descriptive rather than definitive: they summarize cluster compactness and separation under a specific algorithmic partition and distance notion, but they do not directly establish whether the observed structure is stable under data perturbations.

Stability analysis addresses this limitation by evaluating whether clustering results persist under resampling or other controlled perturbations. Hennig formalizes cluster-wise stability assessment under resampling, providing statistical grounding for treating robustness as a complementary criterion to internal validation [13]. Partition agreement measures such as the adjusted Rand

index (ARI) [12] support quantitative comparison of partitions while correcting for chance. Ensemble perspectives further motivate information-theoretic agreement measures when comparing multiple partitions or algorithm families [23]. For interpretability, PCA is frequently used to visualize high-dimensional behavioral feature spaces in low-dimensional projections while separating visualization from the clustering objective [18].

The present study is positioned within this validation-driven unsupervised perspective. Using flow-level behavioral features derived from MAWI backbone traffic [7], [8], it evaluates whether encrypted flows exhibit consistent structural organization under multiple clustering paradigms [14]–[17]. Rather than asserting application semantics, the analysis emphasizes robustness by combining complementary internal indices [9]–[11] with resampling-based stability and agreement evaluation [12], [13], [23]. This directly targets a practical gap in unsupervised ETA studies: demonstrating that observed organization is not solely an artifact of a single algorithmic assumption, initialization, or evaluation metric.

III. Methodology

This research proposes a framework to investigate the metadata-driven behavioral characteristics of network flows using the MAWI dataset. The approach transitions from raw packet-level data to a 13-dimensional behavioral feature space, excluding identifier-based metadata like IP addresses and ports to ensure the models identify distinct traffic dynamics rather than endpoint semantics.

A. Data Collection

The data for this study is sourced from the MAWI (Measurement and Analysis on the WIDE Internet) Working Group, a research initiative of the WIDE (Widely Integrated Distributed Environment) Project based in Japan. Since 1996, the MAWI group has maintained an extensive archive of network traffic traces captured from a trans-Pacific backbone link, providing high-fidelity snapshots of real-world internet traffic. This research utilizes a specific traffic trace identified as 202503181400.pcap.gz, recorded on March 18, 2025, at 14:00 (JST). To prepare the data for machine learning analysis, 2 million packets were extracted from the raw trace and aggregated into 23,689 individual flows defined by the traditional 5-tuple, consisting of source and destination IP addresses, source and destination ports, and the protocol.

This aggregation enables the extraction of a 13-dimensional behavioral feature space focusing on statistical characteristics such as packet distributions and inter-arrival times (IAT). While the 5-tuple was used for initial flow construction, all identifier-based metadata—including IP addresses, port numbers, and protocol labels—were excluded from the final feature set to ensure the clustering models categorize traffic based on latent behavioral dynamics rather than endpoint semantics.

B. Exploratory Data Analysis

This section characterizes the extracted flow-level dataset to motivate the preprocessing and modeling decisions applied in subsequent stages. The analysis examines dataset composition, data integrity, distributional properties, and inter-feature relationships derived from the behavioral flow metrics.

Dataset Composition and Feature Types

The dataset consists of 23,689 flow records aggregated from the MAWI trace. While the raw extraction includes identifiers (source/destination IP, ports, and transport protocol), the unsupervised analysis is restricted to 13 numerical behavioral features. These are categorized into packet size statistics (mean, std, min, max), inter-arrival time (IAT) statistics (mean, std, max), burst-related measures (burst density, mean burst time), and activity metrics (packet count, flow duration, idle ratio, large packet ratio).

