

SNAP Framework: Linked Prediction Based Anomaly Prevention With Suspicious Nodes on Social Network Graph

Vahide Nida Kılıç

(Adana Alparslan Türkeş Science and Technology University, Adana, Turkey
<https://orcid.org/0000-0003-2181-9309>, vnuzel@atu.edu.tr)

Esra Saraç Eşsiz

(Adana Alparslan Türkeş Science and Technology University, Adana, Turkey
<https://orcid.org/0000-0002-2503-0084>, esarac@atu.edu.tr)

Abstract: In previous studies, the focus has predominantly been on anomaly detection, with minimal attention given to anomaly prevention. However, anomaly prevention holds greater significance than anomaly detection. Preventing anomalous behavior before it occurs and identifying potential anomalies in advance to enable timely intervention is both challenging and crucial. In this study, a Suspicious Nodes Anomaly Prevention framework for anomaly prevention has been developed. First, a novel K-medoid based Salp Swarm Anomaly Detection method is proposed within the framework. This method reveals unclustered data by applying clustering and determines the boundaries of clusters using a nature-inspired algorithm that optimizes the threshold. Since threshold determination is an optimization problem, it aligns well with nature-inspired algorithms. Additionally, the Enron email dataset was selected as it is a real-world dataset with accessible content information. Initially, content and node features were extracted from the Enron email dataset. The proposed anomaly detection method was then applied separately to each of these features. Nodes identified as anomalous by one feature but normal by others were of particular interest. These nodes were labeled as “suspicious nodes,” and their connections were analyzed to detect potentially harmful email content. This framework fills a significant gap in the anomaly detection literature by contributing an unprecedented approach to anomaly prevention, offering early intervention capabilities in various sectors by identifying risks in advance. In this study, the proposed framework demonstrates high efficacy in detecting anomalies, achieving a True Positive Rate of 94% in node-based anomaly detection and 78% in content-based anomaly detection, indicating a robust capability for early intervention and risk identification.

Keywords: Anomaly Prevention, Linked Prediction, Social Network Graph, Nature Inspired Algorithms, Enron Dataset

Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1

DOI: 10.3897/jucs.152114

1 Introduction

Data that exhibit behavior other than normal behavior or data that has a very different distribution compared to other data are defined as an anomaly. Anomaly data can exist in various fields such as fraud, spam, event detection, and medical diagnosis. For this reason, anomaly detection is a very important issue. Thanks to anomaly detection, zero-day attacks, which are very dangerous for security, can be detected or a new drug can be

discovered. It can also be applied to text data or graph data. Anomaly detection in graph data is more applicable because it is visualizable.

Social networks as a graph structure have become a part of our real lives over the past decade. With the explosive growth and development of social platforms the power of reaching someone in the real world has been dramatically enhanced. Almost everyone has access to social network. Among these people, many people are malicious and harm others. These people live inside us like normal people. In fact, it is very important to detect these abnormal people in the social network. It is necessary to take precautions for the actions to be taken rather than just being detected. This is only possible before the action takes place. Such people should be identified in advance and psychological treatment should be provided if necessary.

Although Anomaly detection has been studied for years, it is still an unresolved issue. This is because anomalies always change shape. If these anomalies are human, they have discovered new methods to hide. For example, some methods treat star graphs as an anomaly. If the person with an anomaly fits the definition of a star graph, they can use various simple methods to change it. In this way, it easily bypasses the anomaly detection method. Anomaly detection will continue to work as long as these changes exist. Anomaly detection methods that are not dependent on change should be developed.

As mentioned before, there are many studies in anomaly detection and studies are still ongoing. Since anomaly prevention is a more difficult problem than detection, in anomaly prevention, a branch of anomaly detection, there is very little work compared to anomaly detection. The biggest difference between anomaly prevention from detection is that it can be intervened, and precautions taken before the event occurs. There are very critical anomalies that need to be detected on time, where it is useless to detect them after the fact. That's why anomaly prevention is so important.

While previous studies have generally focused on anomaly detection, this study is centered on anomaly prevention. In this work, a new clustering and threshold-based anomaly detection method is proposed. This approach aims to first perform anomaly detection more accurately. Subsequently, for the anomaly prevention phase, anomaly detection is carried out using node-based and content-based methods, each tailored to the nature of the data. The key distinction of this study from others is that it does not merely detect anomalies and label them as such. Instead, suspicious nodes are identified, and it is emphasized that expert intervention is necessary to assess whether these nodes may pose potential risks and require preventive measures.

This study investigates anomaly prevention in the context of email communication. Given the irreversible nature of email data, where once an email is sent it cannot be retrieved, and considering that traditional filtering methods frequently fail to identify numerous anomalies, the email dataset is employed for analysis.

Enron corpus, which has been used in many studies, is used as a dataset in this study, to compare with more studies and to highlight the success of our work. Some reasons for choosing the Enron corpus can be listed as follows:

1. Sharing the content of e-mails in a public way,
2. There is information about who sent an e-mail to whom, day and time,
3. It is a real dataset.

The main contributions of this paper are summarized as follows:

1. Unlike other studies, this paper focuses on anomaly prevention rather than anomaly detection.

2. It extracts both node-based and content-based features, identifies suspicious nodes, and prevents their potential connections through link prediction. This approach provides a comprehensive framework.
3. The paper employs nature inspired algorithms, to find optimum threshold value, in a novel manner for the purpose of anomaly prevention.

The structure of this paper is as follows: Section 2 presents a literature review covering Link Prediction, Anomaly Detection, and the Enron email dataset. Section 3 provides the background of the study. Section 4 outlines the system overview, while Section 5 details the experimental results. Finally, Section 6 discusses the conclusions and potential directions for future research.

2 Literature Review

Traditionally, much of the literature has focused on anomaly detection, which seeks to identify unusual or unexpected behaviors within systems to uncover vulnerabilities and potential threats [Rahman et al. 2021, Degirmenci and Karal 2022, Xu et al. 2022, Xu et al. 2023, Min et al. 2024]. These approaches are predominantly reactive, taking action only after the threat has emerged. However, moving beyond anomaly detection to anomaly prevention is crucial, as it shifts the focus towards preventing potential threats before they occur. Few studies have focused on anomaly prevention, making it an underexplored area despite its significance in preemptively mitigating risks [?, Van Vlasselaer et al. 2015, Kirchner and Gade 2011]. In this study, we aim to bridge this gap by proposing a proactive approach that emphasizes anomaly prevention over post-event intervention. Such a strategy not only enhances system security but also reduces the risk of costly security breaches.

To extend beyond traditional anomaly detection methods, our study employs linked prediction techniques, which are rarely used for anomaly prevention. In the existing literature, linked prediction has typically been applied in social network analysis, where it is used to predict the future relationships and connections between individuals [Samad et al. 2021, Ott. et al. 2021, Saxena et al. 2022, ?, Lande et al. 2020, Zhou 2023, Jiao et al. 2024]. These studies aim to model the evolution of relationships over time and forecast future interactions. However, the application of linked prediction methods for anomaly prevention remains largely unexplored. This gap represents one of the innovative aspects of our study. By analyzing relationships within data through linked prediction, we contribute to enhancing the process of anomaly prevention, particularly in social network-like datasets.

The Enron email dataset, a widely utilized real-world dataset, has been employed extensively in anomaly detection research. It offers a rich set of data capturing the email communications within the Enron Corporation, making it highly suitable for building robust anomaly detection models [Jáñez-Martino et al. 2023, Bountakas and Xenakis 2023, Poobalan et al. 2025]. The dataset's structure can also be modeled as a social network, enabling the analysis of user connections and interactions. Previous research has demonstrated the efficacy of the Enron dataset in detecting anomalies across various domains, including classification, clustering, and text analysis [Jáñez-Martino et al. 2023, Bountakas and Xenakis 2023, Poobalan et al. 2025, Corneli et al. 2019, Assouli et al. 2021]. However, existing studies predominantly emphasize detection, with limited focus on preventive measures. By leveraging our proactive anomaly prevention

framework in the context of the Enron dataset, this study aims to address this gap and explore the potential of applying preventive strategies to real-world scenarios.