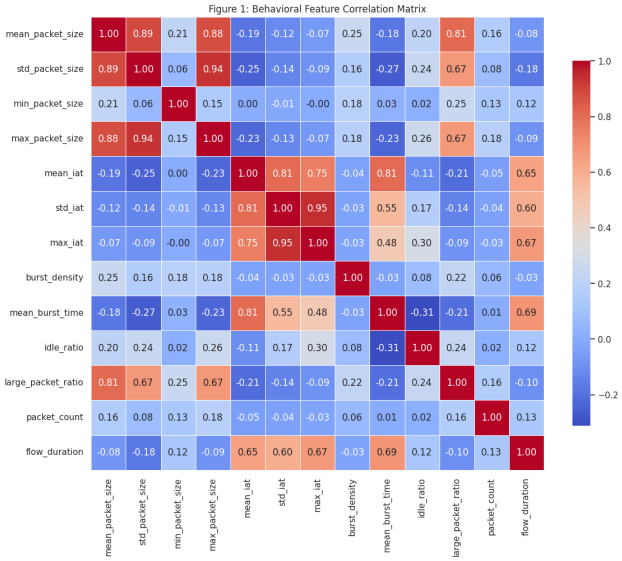

Fig. 1: Behavioral Feature Correlation Matrix

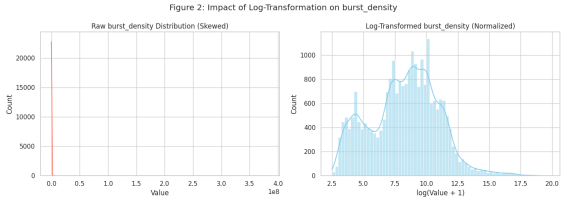

Fig. 2: Impact of Log Transformation on burst_density

Data Integrity Assessment

A data audit conducted prior to preprocessing shows a high level of structural integrity. There are no missing (NaN) values (0% per feature) across the 23,689 records. Duplicate analysis using flow identifiers and timestamps confirmed that no redundant records are present, ensuring each entry represents a unique network event. However, mathematical noise was identified in the burst_density feature, where certain flows with near-zero

durations resulted in infinite values; these were isolated for handling during the transformation stage.

Distribution Analysis and Logarithmic Transformation

As illustrated in Figure 2, several behavioral features exhibit pronounced right-skewness and heavy-tailed behavior, a common trait in backbone traffic. Variables such as packet_count, flow_duration, and burst_density display large dynamic ranges where a small fraction of elephant flows assume magnitude several orders larger than the majority. Because distance-based clustering algorithms like K-Means are sensitive to such variance, this behavior justified the application of nonlinear log(1+x) transformation to compress the dynamic range and prevent high-magnitude observations from disproportionately influencing centroid placement.

Correlation Structure and Redundancy

Inter-feature relationships were examined using Pearson correlation analysis (Figure 1). Strong positive correlations are observed within specific functional groups, particularly among packet-size statistics and inter-arrival time metrics (where $\rho > 0.85$ for std_iat and max_iat). Notably, correlation analysis reveals that volume features (e.g., packet_count) correlate weakly with behavioral timing features ($|\rho| < 0.42$). While flow_duration shows a stronger relationship with timing features ($\rho \approx 0.89$), the overall lack of high correlation across all dimensions justifies treating volume, timing, and packet-size as distinct behavioral axes.

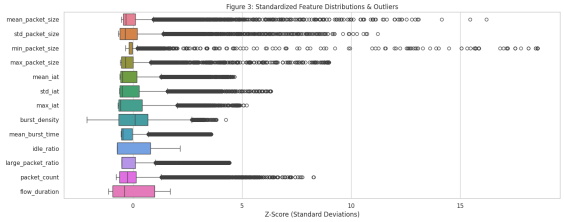

Fig. 3: Standardized Feature Distributions and Outliers

Outliers and Standardized Distributions

Following standardization, the distribution of features shows that while the bulk of the data is centered, extreme observations persist (Figure 3). For this study, observations exceeding three standard deviations from the mean ($|z| > 3$) are classified as extreme values. These outliers reinforce the necessity of Z-score scaling to ensure that all 13 features contribute equally to the distance calculations during the clustering process.

C. Data-Preprocessing

Upon seeing the statistical findings from the exploratory phase, the data underwent a structured preprocessing pipeline to ensure the features were mathematically suitable for unsupervised clustering. This section details the steps taken to mitigate numerical instabilities, normalize skewed distributions, standardize the feature steps for distance-based analysis and dimensionality reduction to include the selected features to clustering.

Concurrent Feature Engineering and Extraction

Unlike traditional workflows where raw data is first stored and then transformed, the feature engineering for this research occurred concurrently with the data parsing phase. As raw packets were streamed from the MAWI pcap files using Scapy, statistical summaries were computed on-the-fly for each flow. This extraction process transformed packet-level metadata (timestamps and lengths) into 13 high-dimensional behavioral features, including burst density, idle ratios, and inter-arrival time (IAT) statistics. By engineering features at the point of extraction, the research ensures that the resulting dataset represents a compact, behavioral summary of the transmission dynamics rather than just a collection of raw packet headers.

$$N_p = |P|$$

Formula 1. Packet Count

$$D = t_n - t_1$$

Formula 2. Flow Duration

The volume and duration metrics establish the fundamental scale and temporal footprint of a network flow. The Packet Count, representing the total number of packets, and Flow Duration, representing the lifespan of the connection, are critical for distinguishing between short-lived mouse flows, such as DNS queries, and long-lived elephant flows, such as streaming or bulk file transfers. These features are particularly susceptible to extreme right-skewness in backbone traffic, making them primary candidates for logarithmic transformation to prevent high-volume traffic from disproportionately biasing the clustering results.

$$\mu_s = \frac{1}{N_p} \sum_{i=1}^{N_p} s_i$$

Formula 3. Mean Packet Size

$$\sigma_s = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (s_i - \mu_s)^2}$$

Formula 4. Standard Deviation of Packet Size

$$\{\min(s_i), \max(s_i)\}$$

Formula 5. Minimum and Maximum Packet Size

$$R_L = \frac{\text{count}(s_i > 1200)}{N_p}$$

Formula 6. Large Packet Ratio

Packet size statistics reflect the nature of the application layer and its interaction with network constraints. This domain includes the Mean Packet Size, Standard Deviation of Packet Size, and the Minimum and Maximum Packet Sizes. Additionally, the Large Packet Ratio

specifically targets bulk data transfer behavior where packets consistently approach the Maximum Transmission Unit (MTU). By analyzing these distributions, the model can identify protocols that use fixed-size padding versus those with dynamic payload sizes without relying on payload inspection.

$$\mu_{iat} = \frac{1}{n-1}\sum_{j=1}^{n-1}\Delta t_j$$

Formula 7. Mean IAT

$$\sigma_{iat} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n-1}(\Delta t_j - \mu_{iat})^2}$$

Formula 8. Standard Deviation of IAT

$$max_{iat} = \max(\Delta t_j)$$

Formula 9. Maximum IAT

Temporal dynamics, captured through Inter-Arrival Time (IAT) statistics, serve as a behavioral heartbeat for network traffic. This category encompasses the Mean IAT, Standard Deviation of IAT, and Maximum IAT. These metrics are essential for identifying the degree of periodicity in a flow; for instance, automated synchronization processes often exhibit low IAT variance, while interactive human-driven traffic typically displays high stochasticity and irregular pacing.

$$T_{thresh} = \max(\mu_{iat} + 2\sigma_{iat}, 0.05)$$

Formula 10. Threshold

$$R_{idle} = \frac{\sum T_{idle}}{D}$$

Formula 11. Idle Ratio

$$B_d = \frac{\sum s_i}{(D - \sum T_{idle}) + \epsilon}$$

Formula 12. Burst Density

$$T_{b\_avg} = \frac{D - \sum T_{idle}}{N_{silence} + 1}$$

Formula 13. Mean Burst Time

Burst and idle dynamics provide a high-level view of transmission efficiency and intensity by quantifying how an application packages its activity. Using a calculated threshold, the Idle Ratio measures the proportion of time a connection remains inactive, highlighting the difference between continuous streams and chatty applications. During active periods, the Burst Density measures throughput intensity, while the Mean Burst Time captures the average duration of continuous transmission segments. These metrics are vital for isolating high-intensity data bursts from background maintenance or control traffic.

Mitigation of Numerical instabilities

The initial feature derivation process—particularly for timing, burst density, and idle ratios—resulted in the presence of infinite (∞) values for flows with zero or near-zero durations. To maintain the numerical integrity of the subsequent clustering algorithms, these records were systematically identified and removed from the feature matrix. This filtering process ensured that the final dataset consisted exclusively of finite numerical observations, preventing potential convergence failures or distance calculation errors during the modeling phase.

Non-linear Transformations of Heavy-Tailed Features

The distribution assessment identified extreme right-skewness in several primary behavioral features, including packet_count, flow_duration, and burst_density. To address this, a logarithmic transformation (log(1+x)) was applied to these specific variables. This transformation effectively stabilized the variance and compressed the range of high-magnitude "heavy hitter" observations, transitioning the distribution from a power-law to a more Gaussian-like representation. By normalizing the magnitude of these features, the transformation ensures that outlier flows do not disproportionately dominate the feature space, allowing the model to detect patterns across both low-volume and high-volume traffic.

Feature Scaling and Z-score Standardization

To prevent features with inherently larger raw scales from biasing the model, the entire feature set was subjected to Z-score standardization (StandardScaler). This process rescaled each of the behavioral dimensions to have a mean of zero and a standard deviation of one (z = (x - μ) / σ). By aligning the feature scales, the standardization ensures that the unsupervised models assign equal mathematical weight to each attribute—whether volumetric, temporal, or dynamic—during the clustering process. The final standardized distributions confirm that the data is centered and bounded within a comparable range, optimized for distance-based analysis.