A critical component of anomaly detection is the accurate setting of anomaly thresholds, which help distinguish normal from abnormal behavior. Establishing the correct threshold is crucial, as it minimizes false positive and false negative rates, leading to more effective anomaly detection. Our study leverages nature-inspired algorithms to determine these thresholds. These algorithms have proven successful in complex optimization problems [Bastami et al. 2021, Alsaleh and Binsaeedan 2021, Alzaqebah et al. 2023, Khayyat 2023, Shao et al. 2022]. However, their application in the context of anomaly threshold setting is limited in the literature, adding another novel dimension to our work. By utilizing nature-inspired algorithms, our approach enhances anomaly detection by optimizing the system for lower false positive rates.

While the bulk of the literature centers on anomaly detection, this study introduces a novel framework for anomaly prevention, offering a proactive solution to potential security threats. Our approach, which combines linked prediction and nature-inspired algorithms, goes beyond traditional detection methods and seeks to prevent threats before they occur. By utilizing the Enron email dataset, this study provides a real-world application, offering insights into anomaly prevention that can be applied to a range of domains. Furthermore, the use of nature-inspired algorithms and linked prediction methods represents a rare and innovative solution, offering the potential for more secure and efficient systems in practice.

Feature	Previous Studies	Proposed Study
Techniques Used	Clustering and statistical methods	Clustering and threshold-based nature-inspired algorithms
Linked Prediction	Relationship prediction	Applied for anomaly prevention
Threshold Setting	Fixed threshold values, leading to high false positive rates	Optimized threshold values to reduce false positives
Innovative Aspects	Focus on anomaly detection only	Focus on anomaly prevention, with active intervention through suspicious node identification

Table 1: Comparison with Previous Studies

In Table-1, The comparison presented highlights the key differences between existing anomaly detection studies and the proposed approach, which focuses on anomaly prevention. While the majority of previous research has centered on reactive methods that identify anomalies only after they occur, this study shifts the focus towards proactive prevention. Traditional approaches often rely on clustering and statistical methods to detect unusual patterns; however, they do not take additional steps to mitigate potential threats. In contrast, the proposed framework not only identifies anomalies but also emphasizes preventing them by flagging suspicious nodes and recommending expert intervention. Moreover, the integration of nature-inspired algorithms for threshold optimization enhances detection accuracy by reducing false positives. By incorporating

linked prediction techniques, the study goes a step further, enabling the analysis of data relationships to preemptively address risks. This forward-thinking and novel approach fills a significant gap in the literature, offering a more comprehensive and effective solution for anomaly detection and prevention.

Recent studies on mobile edge networks and dynamic offloading have focused on improving performance under varying network conditions. [Mohajer et al. 2024] introduced FlexSlice, a dynamic offloading framework utilizing a multi-head graph attention mechanism and TD3 algorithm to optimize traffic prediction and network slicing, achieving better service demand handling and quality of service. [Yang and Mohajer 2025] developed a framework combining PD-NOMA and DRL for multi-objective optimization in satellite communication networks, enhancing throughput, energy efficiency, and reliability. [Zhou and Mohajer 2024] explored UAV integration in edge networks, using IRS and cell-free communication strategies for improved backhaul traffic and resource management, further advancing dynamic offloading techniques for optimized network performance.

3 Backgrounds

3.1 K-medoids

K-medoids [Kaufman and Rousseeuw 2009] is a clustering technique that selects data points from the dataset as cluster representatives. Similar to other clustering techniques, its goal is to minimize the dissimilarity of points within the same cluster by utilizing medoids. Unlike K-means, K-medoids uses actual data points as the cluster centers. To calculate the distances between points, metrics such as Manhattan or Euclidean distance are employed. Compared to K-means, K-medoids has a stronger capability to handle outliers.

$$\text{minimize } \sum_{i=1}^n \sum_{j=1}^k d(x_i, m_j) \quad (1)$$

In Equation-1, x_i represent a point within the dataset and m_j represent a medoid, with a total of k clusters. The goal is to calculate and minimize the distances between x_i and the k clusters, where the distance function $d(x_i, m_j)$ is minimized. The point where this minimization occurs is considered the optimal solution.

3.2 Silhouette Score

The Silhouette Score [Rousseeuw 1986] is a metric used to evaluate the effectiveness of clustering by measuring how well-separated clusters are. It assesses the compactness and separation of clusters by comparing the average distance of a sample to other samples within the same cluster $a(i)$ with the average distance to the samples in the nearest neighboring cluster $b(i)$. The Silhouette Score for a sample i is given by:

$$\text{Silhouette Score}(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (2)$$

In this formula, $a(i)$ represents the average distance between the sample i and all other samples in its cluster, while $b(i)$ denotes the minimum average distance between

the sample i and all samples in any other cluster. A higher Silhouette Score indicates that the sample is more appropriately clustered, showing that it is well-separated from other clusters while being closer to its own cluster.

3.3 Anomaly Detection Methods

3.3.1 Interquartile Range

The Interquartile Range (IQR) [Wan et al. 2015] is a statistical measure used to quantify the dispersion of a dataset. It is obtained by subtracting the lower quartile from the upper quartile. Specifically, the IQR represents the range between the 25th percentile (Q_1) and the 75th percentile (Q_3) of the data. In Equation-3, Q_3 denotes the upper quartile, which corresponds to the 75th percentile, while Q_1 represents the lower quartile, or the 25th percentile.

$$\text{IQR} = Q_3 - Q_1 \quad (3)$$

To identify outliers, thresholds are established using the Interquartile Range (IQR). Equations-4 and 5 define the upper and lower limits as 1.5 times the IQR added to and subtracted from the third quartile (Q_3) and first quartile (Q_1), respectively. Points that fall above or below these thresholds are considered outliers.

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR} \quad (4)$$

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR} \quad (5)$$

3.3.2 Local Outlier Factor (LOF)

LOF [Breunig et al. 2000] detects anomalies by evaluating the local density of data points in relation to their neighbors. Unlike global methods, LOF focuses on the local structure of the data. It assesses how isolated a data point is compared to its neighbors, with outliers being those that have significantly lower local density than their neighbors.

The LOF score for a data point p is defined as:

$$\text{LOF}(p) = \frac{\sum_{o \in N_k(p)} \frac{\text{lrd}(o)}{\text{lrd}(p)}}{|N_k(p)|} \quad (6)$$

where $N_k(p)$ denotes the set of k -nearest neighbors of p , and $\text{lrd}(p)$ represents the local reachability density of p . The local reachability density is calculated as:

$$\text{lrd}(p) = \frac{1}{\sum_{o \in N_k(p)} \frac{\text{reach-dist}_k(o,p)}{|N_k(p)|}} \quad (7)$$

where $\text{reach-dist}_k(o,p)$ is the reachability distance between o and p , given by:

$$\text{reach-dist}_k(o,p) = \max(\text{dist}(o,p), \text{dist}(o, N_k(p))) \quad (8)$$

Here, $\text{dist}(o,p)$ denotes the distance between points o and p , and $\text{dist}(o, N_k(p))$ is the distance from o to the k -th nearest neighbor of p .

3.3.3 Isolation Forest (iForest)

iForest [Liu et al. 2008] is an anomaly detection method that isolates data points rather than profiling normal data distributions. It constructs multiple random binary trees by recursively splitting the data on randomly chosen features and split values. Anomalies are isolated more quickly and thus have shorter path lengths in these trees.

The anomaly score for a data point is calculated using the formula:

$$\text{score}(a) = 2^{-\frac{E(h(a))}{c(n)}} \quad (9)$$

In Equation-9, $h(a)$, n , $E(h(a))$, $c(n)$ represents the path length of a data point, the number of external nodes, the average path length, the average path length of an unsuccessful search in a Binary Search Tree (BST), respectively.

3.4 Link Prediction Methods

Link prediction techniques are employed to anticipate future relationships between entities, such as people or objects. These methods identify potential connections between nodes that are not currently linked, enabling proactive measures to prevent undesirable associations when necessary.

3.4.1 Common Neighbors (CN)

The CN method [Liben-Nowell and Kleinberg 2003] is a straightforward yet effective approach for link prediction in networks. This technique is based on the principle that the likelihood of a direct connection between two nodes increases with the number of shared neighbors. It quantifies this probability by computing the number of common neighbors between two nodes, X and Y, which can be mathematically represented as:

$$CN(X, Y) = |N(X) \cap N(Y)| \quad (10)$$

where $N(X)$ and $N(Y)$ represent the sets of neighboring nodes for X and Y, respectively. This measure is commonly utilized in social and biological networks due to its straightforwardness and ease of interpretation. However, it has limitations in capturing intricate network structures.

3.4.2 Jaccard Coefficient (JC)

The Jaccard Coefficient [Kackson et al. 1989] quantifies the similarity between two sets by computing the ratio of shared neighbors to the total number of distinct neighbors. It is defined as the cardinality of the intersection of the sets divided by the cardinality of their union:

$$JC(X, Y) = \frac{|N(X) \cap N(Y)|}{|N(X) \cup N(Y)|} \quad (11)$$

where $N(X)$ and $N(Y)$ represent the respective sets of neighboring nodes for X and Y.

3.4.3 Adamic Adar (AA)

The Adamic-Adar index [Adamic and Adar 2003] estimates the probability of a missing link between two nodes by considering their shared neighbors. It is computed as the sum of the inverse logarithm of the degree of each common neighbor:

$$AA(X, Y) = \sum_{u \in N(X) \cap N(Y)} \frac{1}{\log(|N(u)|)} \quad (12)$$

where $N(X)$ and $N(Y)$ denote the sets of neighboring nodes for X and Y , respectively, while $N(u)$ represents the set of neighbors associated with a common node u .

3.4.4 Preferential Attachment (PA)

The Preferential Attachment [Newman 2001] mechanism posits that nodes with a greater degree, meaning a higher number of connections, have an increased likelihood of forming new links. This measure is determined by computing the product of the degrees of the two nodes:

$$PA(X, Y) = |N(X)| \cdot |N(Y)| \quad (13)$$

where $N(X)$ and $N(Y)$ represent the degrees of nodes X and Y , respectively.

3.5 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF [Sparck Jones 1972] is a prevalent weighting scheme used to assess a term's (or feature's) significance within a document. This approach merges two components: TF and IDF. The TF component quantifies the number of times a term appears in a given document by analyzing the frequency of all terms present. To adjust for variations in document lengths, the term frequency is normalized by dividing it by the total number of terms in the document. The TF is computed as follows:

$$TF(t) = \frac{\# \text{ of times term } t \text{ appears in document } D}{\text{Total occurrences of all terms in document } D} \quad (14)$$

Inverse Document Frequency (IDF) quantifies the significance of a term across an entire corpus of documents. Terms that occur in a limited number of documents receive a higher IDF value, underscoring their distinctiveness, whereas terms that appear in many documents yield a lower IDF, implying reduced significance. The IDF is calculated as follows:

$$IDF(t) = \log \left(\frac{\text{total\# of documents}}{\# \text{ of documents with term } t \text{ in it}} \right) \quad (15)$$

Multiplying the TF and IDF produces the TF-IDF score, a metric that quantifies the importance of a term in a document relative to its occurrence across the entire corpus.

4 System Overview

4.1 Parameters Determination

All parameters were maintained at their default settings, with the exception of the k value and the SSA fitness function. Consistent k values were applied, as indicated in

[Degirmenci and Karal 2022], and various formulations of the SSA fitness function were evaluated to identify the most appropriate option.

FORMULA	a_dist/diff	a_dist	a_diff/ diff	a_diff - diff
MUSK	0.9465	0.9341	0.9494	0.9494
IONOSPHERE	0.6367	0.6702	0.6367	0.7733
SATIMAGE-2	0.9400	0.9396	0.9400	0.9396
CARDIO	0.6752	0.7620	0.6755	0.6755
THYROID	0.8651	0.8651	0.8669	0.8669
ANNTHYROID	0.6811	0.6811	0.6827	0.6811
MAMMOGRAPHY	0.7603	0.7490	0.7681	0.7490
PIMA	0.5457	0.5457	0.5326	0.5326
GLASS	0.6705	0.6680	0.7119	0.7095
VOWELS	0.8663	0.8663	0.8680	0.8680

Table 2: Results of Different Fitness Functions for K-SAD

a_dist denotes the mean distance of the anomaly points, while diff represents the average deviation of normal points from the overall mean. Similarly, a_diff indicates the mean deviation of anomaly points from the total average.

The formulas are based on the principle that anomalous points should be as distant from the center as possible, while normal points should remain close to it. As a result, the fitness function is maximized by reducing the deviation among normal points and increasing the distance of anomalies. A division operation is used to represent the inverse relationship between these measures. This particular formula was chosen over alternatives because, despite only minor differences, it is more effective at detecting anomalies. Although other formulas may perform better with K-SAD, this one was selected with the potential for superior performance in other methods.

Parameter	Description	Value
num_salps	Number of salps	30
num_iters	Maximum iterations	100
objective_fn	Objective function	a_dist/diff
cluster_range	Determining the # of clusters	[2, 30]
threshold_range	Threshold range	Min-Max

Table 3: Parameters Determination for K-SAD

The parameters utilized in this study, along with their descriptions and assigned values, are presented in Table-3. Each parameter was carefully selected to ensure the effectiveness and precision of the proposed clustering and threshold-based anomaly detection approach.

4.2 Salp Swarm Algorithm

According to Mirjalili [Mirjalili et al. 2017], salps are deep-ocean inhabitants that form chains, resulting in swarm-like behavior. In the Salp Swarm Algorithm, two types of salps are distinguished: leader and follower salps. The leader's position is updated based on the formula provided in Equation-16.

$$x_j^1 = \begin{cases} F_j + c_1((ub_j - lb_j)c_2 + lb_j) & c_3 \geq 0 \\ F_j - c_1((ub_j - lb_j)c_2 + lb_j) & c_3 < 0 \end{cases} \quad (16)$$

In Eq.(16), x_j^1 denotes the position of the leader salp in the j -th dimension. F_j represents the food source, while ub_j and lb_j are the upper and lower bounds, respectively. The variables c_1 , c_2 , and c_3 are random numbers.

$$c_1 = 2e^{-(\frac{4}{L})^2} \quad (17)$$

As shown in Equation-17, c_1 plays a crucial role in balancing exploration and exploitation. L and l denote the total number of iterations and the current iteration count, respectively. The algorithm initially emphasizes exploration and shifts towards exploitation as the iterations advance.

c_2 and c_3 are random values between 0 and 1. These values determine the direction of movement, whether towards positive or negative infinity, and also influence the step size.

$$x_j^i = \frac{1}{2}(x_j^i + x_j^{i-1}) \quad \text{where } i \geq 2 \quad (18)$$

The new position of the i -th follower salp is calculated as the average of its current position and that of the previous salp. The leader salp is represented by 1, and the follower salps start numbering from 2.

Algorithm-1 illustrates the process of determining the threshold stage within the methodology. Once the threshold is identified, the fitness function assesses it to determine the optimal threshold value.

4.3 K-Salp Swarm Anomaly Detection (K-SAD)

Algorithm 2 presents the complete KSAD algorithm, which integrates the K-Salp anomaly detection method and the LOF anomaly detection method using the logical OR operation within each K-Medoid cluster. The dataset is preprocessed in steps [1–3], which include scaling the data and removing the upper and lower bounds. In steps [4–9], the optimal k value is determined based on the silhouette score. In step 11, clusters are identified using the K-medoid algorithm. Steps [12–30] then assess whether the distances exceed the salp threshold and if the LOF score equals -1 ; instances meeting these criteria are classified as anomalies within each cluster. The anomalies detected by both the LOF and Salp methods are combined using a logical OR operation. Finally, the combined labels and anomalies, previously refined using IQR, are integrated, and the ROC-AUC score is computed by comparing the final labels with the true labels.