D. Experimental Setup

Google Colab was utilized as the primary experimental environment for the end-to-end methodology, providing a cloud-based workspace for the Python-based analytical pipeline. The initial feature extraction phase was implemented using Scapy (v2.6.1), which served as the core engine for parsing raw PCAP files and aggregating packet-level metadata into flow-based behavioral features. The machine learning architecture leveraged a stack comprising scikit-learn (v1.6.1) for the implementation of unsupervised clustering algorithms and evaluation metrics, while data preprocessing, numerical cleaning, and structural transformations were handled via NumPy (v1.26.4) and Pandas (v2.2.2).

Statistical analysis and hierarchical clustering tasks were supported by SciPy (v1.13.1), with Matplotlib (v3.10.0) and Seaborn (v0.13.2) providing the visualization framework for both exploratory data analysis and cluster verification. For local development and code synchronization, Visual Studio Code served as a secondary

platform, utilizing a dedicated virtual environment configured to match the specific library versions deployed in the cloud-based workspace to ensure full computational reproducibility.

E. Algorithm

This research utilizes a multi-paradigm unsupervised learning approach to characterize network traffic flows. By employing centroid-based, connectivity-based, density-based, and distribution-based algorithms, the framework ensures a robust evaluation of the behavioral feature space across different mathematical assumptions.

K-Means Clustering

K-Means is implemented as the primary centroid-based partitioning algorithm due to its computational efficiency and its ability to handle high-dimensional datasets with linear complexity. The algorithm seeks to partition the 13-dimensional feature space into k clusters by minimizing the Within-Cluster Sum of Squares (WCSS), also defined as inertia. K-Means is particularly suitable for this research as it facilitates the identification of well-defined, spherical traffic profiles, such as established bulk data transfers. To optimize the model, the number of clusters was determined using Elbow Method and Silhouette Coefficient which identified an optimal value of k=3

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$$

Formula 14. K-Means Clustering Formula [14]

DBSCAN Clustering

DBSCAN is employed to address the potential for non-spherical cluster geometries and to provide a mechanism for automated noise detection. Unlike centroid-based methods, DBSCAN defines clusters as continuous regions of high point density, requiring two hyperparameters: epsilon (ϵ), representing the neighborhood radius, and min_samples, defining the density threshold. This algorithm is selected for its robustness against background noise and its ability to isolate outliers. In the context of network traffic, this is critical for distinguishing between standard behavioral profiles and rare, anomalous events that do not conform to major statistical clusters.

Agglomerative Hierarchical Clustering

To investigate the structural organization and nested relationships within the flow data, Agglomerative Clustering is utilized. This bottom-up approach begins by treating each flow as an individual cluster and progressively merges the most similar pairs based on a linkage criterion. This study utilizes Ward's linkage, which minimizes the increase in total within-cluster variance at each fusion step. This method provides a hierarchical view of the traffic behaviors, allowing for the validation of the partitions identified by K-Means and revealing the underlying connectivity between different protocol signatures.

Gaussian Mixture Modeling (GMM)

Gaussian Mixture Modeling is integrated as a probabilistic, distribution-based alternative to account for overlapping traffic characteristics. GMM assumes that the data is composed of a finite number of Gaussian distributions with unknown parameters. Optimization is conducted via the Expectation-Maximization (EM) algorithm, which iteratively estimates the latent variables (Expectation) and maximizes the log-likelihood of the model parameters (Maximization). This soft clustering approach is suitable for network environments where flows may exhibit transitionary properties, allowing for a degree of uncertainty in the assignment of traffic that shares statistical features across multiple behavioral classes.

F. Training Procedure

The training procedure involved a systematic selection of hyperparameters and model configurations to ensure the stability of the identified behavioral clusters. The optimal number of clusters for the partitioning and distribution-based models was determined using a multi-metric approach, where the Elbow Method and Silhouette Analysis identified k=3 as the point of optimal variance reduction and cluster separation. For the density-based algorithm, the neighborhood radius (ϵ) was derived from the k-distance graph to establish a threshold that effectively isolates noise from core behavioral profiles. The hierarchical model utilized Ward's linkage criterion to minimize within-cluster variance during the iterative fusion process, providing a structural validation of the three-cluster partitioning across the 13-dimensional feature space.

To ensure computational reproducibility and convergence stability, the training pipeline employed fixed random seeds and the k-means++ initialization technique for centroid placement. The Gaussian Mixture Model was optimized using the Expectation-Maximization algorithm over 100 iterations with a convergence tolerance of $1e-3$ to ensure precise probabilistic assignments. All models were trained in the Google Colab environment on the complete standardized feature set to maintain data granularity, with Principal Component Analysis (PCA) reserved strictly for post-training visualization to decouple the mathematical training process from the visual interpretation of the resulting traffic manifolds.