5 Results and Discussions

Figure 2 presents a flowchart of the Suspicious Nodes Anomaly Prevention (SNAP) Framework. Initially, both node-based and content-based features are extracted from the

Algorithm 1 Salp Swarm Optimization Algorithm for Threshold

```

1: num_salps ← 30
2: num_iterations ← 100
3: min_dist ← minimum distance between salps
4: max_dist ← maximum distance between salps
5: salps ← [random(min_dist, max_dist) for _ in range(num_salps)]
6: for iter in iterations do
7:   best_salp_position ← objective_function(salp, distances)
8:   for i in salps do
9:      $c1 \leftarrow 2 \times \exp \left( - \left( 4 \times \left( \frac{\text{iter}}{\text{num\_iterations}} \right)^2 \right) \right)$ 
10:    c2 ← random values between[0, 1]
11:    c3 ← random values between[0, 1]
12:    ub ← max_dist
13:    lb ← min_dist
14:    a ← c1 × ((ub − lb) × c2 + lb)
15:    if c3 ≥ 0.5 then
16:      new_salp ← int(best_salp_position + a)
17:    else
18:      new_salp ← int(best_salp_position − a)
19:    end if
20:    if new_salp > max_dist then
21:      salps[i] ← max_dist
22:    else if new_salp < min_dist then
23:      salps[i] ← min_dist
24:    else
25:      salps[i] ← new_salp
26:    end if
27:    score ← objective_function(salps[i], distances)
28:    if score > global_best_score then
29:      global_best_score ← score
30:      global_best_threshold ← salps[i]
31:    end if
32:  end for
33: end for

```

Enron dataset. The K-SAD method is then employed to classify nodes as either anomalous or normal for each set of features. A node is labeled as anomalous if both classifications agree on its anomalous nature, and as normal if both concur on its normality. In cases where one classification indicates an anomaly while the other does not, the node is deemed suspicious. Link prediction is subsequently applied to these suspicious nodes to identify potential future connections, and connections exceeding a predefined threshold are further scrutinized by experts.

Algorithm 2 K-medoid Salp Swarm Anomaly Detection Algorithm

Input: Infinite data stream $D = \{x_1, x_2, \dots, x_n, \dots\}$ where x_i is a vector of more than one dimension.

Output: Anomaly labeled data $A = \{0, 0, 1, \dots\}$ (1 indicates anomalies in the data). _

```

1:  $D_{\text{scaled}} \leftarrow$  scale the  $D$ 
2:  $D_{\text{IQR}} \leftarrow$  upper and lower in  $D$ 
3:  $D_{\text{rest}} \leftarrow D_{\text{scaled}} - D_{\text{IQR}}$ 
4: for  $k$  in range  $[2, 30]$  do
5:    $k_{\text{medoids}} \leftarrow$  KMedoids( $n_{\text{clusters}} = k$ )
6:   if silhouette_avg > best_silhouette_score then
7:     best_silhouette_score  $\leftarrow$  silhouette_avg
8:     best_k  $\leftarrow k$ 
9:   end if
10: end for
11: Clusters  $\leftarrow$  K_medoid(best_k,  $D_{\text{rest}}$ )
12: for cluster in Clusters do
13:   cluster_threshold  $\leftarrow$  Salp_Swarm_Algorithm(distances)
14:   if (distances > cluster_threshold) then
15:     salp_label  $\leftarrow$  1
16:   else
17:     salp_label  $\leftarrow$  0
18:   end if
19:   lof_model  $\leftarrow$  LocalOutlierFactor()
20:   lof_scores  $\leftarrow$  list(lof_model_predict(cluster))
21:   if (lof_scores == -1) then
22:     lof_label  $\leftarrow$  1
23:   else
24:     lof_label  $\leftarrow$  0
25:   end if
26: end for
27: combined_label  $\leftarrow$  logical_or(lof_label, salp_label)
28: final_anomalies  $\leftarrow$  merge( $D_{\text{IQR}}$ , combined_label)
29: roc_auc_score  $\leftarrow$  calculate_roc_auc(final_anomalies, true_labels)

```

5.1 Dataset

The Enron corpus, introduced by Klimt and Yang [Klimt and Yang 2004], comprises 619,446 emails from 158 users. Following preprocessing steps such as cleaning and filtering, the dataset was refined to 200,399 messages. This corpus was chosen due to its extensive size, diversity, and authenticity, as it accurately represents real-world corporate communication patterns. It serves as a valuable resource for extracting both content-based and node-based features, which play a crucial role in anomaly detection and analysis. With 30,091 identified threads, the dataset provides a comprehensive foundation for studying communication sequences and thread dynamics, where 61.63% of emails are part of threads, typically involving brief exchanges. Given its scale and variability, this

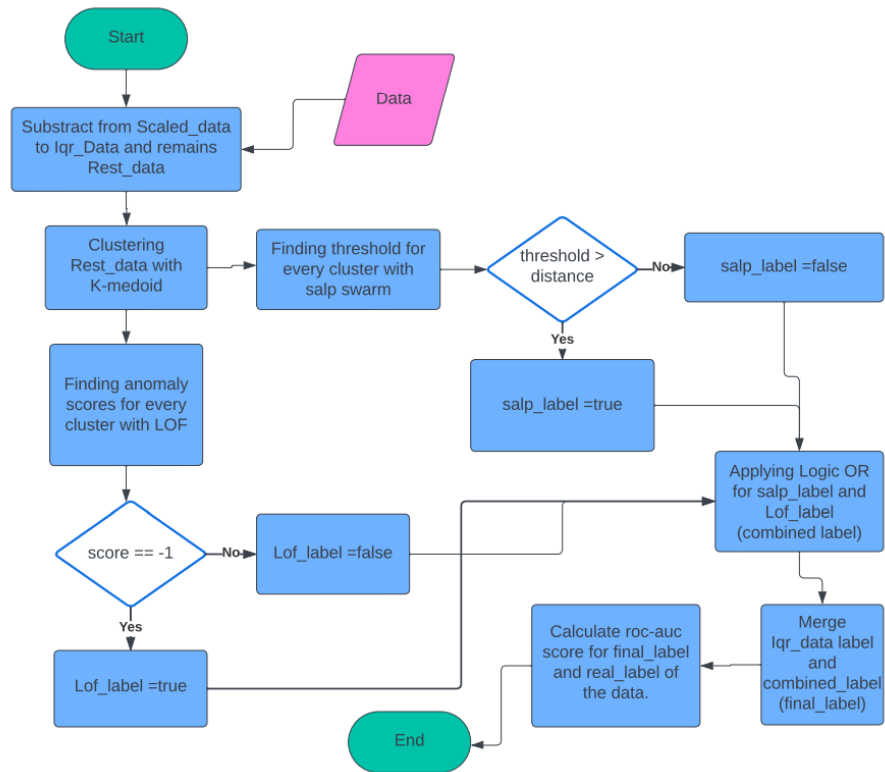


Figure 1: Flowchart of a K-SAD Method

dataset offers meaningful insights into email interactions, making it highly suitable for investigating advanced anomaly detection methodologies.

5.2 Performans Criteria

5.2.1 Accuracy

Accuracy [Wu et al. 2005] is a fundamental evaluation metric derived from the confusion matrix. It quantifies the proportion of correctly classified instances, including both true positives and true negatives, relative to the total number of instances in the dataset. The mathematical formulation of accuracy is provided in Equation-19.

$$Accuracy = \frac{True\ Positive(TP) + True\ Negative(TN)}{(TP) + (TN) + False\ Positive(FP) + False\ Negative(FN)} \quad (19)$$

5.2.2 F1-score

The F1 Score [Powers 2020] is the harmonic mean of Precision and Recall, offering a balanced assessment of these two metrics. Precision, or positive predictive value,

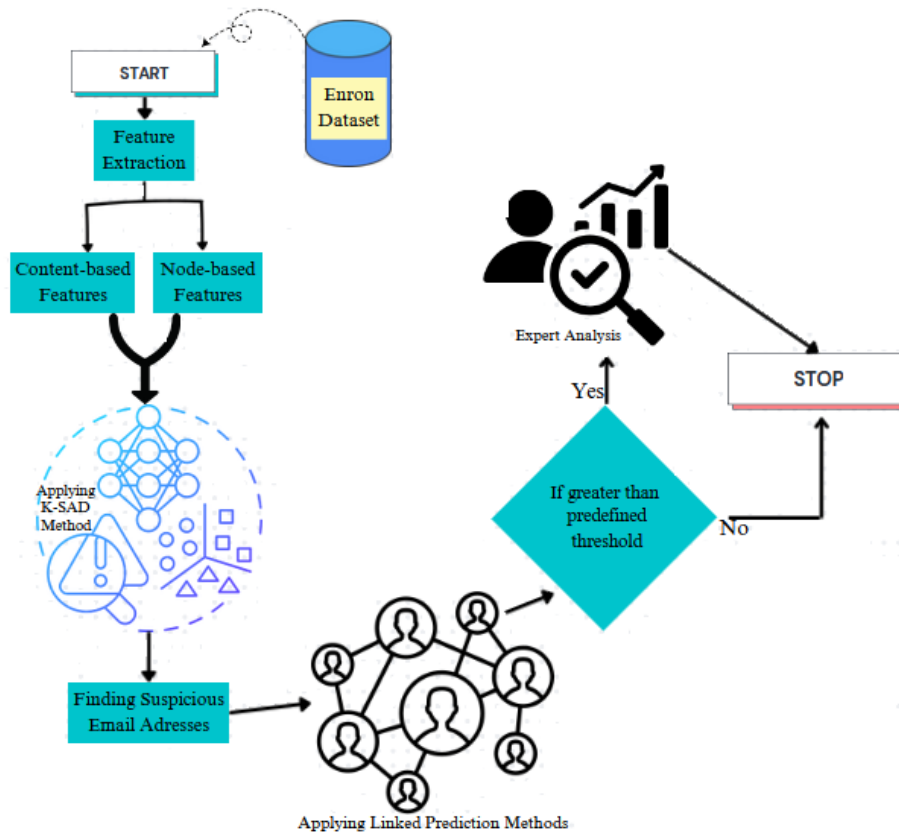


Figure 2: Flowchart of a SNAP Framework

quantifies the proportion of correctly predicted positive instances among all predicted positives, while Recall, also referred to as sensitivity, measures the proportion of actual positive instances that were correctly identified. The F1 Score is particularly important in scenarios with imbalanced class distributions, as it effectively captures the trade-off between Precision and Recall.

5.2.3 ROC-AUC Score

The Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) score [Metz 1978] is a widely utilized metric for evaluating the effectiveness of anomaly detection models. It measures the model's capability to differentiate between true positives and false positives by analyzing both sensitivity and specificity simultaneously.

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (20)$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} \quad (21)$$

The True Positive Rate (TPR), or sensitivity, indicates the proportion of actual anomalies that are correctly identified by the model, reflecting its ability to detect abnormal instances. Conversely, the False Positive Rate (FPR) represents the fraction of normal data points that are incorrectly classified as anomalies, serving as the complement of specificity. The ROC curve visualizes the relationship between TPR and FPR, while the Area Under the Curve (AUC) quantifies the model's overall performance.

A higher ROC-AUC score signifies that the model effectively distinguishes anomalies while reducing false positive classifications.

5.3 Comparison

To validate the effectiveness and acceptance of the proposed K-SAD method, comparisons were made with 10 different datasets and 11 distinct methods from the study by Degirmenci and Karal [Degirmenci and Karal 2022]. As shown in Table-4, the proposed method achieved the best or near-best results on 5 datasets and performed above average on 3 datasets. These results demonstrate the validity of the proposed method.

Dataset	K-SAD	Scaling	Comparison
Musk	0.9987	y	almost same
Ionosphere	0.6568	n	below average
Satimage	0.9664	y	best except LODA
Cardio	0.6879	both	above average
Thyroid	0.8651	y	best
Anthyroid	0.6811	y	below average
Mammography	0.7603	both	best except LODA
Pima	0.5623	n	above average
Glass	0.7499	n	best
Vowels	0.8683	n	above average

Table 4: Comparison with Other Methods

5.4 Results

The Enron dataset contains multiple folders, which can result in duplicate emails appearing across different locations. For example, some emails are found in both the sent and received folders. To streamline the analysis, only emails from the sent folder were considered, following the approach adopted in prior research that also focused exclusively on this folder [Styler 2011].

A NetworkX graph was generated based on the "from-to" relationships in the sent emails, initially comprising 15,868 nodes and 32,433 edges. However, some nodes were isolated, leading to the formation of five separate subgraphs with sizes of 15,852, 10, 2, 2, and 2. The four smaller subgraphs were excluded, and the corresponding email addresses were removed from the analysis. The excluded addresses included rjbaker@ttu.edu, lcampbe, tim.derrick, patricia.hunter, tom.nemila, denise.williams, mark.fisher, andre.rast, kurt.anderson, jeff.duff, joe.chapman, julie.johnson, all.states,

click.home, kid.mailing, and enron.announcements. Additionally, email addresses without a specified domain were assumed to belong to @enron.com. The segmentation of the graph into five distinct components is illustrated in Figure-3.

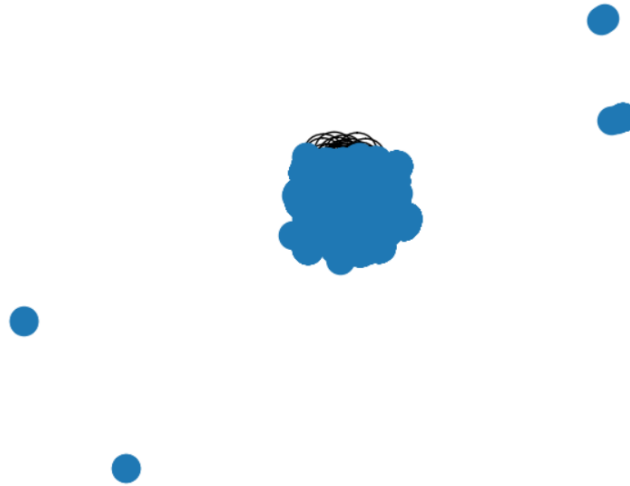


Figure 3: Enron Graph Visualization

Several node-based features were derived from the NetworkX graph, including degree, eigenvector centrality, closeness centrality, and betweenness centrality [Freeman et al. 2002]. Additional extracted features included the clustering coefficient [Soffer 2005], PageRank [Page 1999], HITS (hubs and authorities) [Kleinberg 1999], eccentricity [Hage and Harary 1995], and community structure. Furthermore, network-based similarity and connectivity measures such as average common neighbors, average Jaccard coefficient, average Adamic-Adar score, preferential attachment, and effective size were also computed.

Additionally, we utilized partially labeled Enron nodes, where a previous study [Noever 2020] identified certain individuals as "persons of interest" (POIs). By comparing the labels generated by our approach with these POI annotations, which represent anomalous emails, we obtained a ROC-AUC score of 0.6781. Among the 18 identified anomalies, our method misclassified only one as normal. The confusion matrix illustrating the classification performance for node-based features is presented in Figure-4.

While our approach effectively detected anomalies, it exhibited a tendency to misclassify normal instances as anomalies. Future research will focus on addressing this limitation by reducing false positives and enhancing the accuracy of normal data classification.

In a similar approach, TF-IDF was applied to the email content, and duplicates were removed, resulting in a reduced dataset of 96,148 emails. The proposed anomaly detection method was subsequently used to determine if these connections were anomalous. As illustrated in the confusion matrix in Figure-5, the method showed a high success rate in anomaly detection, although its performance in classifying normal instances was slightly less accurate.

The results presented in Table-5 highlight the performance differences between

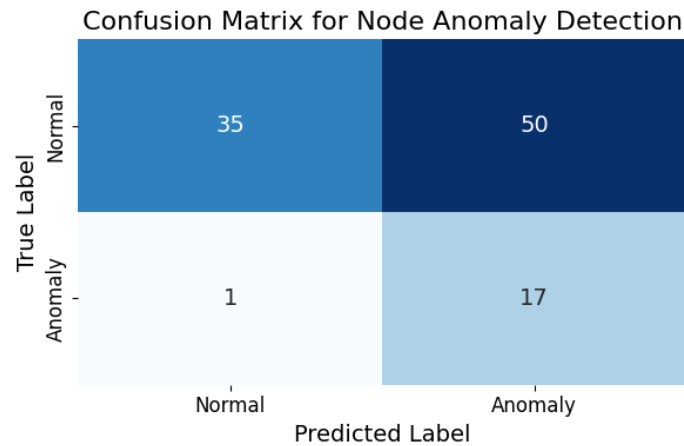


Figure 4: Anomaly Detection Results with Node-based Features

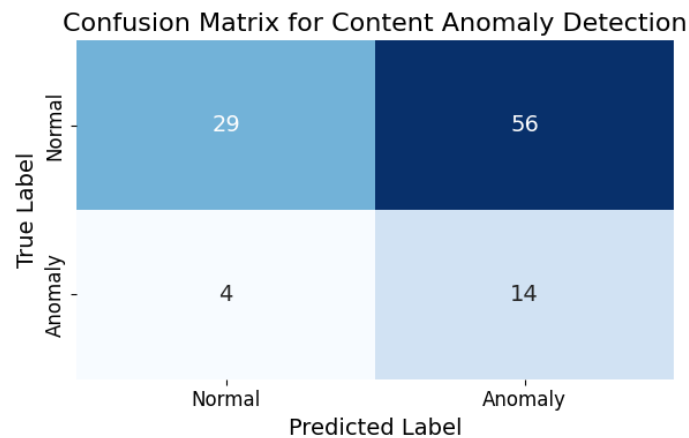


Figure 5: Anomaly Detection Results with Content-based Features

Content Anomaly Detection (CAD) and Node Anomaly Detection (NAD). NAD demonstrates a clear advantage in all metrics, achieving higher accuracy (0.5049), F1 score (0.5473), and ROC-AUC score (0.6781). Both methods exhibit strong precision, with NAD at 0.8467 and CAD at 0.7602, but NAD proves more effective at minimizing false positives. However, the recall values suggest there is still potential for improvement in capturing all true anomalies. Additionally, NAD identifies more true positives (17) and generates fewer false positives (50) compared to CAD, indicating its superior ability to utilize the relationships within the data.