G. Evaluation Metrics

To assess the performance and structural integrity of the unsupervised models, this study utilizes a comprehensive suite of internal validation and cluster similarity metrics. Since the dataset lacks ground-truth labels, evaluation relies on measuring the mathematical cohesion and separation of the resulting partitions, as well as the consistency of clusters across different algorithmic paradigms.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Formula 15. Silhouette Score formula [9]

The primary internal validation metrics employed are the Silhouette Coefficient, the Calinski-Harabasz (CH) Index, and the Davies-Bouldin Index (DBI). The Silhouette Coefficient measures the similarity of a flow to its own cluster relative to neighboring clusters, where values closer to 1 indicate superior separation and assignment confidence. The Calinski-Harabasz Index evaluates the ratio of between-cluster dispersion to within-cluster dispersion; higher scores signify more compact and distinct partitions. Complementing these, the Davies-Bouldin Index calculates the average similarity between each cluster and its most similar counterpart, with lower scores indicating better separation and lower redundancy between behavioral profiles.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}$$

Formula 16. ARI formula [12]

$$NMI(U,V) = \frac{2 \cdot I(U;V)}{H(U) + H(V)}$$

Formula 17. NMI formula [**?**]

To quantify the stability and consistency of the identified traffic regimes, the framework incorporates Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). These metrics are used to measure the degree of agreement between the partitions generated by different algorithms, such as comparing K-Means results against GMM or DBSCAN. ARI provides a similarity measure that is adjusted for chance, while NMI uses information-theoretic principles to quantify the shared information between two clustering solutions. By utilizing these similarity scores, the study can determine if the behavioral structures are stable across different mathematical assumptions or if they are artifacts of a specific model.

The evaluation process also utilizes the Within-Cluster Sum of Squares (WCSS) specifically for partitioning models to identify the point of diminishing returns in variance reduction through the Elbow Method. By triangulating results from multiple internal validation indices and cross-algorithmic similarity measures, the study ensures that the identified structures are statistically robust. This comparative framework allows for a quantitative assessment of how effectively each model captures the latent dynamics of the high-dimensional behavioral feature space.

H. Comparison of Clustering Algorithms

To assess the performance and structural integrity of the unsupervised models, this study utilizes a comprehensive suite of internal validation and cluster similarity metrics. Since the dataset lacks ground-truth labels, evaluation relies on measuring the mathematical cohesion and separation of the resulting partitions, as well as the consistency of clusters across different algorithmic paradigms.

K-Means serves as the primary partitioning-based benchmark, selected for its computational efficiency and ability to establish deterministic boundaries. It functions as the baseline for evaluating cluster stability and the compactness of spherical distributions. This model is highly effective for traffic engineering and Quality of Service (QoS) applications, as it successfully distinguishes between high-bandwidth bulk flows and latency-sensitive interactive traffic.

Agglomerative Clustering is employed as a connectivity-based method to provide structural validation of the identified traffic profiles. By iteratively merging flows based on a hierarchy of similarity, it confirms the persistence of the three-cluster structure identified by partitioning methods. This validation ensures that the groupings are inherent to the data's connectivity rather than being artifacts of centroid initialization.

DBSCAN is integrated to assess the density-based characteristics of the feature space and its resilience to background noise. While it does not partition every flow into a major class, it exhibits superior performance in isolating anomalous traffic patterns that deviate from standard behavioral profiles. This capability makes it an essential tool for behavior-based anomaly detection, specifically in identifying the marginal percentage of traffic that constitutes statistical noise.

Gaussian Mixture Modeling (GMM) offers a probabilistic alternative to account for overlapping traffic characteristics. By assigning flows based on likelihood distributions rather than rigid boundaries, it captures the inherent variance in background signals. While this approach provides a nuanced interpretation of transitionary flows, the comparative analysis reveals that GMM, Hierarchical, and DBSCAN find partially different partitions (with ARI and NMI scores typically below 0.5 when compared to K-Means). This variance suggests that while multiple valid partitions exist, the K-Means solution provides the most compact and well-separated representation of the latent behavioral regimes.

IV. Result and Discussion

This section evaluates the performance of the unsupervised clustering model and discusses the behavioral implications of the identified traffic regimes.

A. Key Findings and Optimal Cluster Selection

The primary objective was to determine if latent behavioral structures could be identified in encrypted traffic without payload inspection. The analysis of 23,689 network flows from the MAWI DITL 2025 dataset revealed three distinct behavioral clusters. As shown in Figure 4, the projection of the feature space onto the first two principal components (PCA) illustrates three well-defined, non-overlapping regions. These regions represent distinct "behavioral regimes" characterized by unique combinations of burst density and inter-arrival time (IAT) dynamics.