The relatively lower performance of content-based anomaly detection (CAD) can be attributed to its reliance on TF-IDF. While TF-IDF is computationally efficient, it lacks the ability to capture semantic relationships within the data, which limits its effectiveness in distinguishing anomalies in complex, context-dependent scenarios. In contrast, node-based anomaly detection (NAD) demonstrates superior performance due to its ability

to leverage the relational structure of the data through graph-based features such as PageRank, betweenness centrality, and Adamic-Adar score. These features exploit the interconnected nature of nodes, enabling the model to identify anomalies based on structural deviations. This advantage explains the higher recall and ROC-AUC scores observed in NAD, highlighting its robustness compared to CAD in detecting anomalies accurately.

CAD exhibits lower performance largely because it depends on TF-IDF. Although TF-IDF is computationally efficient, it fails to capture semantic relationships within the data, which limits its ability to differentiate anomalies in complex, context-dependent scenarios. In contrast, NAD outperforms CAD by capitalizing on the relational structure of the data, utilizing graph-based features such as PageRank, betweenness centrality, and the Adamic-Adar score. These features harness the interconnected nature of nodes, enabling the model to identify anomalies through structural deviations. This advantage is reflected in the higher recall and ROC-AUC scores observed for NAD, underscoring its robustness in accurately detecting anomalies.

These findings underscore the advantages of a node-based methodology in this context, highlighting its potential for more dependable anomaly detection. Moreover, the ROC curve presented in Figure-6 clearly demonstrates that anomalies identified using node-based features outperform those detected via content-based features.

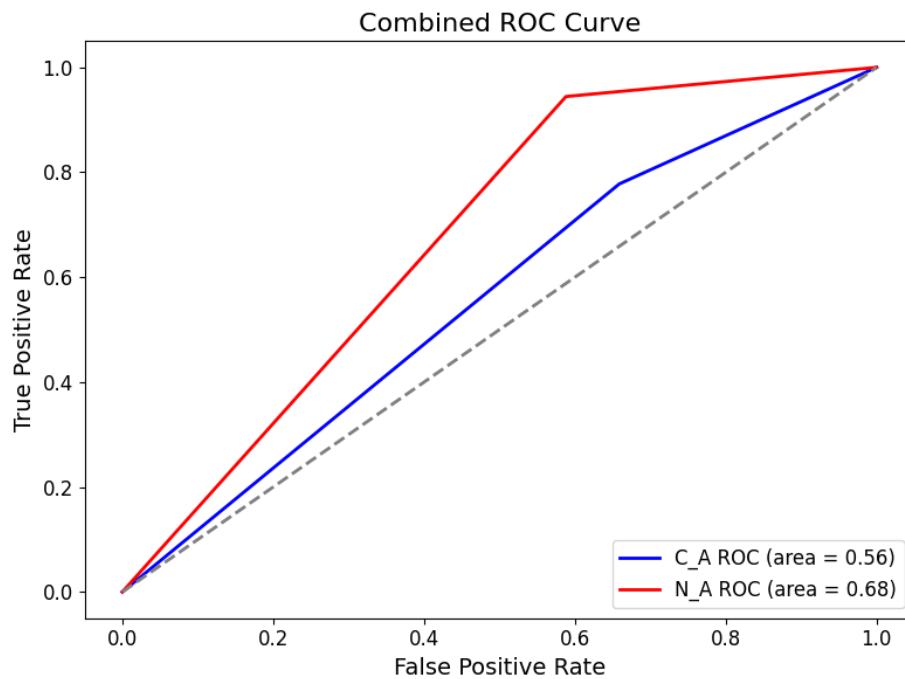


Figure 6: Comparison ROC Curve Results for Content and Node-based Features

In the analysis of labeled data, 56 instances were consistently identified as anomalies by both node-based and content-based feature methods. Within the POI labels, there

Method	Accuracy	F1 Score	ROC-AUC	Precision	Recall	TP	FP
CAD	0.4175	0.4612	0.5595	0.7602	0.4175	14	56
NAD	0.5049	0.5473	0.6781	0.8467	0.5049	17	50

Table 5: Content and Node-based Features Results

were 18 anomalies, with the remaining instances classified as normal email addresses. In contrast, both the node-based and content-based approaches detected an additional 38 anomalies compared to the POI labels. Furthermore, while 22 nodes were categorized as normal by both methods, the POI labels identified only one node as anomalous with the rest deemed normal. Instances where one method classified an instance as anomalous and the other as normal—termed suspicious email addresses—are detailed in Table-6.

Email_Address	C_L	N_L	R_L
kenneth.lay	0	1	1
richard.shapiro	0	1	0
andrew.fastow	0	1	1
bill.cordes	1	0	0
david.haug	0	1	0
diomedes.christodoulou	1	0	0
gene.humphrey	1	0	0
james.bannantine	1	0	0
joe.hirko	0	1	1
joe.kishkill	0	1	0
john.buchanan	1	0	0
ken.powers	1	0	0
lou.pai	0	1	0
mark.pickering	1	0	0
marty.sunde	0	1	0
matthew.scrimshaw	1	0	0
michael.moran	1	0	0
mittell.taylor	1	0	0
rebecca.mcdonald	0	1	0
richard.lewis	1	0	0
rick.bergsieker	1	0	0
robert.hayes	1	0	0
tod.lindholm	0	1	0
tracy.foy	1	0	0
w.duran	0	1	0

Table 6: Suspicious Nodes Content, Node, and Real Labeled Data

Table-7 presents the risky connections identified through link prediction for suspicious nodes. The 'Between' column lists connections between suspicious nodes and all

nodes in the dataset, while the 'Only' column details connections solely among suspicious nodes. The 'Anomalies' column highlights connections between suspicious nodes and anomalies. Link prediction was carried out using four methods (CN, JC, AA, PA) with specific thresholds: a threshold of 0.7 was applied for 'All' connections, and 0.5 for the remaining cases. This higher threshold was required because no connections were observed below that level. These threshold settings ensured that the connections predicted by all four methods were reliable, thereby enhancing the accuracy of the link prediction process.

From	To	Threshold	Email Chain	Between	Num
diomedes.christodoulou	sanjay.bhatnagar	0.7	No	All	1
james.bannantine	diomedes.christodoulou	0.5	Yes	Only	2
ken.rice	david.w.delaine	0.5	Yes	Anomalies	3
andrew.fastow	mark.koenig	0.5	No	Anomalies	4

Table 7: Suspicious Nodes Contacts

Table-8 displays the email contents. Although initial detection identified a higher number of connections—four for "All," three for "Only," and nine for "Anomalies"—a thorough review subsequently pinpointed specific emails as anomalous. These emails were flagged because their content pertained to financial and governmental matters, reflecting the corruption context inherent in the Enron dataset. For instance, one email discusses discounts and media while implying a shift to telephone communication to avoid creating a record. Another email appears to involve governmental approval procedures, suggesting attempts to expedite acceptance by leveraging influential individuals. A third email congratulates a recipient on a workplace role while proposing an offer that could be interpreted as a bribe. Finally, a fourth email subtly raises concerns about low credit ratings and an increased likelihood of engaging in anomalous activities.

To evaluate the computational efficiency of the proposed SNAP framework, we specifically focused on its content-based anomaly detection component, measuring both runtime and memory consumption across its key processing stages. As summarized in Table 9, the TF-IDF vectorization and cosine similarity computation stage completed in 11.43 seconds, utilizing approximately 54.04 MB of memory. The K-Medoids clustering, which facilitates the structural grouping of message content, required 33.24 seconds with a peak memory usage of 5561.41 MB, primarily due to the overhead of distance matrix computations. The SSA process was applied separately to determine a threshold for each cluster, completing in a total of 3.62 seconds using approximately 0.163 MB of memory. During this process, the average time per cluster was 1.21 seconds, and the average memory usage was 0.055 MB. This cluster-specific thresholding approach enabled more precise and distinct analyses within the content-based anomaly detection process. The link prediction phase, which includes proximity-based metrics applied on content similarity graphs, was the most computationally intensive, taking 433.05 seconds and consuming approximately 121.81 MB of memory. These results confirm that, despite

Num	Suspicious Mail Content
1	Joe asked me to get in touch with you regarding Indian media reports/inquiries on possible sale or sell down of Dabhol and to discuss a possible statement in the event we get additional inquiries. Give me a call - 713-853-1586.
2	James emphasizes the need for early action to avoid delays in the governmental approval stage of Project California. He notes that Jose, with his expertise and local knowledge, is best suited to handle government-related assessments and regulatory approvals. James recommends including Jose in the process to effectively manage governmental interactions and prevent potential delays.
3	Ken Rice congratulates David Delaney on his new role and offers a building in Bellaire for lease or purchase. The space is suggested for car storage, with a rental cost of \$12-\$15 per square foot annually. Leasing one-third of the building would cost about \$1,600 per month, potentially less than a climate-controlled storage unit. Ken invites David to discuss further if interested.
4	A mid-day summary compares the Enron Online (EOL) trades on October 22, 2001, with the full-day transaction counts from September 21, 2001. A detailed recap can be provided at the end of the day if needed. The counterparties listed have higher credit ratings than Enron. It is suggested to contact a relevant person for any credit-related questions.

Table 8: Suspicious Nodes Contacts Contents

Module	Runtime (s)	Peak Memory (MB)
TF-IDF + Cosine Distance	11.43	54.04
K-Medoids Clustering	33.24	5561.41
SSA Threshold Optimization	1.21	0.055
Link Prediction	433.05	121.81

Table 9: Runtime and Memory Analysis

the computational demands of clustering and link prediction, the content-based module of the SNAP framework remains efficient and scalable for medium-sized datasets such as Enron.

5.5 Discussions

In this study, we introduced and assessed a framework for anomaly prevention using a real-world dataset. We developed a novel, nature-inspired method for anomaly detection that leverages both node and content features. Although nature-inspired algorithms are traditionally applied to clustering tasks, their use for threshold estimation in our approach represents a significant innovation. Anomalies, by definition, are data points that do not conform to any predefined grouping. By first clustering the data and then establishing an optimal threshold, we can classify data points that fall outside this threshold as anomalies. However, our current implementation may use an overly strict threshold, which results in a higher rate of false positives—normal data points mistakenly labeled as anomalies. Future research will focus on refining the clustering process and optimizing the threshold to improve anomaly detection accuracy.

Building upon this, One promising direction for improvement is to formalize the anomaly detection thresholding as a constrained multi-objective optimization problem that jointly minimizes false positives and maximizes ROC-AUC. Additionally, deep reinforcement learning (DRL) techniques like Twin Delayed Deep Deterministic policy gradient (TD3) could be explored in future work to learn adaptive threshold policies in dynamic graph environments

The higher false-positive rates observed in both methods can be partially attributed to the challenges in setting optimal thresholds for anomaly classification. To address this, future work could explore the use of advanced optimization techniques, such as genetic algorithms or particle swarm optimization, to fine-tune threshold values dynamically. Additionally, integrating semantic models like BERT for content analysis could further enhance CAD's ability to distinguish anomalies more accurately, reducing the likelihood of false positives.

For content analysis, we employed the traditional TF-IDF method, chosen for its efficiency and broad acceptance as a conventional technique. Although more semantically advanced methods like BERT offer superior semantic understanding, they tend to be slower, which was a critical consideration given the large volume of emails in the dataset. By applying the same anomaly detection approach to both node-based and content-based features, instances concurrently identified as anomalies were classified as confirmed anomalies, while those consistently recognized as normal were considered normal. Our primary focus, however, was on cases where one method flagged an instance as anomalous while the other did not; these were designated as suspicious nodes.

Further analysis was conducted on the potential future connections of these suspicious nodes, as such links might entail additional anomaly risks. These connections were flagged for deeper investigation, enabling preventative measures to be implemented before suspicious activities evolve into major anomalies. For example, within the Enron dataset, suspicious financial-related emails could be referred to experts for closer scrutiny, thereby aiding in the clearer identification of individuals engaged in unethical practices. Although several high-level executives have already been implicated, this framework may also reveal previously undetected offenders.

In selecting link prediction methods for this framework, traditional techniques were favored due to their speed and low computational cost, aligning well with the runtime-sensitive design of our system. While graph neural network (GNN)-based link predictors such as GraphSAGE or GAT offer greater predictive capabilities, they require significantly more resources and time. Future versions of the framework may integrate GNNs, provided that scalability challenges are addressed.

6 Conclusions

In this study, a novel framework was proposed to contribute to anomaly prevention. As part of this framework, a new anomaly detection technique, based on clustering and nature-inspired algorithms, was introduced to identify potentially anomalous data. The study utilized email data as the social network dataset. The proposed anomaly detection technique enabled both node-based and content-based analyses. In both approaches, instances identified as anomalies were labeled as anomalies, and those deemed normal were labeled as normal. However, instances classified as normal by one method but as anomalous by the other were labeled as suspicious nodes. Subsequently, the connections established by these suspicious email addresses were examined, and those exceeding a certain threshold were flagged as potential anomalies. This approach allowed for the prevention of anomalous activities by identifying potential anomalies before any actual anomalous behavior occurred. Unlike previous studies in the literature, this research focused on anomaly prevention rather than mere detection.

Moreover, unlike conventional anomaly detection methods that primarily identify anomalies and conclude the process, this study introduced a proactive strategy by not only detecting anomalies but also identifying suspicious nodes and assessing their potential

risk. By incorporating expert intervention to evaluate these suspicious nodes, the study contributes to a more comprehensive anomaly prevention strategy, ensuring that actions can be taken before the anomaly escalates. This research advances the field by addressing the challenge of anomaly prevention, an area that has been less explored in comparison to anomaly detection, offering new perspectives and methodologies for safeguarding against anomalous behaviors in static networks.

7 Future Work

For future research, several promising directions exist for enhancing the proposed anomaly prevention framework. One key avenue is the integration of advanced semantic models, such as BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa, for content-based anomaly analysis. While this study employed TF-IDF for its speed and efficiency, particularly given the large volume of email content, we acknowledge its limitations in capturing deeper semantic relationships, including challenges like polysemy and synonymy. Contextual language models such as BERT and RoBERTa are better equipped to detect subtle semantic anomalies in textual communication and could significantly improve the accuracy of content-based anomaly detection, despite their higher computational costs.

The scalability of the proposed methods remains a critical area for improvement. As datasets grow in size and complexity, leveraging distributed computing frameworks like Apache Spark or TensorFlow Distributed could be beneficial in enhancing performance and managing large-scale data effectively.

To validate the generalizability of the framework, future experiments could involve additional datasets from diverse domains, such as finance, healthcare, and cyber security. Testing the methods on these datasets would not only demonstrate their adaptability but also reveal domain-specific challenges, guiding further refinements. Such experiments would provide a broader perspective on the applicability and robustness of the anomaly prevention framework.

The current implementation adopts a static representation of the graph, primarily for computational efficiency and conceptual clarity. This design choice enables rapid detection and highlights the core principles of the proposed SNAP framework. However, future iterations of the framework will aim to incorporate temporal graph structures, dynamic behavior modeling, and adaptive thresholding mechanisms to better reflect real-world complexities.

References

- [Adamic and Adar 2003] Adamic, L. A., Adar, E. (2003). Friends and neighbors on the web. *Social Networks*, 25(3), 211–230.
- [Alsaleh and Binsaeedan 2021] Alsaleh, A., Binsaeedan, W. (2021). The influence of salp swarm algorithm-based feature selection on network anomaly intrusion detection. *IEEE Access*, 9, 112466–112477.
- [Alzaqebah et al. 2023] Alzaqebah, A., Aljarah, I., Al-Kadi, O. (2023). A hierarchical intrusion detection system based on extreme learning machine and nature-inspired optimization. *Computers Security*, 124, 102957.
- [Apoorva and Sangeetha 2021] Apoorva, K. A., Sangeetha, S. (2021). Deep neural network and model-based clustering technique for forensic electronic mail author attribution. *SN Applied Sciences*, 3(3), 348.