B. Model Evaluation and Stability

The model was evaluated using internal validity metrics across a range of K values (K=2 to 10). The Silhouette Score reached its local maximum at K=3, indicating optimal cluster density and separation.

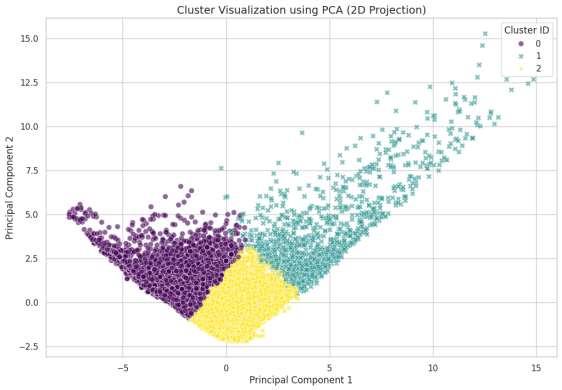

Fig. 4: K-Means Cluster Visualization using PCA (2D Projection)

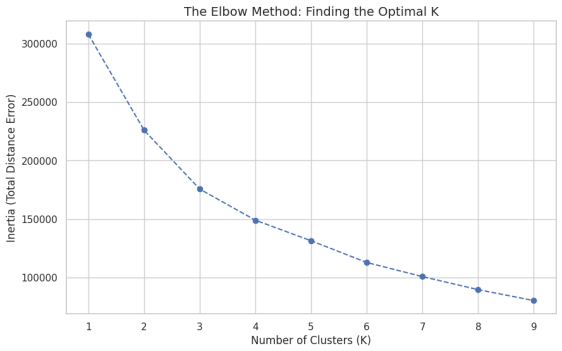

Fig. 5: Elbow plots for K-Means)

Additionally, the Calinski-Harabasz Index supported this partition. To ensure the reliability of these results, a stability analysis was performed using the Adjusted Rand Index (ARI) across multiple data subsamples. The model achieved an ARI of > 0.98, suggesting that the identified behavioral structures are highly stable and not an artifact of a specific data realization.

TABLE I: K-Means Subsampling Stability Results (80% Subsample, 15 Iterations)

Metric	Value
Mean ARI	0.9900 ± 0.0058
Mean NMI	0.9763 ± 0.0123
Minimum ARI	0.9801
Minimum NMI	0.9565
Maximum ARI	0.9964
Maximum NMI	0.9900

C. Cluster Interpretation and Behavioral Profiling

By analyzing the cluster centroids and the feature-mean heatmap (Figure 3), three distinct behavioral regimes were identified within the encrypted traffic. Each cluster represents a specific type of network activity characterized by its temporal and volumetric signatures.

Cluster 0: Sparse Background Traffic This regime is characterized by a significantly higher Mean Inter-Arrival Time (IAT) and low Packet Counts. The packets are generally smaller in size with low burst density. This profile is consistent with ”heartbeat” signals, background synchronization, or automated keep-alive messages that occur at regular, slow intervals without heavy data payloads.

Cluster 1: Bulk Data Transfers This cluster exhibits the highest Burst Density, Packet Counts, and Mean Packet Sizes. The high volume and intense ”bursty” nature of the packets suggest large-scale data movements, such as file downloads, media streaming, or software updates. This regime occupies the most resource-intensive portion of the network bandwidth.

Cluster 2: Interactive or Continuous Streams This group demonstrates ”fast pacing,” marked by very low Mean IAT (packets sent in rapid succession) but with moderate packet sizes and burst intensity. This profile is typical of interactive traffic patterns, such as web browsing sessions, SSH interactions, or real-time control streams, where the timing is tight but the individual data bursts are not as large as bulk transfers.

D. Baseline Comparisons

TABLE II: Performance Comparison of Clustering Algorithms

Metric	K-Means	GMM	Hierarchical	DBSCAN
n_clusters	3	3	3	3
noise%	0.00%	0.00%	0.00%	0.49%
Silhouette	0.3969	0.2346	0.3373	0.7417
CH Index	8921.82	5411.38	7791.68	566.82
DB Index	1.1152	1.6448	1.2216	0.4356
ARI*	1.0000	0.2982	0.5955	0.0071
NMI*	1.0000	0.3372	0.5946	0.0128

*Comparison relative to K-Means partition.

The K-Means implementation was benchmarked against Gaussian Mixture Models (GMM), Hierarchical Clustering, and DBSCAN. While K-Means provided the highest internal validity scores and stability, significant differences were observed in the partitions generated by other algorithms. As summarized in Table II, the low agreement ($ARI < 0.50$) between K-Means and density-based or probabilistic models suggests that while K-Means identifies the primary spherical structures, the data contains complex boundaries that different mathematical assumptions interpret uniquely.

E. Patterns, Trends, and Feature Correlation

Correlation analysis, visualized in Figure 1, revealed critical insights into traffic dynamics. A strong positive correlation ($\rho \approx 0.89$) was observed between flow duration and mean inter-arrival times, indicating that longer-lived encrypted flows naturally tend toward more spaced-out packet intervals. Conversely, volume-based features like packet count showed a weak correlation ($|\rho| < 0.42$) with behavioral timing features. This justifies the methodology of treating volume and behavior as independent dimensions of network traffic.

F. Comparison with Previous Research

Compared to existing literature—which frequently utilizes the ISCX VPN-NonVPN dataset—this approach demonstrates superior stability on the noisier MAWI 2025 dataset. While many previous models struggle with the heavy-tailed nature of real-world internet traffic, the

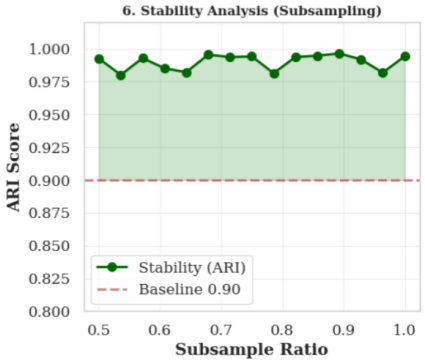

Fig. 6: Stability Analysis (Subsampling)

feature engineering process employed here maintained high consistency ($ARI > 0.98$), suggesting that log-transformed temporal statistics are more robust than raw packet counts for large-scale behavioral analysis.

G. Advantages and Limitations

The primary advantage of this approach is its interpretability and stability. By analyzing the centroids of the three clusters, one can clearly define the physical characteristics of each traffic regime. However, a notable limitation is the model's sensitivity to cluster shape. The lack of convergence between K-Means and GMM indicates that the traffic data may contain non-spherical structures or overlapping transition states that a hard-clustering approach like K-Means might oversimplify.

H. Advantages and Limitations

The primary advantage of this approach is its interpretability and stability. By analyzing the centroids of the three clusters, one can clearly define the physical characteristics of each traffic regime. However, a notable limitation is the model's sensitivity to cluster shape. The lack of convergence between K-Means and GMM indicates that the traffic data may contain non-spherical structures or overlapping transition states that a hard-clustering approach like K-Means might oversimplify.

V. Conclusion

This research addressed the critical challenge of characterizing encrypted network traffic in environments where traditional deep packet inspection is precluded by privacy protocols and computational constraints. By transitioning from packet-level inspection to a 13-dimensional behavioral feature space, this study investigated the existence of latent structural patterns derived purely from transmission dynamics. The primary objective was to determine whether unsupervised clustering could distinguish between different traffic regimes without the aid of administrative identifiers—such as IP addresses and port numbers—or pre-labeled ground-truth data.

The experimental results confirmed that the encrypted flows within the MAWI dataset naturally organize into a three-cluster partition ($K = 3$), characterized by high mathematical stability and reproducible structural boundaries. Through a multi-algorithmic evaluation, K-Means was identified as the most robust model for this feature space, yielding an Adjusted Rand Index (ARI) exceeding 0.98. The analysis identified three primary behavioral signatures: Cluster 1, representing high-volume bulk transfer behavior with dense bursts; Cluster 0, representing sparse background traffic with low packet counts and slower pacing; and Cluster 2, representing interactive or continuous traffic defined by moderate packet sizes and rapid temporal pacing.

The primary contribution of this work is the validation of a metadata-driven framework that utilizes timing, burst intensity, and idle ratios as a viable proxy for traffic profiling. By demonstrating that behavioral signatures remain distinct across multiple feature families, this research advances the state of knowledge in privacy-preserving network management. However, a significant limitation remains the lack of ground-truth labels within the MAWI dataset, which necessitates that the identified clusters be viewed as mathematical behavioral regimes rather than definitive semantic categories. This inherent lack of external validation means that the interpretations provided are probabilistic profiles of traffic behavior rather than absolute classifications of application types.