- [Assouli et al. 2021] Assouli, N., Benahmed, K., Gasbaoui, B. (2021). How to predict crime—informatics-inspired approach from link prediction. *Physica A: Statistical Mechanics and Its Applications*, 570, 125795.
- [Bastami et al. 2021] Bastami, E., Mahabadi, A., Taghizadeh, E. (2019). A gravitation-based link prediction approach in social networks. *Swarm and Evolutionary Computation*, 44, 176–186.
- [Bountakas and Xenakis 2023] Bountakas, P., Xenakis, C. (2023). Helped: Hybrid ensemble learning phishing email detection. *Journal of Network and Computer Applications*, 210, 103545.
- [Breunig et al. 2000] Breunig, M. M., Kriegel, H.-P., Ng, R. T., Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. ACM.
- [Corneli et al. 2019] Corneli, M., Bouveyron, C., Latouche, P., Rossi, F. (2019). The dynamic stochastic topic block model for dynamic networks with textual edges. *Statistics and Computing*, 29, 677–695.
- [Degirmenci and Karal 2022] Degirmenci, A., Karal, O. (2022). Efficient density and cluster based incremental outlier detection in data streams. *Information Sciences*, 607, 901–920.
- [Freeman et al. 2002] Freeman, L. C., Others. (2002). Centrality in social networks: Conceptual clarification. *Social Network: Critical Concepts in Sociology*. Londres: Routledge, 1, 238–263.
- [Hage and Harary 1995] Hage, P., Harary, F. (1995). Eccentricity and centrality in networks. *Social Networks*, 17(1), 57–63.
- [Jáñez-Martino et al. 2023] Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., Alegre, E. (2023). A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artificial Intelligence Review*, 56(2), 1145–1173.
- [Jiao et al. 2024] Jiao, X., Wan, S., Liu, Q., Bi, Y., Lee, Y.-L., Xu, E., ... Zhou, T. (2024). Comparing discriminating abilities of evaluation metrics in link prediction. *Journal of Physics: Complexity*, 5(2), 025014.
- [Kackson et al. 1989] Jackson, D. A., Somers, K. M., Harvey, H. H. (1989). Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? *The American Naturalist*, 133(3), 436–453.
- [Kaufman and Rousseeuw 2009] Kaufman, L., Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley Sons.
- [Khayyat 2023] Khayyat, M. M. (2023). Improved bacterial foraging optimization with deep learning based anomaly detection in smart cities. *Alexandria Engineering Journal*, 75, 407–417.
- [Kirchner and Gade 2011] Kirchner, C., Gade, J. (2011). Implementing social network analysis for fraud prevention. *CGI Gr. Ind.*
- [Klimt and Yang 2004] Klimt, B., Yang, Y. (2004). Introducing the Enron corpus. *CEAS*, 4, 1.
- [Kleinberg 1999] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5), 604–632.
- [Lande et al. 2020] Lande, D., Fu, M., Guo, W., Balagura, I., Gorbov, I., Yang, H. (2020). Link prediction of scientific collaboration networks based on information retrieval. *World Wide Web*, 23, 2239–2257.
- [Liben-Nowell and Kleinberg 2003] Liben-Nowell, D., Kleinberg, J. (2003). The link prediction problem for social networks. *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, 556–559.
- [Liu et al. 2008] Liu, F. T., Ting, K. M., Zhou, Z.-H. (2008). Isolation forest. *2008 Eighth IEEE International Conference on Data Mining*. IEEE.
- [Metz 1978] Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8, 283–298. Elsevier.

- [Min et al. 2024] Min, H., Lei, X., Wu, X., Fang, Y., Chen, S., Wang, W., Zhao, X. (2024). Toward interpretable anomaly detection for autonomous vehicles with denoising variational transformer. *Engineering Applications of Artificial Intelligence*, 129, 107601.
- [Mirjalili et al. 2017] Mirjalili, S., Gandomi, A. H., Mirjalili, S. Z., Saremi, S., Faris, H., Mirjalili, S. M. (2017). Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. *Advances in Engineering Software*, 114, 163–191.
- [Mohajer et al. 2024] Mohajer, A., Hajipour, J., Leung, V. C. (2024). Dynamic offloading in mobile edge computing with traffic-aware network slicing and adaptive TD3 strategy. *IEEE Communications Letters*.
- [Newman 2001] Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2), 025102.
- [Noever 2020] Noever, D. (2020). The Enron Corpus: Where the Email Bodies are Buried? arXiv Preprint arXiv:2001.10374.
- [Ott et al. 2021] Ott, S., Meilicke, C., Samwald, M. (2021). SAFRAN: An interpretable, rule-based link prediction method outperforming embedding models. arXiv Preprint arXiv:2109.08002.
- [Page 1999] Page, L. (1999). The PageRank citation ranking: Bringing order to the web. Technical Report.
- [Poobalan et al. 2025] Poobalan, A., Ganapriya, K., Kalaivani, K., Parthiban, K. (2025). A novel and secured email classification using deep neural network with bidirectional long short-term memory. *Computer Speech Language*, 89, 101667.
- [Powers 2020] Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. arXiv Preprint arXiv:2010.16061.
- [Rahman et al. 2021] Rahman, M. S., Halder, S., Uddin, M. A., Acharjee, U. K. (2021). An efficient hybrid system for anomaly detection in social networks. *Cybersecurity*, 4(1), 10.
- [Rousseeuw 1986] Rousseeuw, P. J. (1986). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- [Samad et al. 2021] Samad, A., Azam, M., Qadir, M. (2021). Structural Importance-based Link Prediction Techniques in Social Network. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, 7(25).
- [Saxena et al. 2022] Saxena, A., Fletcher, G., Pechenizkiy, M. (2022). HM-EIIC: Fairness-aware link prediction in complex networks using community information. *Journal of Combinatorial Optimization*, 44(4), 2853–2870.
- [Shao et al. 2022] Shao, C., Zheng, S., Gu, C., Hu, Y., Qin, X. (2022). A novel outlier detection method for monitoring data in dam engineering. *Expert Systems with Applications*, 193, 116476.
- [Sparck Jones 1972] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.
- [Styler 2011] Styler, W. (2011). The enrnsent corpus. University of Colorado-Boulder, 1–7.
- [Soffer 2005] Soffer, S. N., Vazquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 71(5), 057101.
- [Van Vlasselaer et al. 2015] Van Vlasselaer, V., Akoglu, L., Eliassi-Rad, T., Snoeck, M., Baesens, B. (2015). Guilt-by-constellation: Fraud detection by suspicious clique memberships. 2015 48th Hawaii International Conference on System Sciences, 918–927. IEEE.
- [Wan et al. 2015] Wan, X., Wang, W., Liu, J., Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Medical Research Methodology*, 14, 1–13.

[Witten et al. 2005] Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., Data, M. (2005). Practical machine learning tools and techniques. *Data Mining*, 2, 403–413. Elsevier Amsterdam, The Netherlands.

[Wu et al. 2005] Wu, X. (2019). A trust-based detection scheme to explore anomaly prevention in social networks. *Knowledge and Information Systems*, 60, 1565–1586.

[Xu et al. 2022] Xu, He, Zhang, L., Li, P., Zhu, F. (2022). Outlier detection algorithm based on k-nearest neighbors-local outlier factor. *Journal of Algorithms Computational Technology*, 16, 17483026221078111.

[Xu et al. 2023] Xu, Hongzuo, Pang, G., Wang, Y., Wang, Y. (2023). Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 12591–12604.

[Yang and Mohajer 2025] Yang, J., Mohajer, A. (2025). Multi objective constellation optimization and dynamic link utilization for sustainable information delivery using PD-NOMA deep reinforcement learning. *Wireless Networks*, 31(2), 1839-1859.

[Zhou 2023] Zhou, T. (2023). Discriminating abilities of threshold-free evaluation metrics in link prediction. *Physica A: Statistical Mechanics and Its Applications*, 615, 128529.

[Zhou and Mohajer 2024] Zhou, G., Mohajer, A. (2024). Blind reconfigurable intelligent surfaces for dynamic offloading in fixed-NOMA mobile edge networks. *International Journal of Sensor Networks*, 46(3), 142-160.