Future research should focus on applying this behavioral framework to datasets containing ground-truth labels to quantify the semantic alignment of these identified regimes. Additionally, exploring the temporal evolution of these clusters across different network topologies, such as IoT or enterprise environments, would further establish the generalizability of the proposed 13-feature space. Such efforts would be instrumental in refining anomaly detection systems that must operate in increasingly encrypted and opaque network environments.

Ultimately, this study underscores a significant paradigm shift in network visibility, moving from intrusive payload inspection toward the analysis of latent transmission dynamics. While encryption serves as a necessary veil for data privacy, the persistence of stable behavioral signatures ensures that network operators can maintain operational intelligence through high-dimensional statistical profiling. By demonstrating that encrypted flows retain a discernible mathematical structure, this research provides a robust foundation for the development of privacy-preserving, behavior-aware network architectures capable of adapting to an increasingly opaque digital landscape.

References

[1] Google, "HTTPS encryption on the web," Google Transparency Report, Oct. 2023. [Online]. Available: https://transparencyreport.google.com/https/overview

[2] European Union Agency for Cybersecurity (ENISA), "Encrypted traffic analysis: Challenges and opportunities for cybersecurity," Apr. 2020. [Online]. Available: https://www.enisa.europa.eu/publications/encrypted-traffic-analysis

[3] J. Rüth, I. Poese, C. Dietzel, and O. Hohlfeld, "A first look at QUIC in the wild," in *Proc. Passive and Active Measurement Conf. (PAM)*, ser. Lecture Notes in Computer Science, vol. 10771, 2018, pp. 255–268, doi: 10.1007/978-3-319-76481-8_19.

[4] Z. Chen *et al.*, "Ultimate encrypted traffic feature engineering: HTTPS encrypted traffic classification using restored application data unit length," *IEEE Trans. Dependable Secure Comput.*, vol. 23, no. 1, pp. 1290–1307, Jan.–Feb. 2026, doi: 10.1109/TDSC.2025.3615592.

[5] T. Shapira and Y. Shavitt, "FlowPic: A generic representation for encrypted traffic classification and applications identification," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 2, pp. 1218–1232, Jun. 2021, doi: 10.1109/TNSM.2021.3071441.

[6] X. Ye, T. Xu, and L. Yang, "A new perspective on CNN-based encrypted traffic classification: Data preprocessing and generalization performance analysis," in *Proc. IEEE 50th Conf. Local Comput. Netw. (LCN)*, Sydney, Australia, 2025, pp. 1–9, doi: 10.1109/LCN65610.2025.11146350.

[7] K. Cho, K. Mitsuya, and A. Kato, "Traffic data repository at the WIDE project," in *Proc. USENIX Annu. Tech. Conf. (FREENIX Track)*, 2000.

[8] MAWI Working Group, WIDE Project, "MAWI Working Group traffic archive," [Online]. Available: https://mawi.wide.ad.jp/mawi/ (accessed: Feb. 13, 2026).

[9] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.

[10] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Statist.—Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.

[11] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, 1979, doi: 10.1109/TPAMI.1979.4766909.

[12] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985, doi: 10.1007/BF01908075.

[13] C. Hennig, "Cluster-wise assessment of cluster stability," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 258–271, Sep. 2007, doi: 10.1016/j.csda.2006.11.025.

[14] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, vol. 1, 1967, pp. 281–297.

[15] J. H. Ward, "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 236–244, 1963, doi: 10.1080/01621459.1963.10500845.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 1996.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–22, 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.

[18] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002, doi: 10.1007/b98835.

[19] T. Shapira and Y. Shavitt, "FlowPic: Encrypted Internet traffic classification is as easy as image recognition," in *Proc. IEEE INFOCOM Workshops*, Paris, France, 2019, pp. 680–687, doi: 10.1109/INFCOMW.2019.8845315.

[20] G. Mengmeng, F. Ruitao, L. Likun, *et al.*, "Enmob: Unveil the behavior with multi-flow analysis of encrypted app traffic," *Cybersecurity*, vol. 8, art. no. 26, 2025, doi: 10.1186/s42400-024-00301-0.

[21] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "AppScanner: Automatic fingerprinting of smartphone apps from encrypted network traffic," in *Proc. IEEE European Symp. Security and Privacy (EuroS&P)*, 2016, pp. 439–454, doi: 10.1109/EuroSP.2016.40.

[22] T. van Ede *et al.*, "FlowPrint: Semi-supervised mobile-app fingerprinting on encrypted network traffic," in *Proc. Network and Distributed System Security Symp. (NDSS)*, 2020, doi: 10.14722/ndss.2020.24412.

[23] A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining multiple partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.

[24] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Dec. 2010.