Information Diffusion Analysis
Enron-Email Dataset

Submitted by :

- **Dhrithi K** [22BDS018]
- **Pakhi Singhal** [22BDS042]
- **Preethi Varshala** [22BDS045]
- **Ravi Raj** [22BDS051]

Under the Guidance of :

- Prof. **Dr Utkarsh Khaire**

Abstract

This report provides a detailed analysis of the Enron Email network, focusing on the structure and flow of information to identify key influencers and understand communication dynamics within a large organization. Centrality measures, including degree, betweenness, closeness, eigenvector centrality, PageRank, clustering coefficient, and modularity, are used to identify individuals with significant roles in information dissemination. Community detection techniques, such as the Girvan-Newman method and modularity-based clustering, reveal tightly connected sub-groups and their interaction patterns. Additionally, k-club, k-clan, and k-core analysis are employed to identify groups of individuals that are tightly knit based on a minimum number of connections, uncovering important clusters and key players in the network. The network's density and connected components are also analyzed to assess overall connectivity. Information diffusion models, including the Independent Cascade Model (ICM) and Linear Threshold Model (LTM), simulate the propagation of information, highlighting primary diffusion pathways. The findings uncover critical nodes, pathways, and community structures crucial for effective communication, with implications for organizational analysis, crisis management, and optimizing information flow within networks.

Introduction

- **Background on the Enron Scandal and Email Dataset**

The Enron Corporation, founded in 1985 through the merger of Houston Natural Gas and InterNorth, was once a leading American energy company headquartered in Houston, Texas. Under the leadership of executives like Kenneth Lay and Jeffrey Skilling, Enron expanded rapidly, diversifying into electricity, natural gas, communications, and complex financial instruments such as derivatives. By the late 1990s, Enron had become one of the largest and most influential companies in the United States.

However, in late 2001, Enron's rapid rise was overshadowed by the revelation of massive accounting fraud. Investigations exposed that Enron's executives had employed special purpose entities (SPEs) and off-balance-sheet partnerships to conceal enormous debt and inflate reported profits, misleading investors, regulators, and stakeholders. This manipulation led to a dramatic collapse in investor confidence and Enron's eventual bankruptcy filing in December 2001. The scandal also resulted in the dissolution of Arthur Andersen LLP, Enron's accounting firm, and triggered sweeping regulatory reforms, most notably the Sarbanes-Oxley Act, which aimed to enhance corporate transparency and accountability.

In the aftermath of the scandal, the Federal Energy Regulatory Commission (FERC) released a substantial portion of Enron's internal communications, including over 500,000 emails exchanged between approximately 150 employees from 2000 to 2002. These emails, spanning routine work exchanges to sensitive decision-making discussions, have become a critical resource for academic and industry research. The Enron email dataset is widely used in studies of social network analysis, machine learning, natural language processing, and organizational behavior, providing a real-world lens into the structure and dynamics of communication within a large corporation.

● Importance of Analyzing Social Networks

Social networks are vital for understanding how information, influence, and resources circulate within groups or organizations. Analyzing Enron's email network offers a unique opportunity to uncover communication patterns and power dynamics that influenced the company's operations leading up to and during the crisis. By examining this network, researchers and analysts can gain meaningful insights into several key areas:

- **Understanding Communication Patterns**: Analyzing how individuals within the network communicated reveals the roles they played in decision-making processes and the nature of their relationships. This helps identify key players who were central to discussions and decision-making, shedding light on how information was exchanged and decisions were made.
- **Identifying Influential Nodes**: Network analysis techniques help pinpoint influential individuals—nodes—that served as connectors within the network. These central figures often bridged different groups, facilitating the flow of information and accelerating communication. Recognizing these individuals is crucial for understanding the dynamics of influence and control within the organization.
- **Evaluating Community Structures**: Detecting sub-groups or communities within the network uncovers how different departments or teams were interconnected. This analysis reveals how information flowed between different parts of the organization and how collective behavior within these sub-groups contributed to broader organizational outcomes. Understanding these structures is key to understanding the internal dynamics that shaped Enron's rise and eventual collapse.

Objectives

The goal of this analysis is to gain a comprehensive understanding of the Enron email network by addressing the following objectives, aligned with the methods and tasks undertaken in the project:

1. **Identify Key Influencers through Centrality Analysis:**
   Using in-degree and out-degree, betweenness, closeness, and eigenvector centrality measures, we will identify influential individuals within the network. This analysis will pinpoint those with significant connections, bridging roles, and reach, indicating their importance in information flow.

2. **Examine Information Diffusion Pathways:**
   This analysis will map and visualize the primary pathways through which information spreads within the Enron network. The task will help identify key communicators, bottlenecks, and potential vulnerabilities in information dissemination, providing a clearer understanding of how information reached decision-makers, particularly during critical moments.

3. **Simulate Information Spread using Diffusion Models:**
   Simulations using the Independent Cascade Model (ICM) and Linear Threshold Model (LTM) will model how information propagates across the network, revealing how influential nodes and initial conditions affect the spread.

4. **Community Detection:**
   Applying the Greedy Modularity, Girvan-Newman, and Louvain algorithms, we will identify tightly-knit subgroups within the network, showing how these clusters interact and their role in organizational communication.

5. **Authority and Hub Analysis:**
   Authority and hub scores will be calculated to identify nodes that drive or connect to authoritative sources, enhancing understanding of the network's hierarchy and influence dynamics.

6. **K-Core, K-Club, and K-Clan Analysis:**
   These methods will identify strongly interconnected groups, highlighting core actors and their influence within the organization.

7. **Link Prediction:**
   Link prediction using the Jaccard, Assortativity, and Clustering Coefficients will help predict potential new connections, showing areas for possible network evolution.

8. **PageRank Calculation:**
   PageRank will identify nodes with broad influence across the network, regardless of direct connection frequency, emphasizing key players in overall information flow.

Methodology with Result

1. Random Sample of 500 Nodes for Visualization

To simplify visualization and analysis, a random sample of 500 nodes was selected from the full Enron network of over 36,000 nodes. Sampling allowed for a manageable view of network structure while preserving key characteristics, such as node connectivity and clustering. This subset offered a clearer representation of clusters, central nodes, and overall interaction patterns, enabling effective exploration of core network metrics in a simplified context.

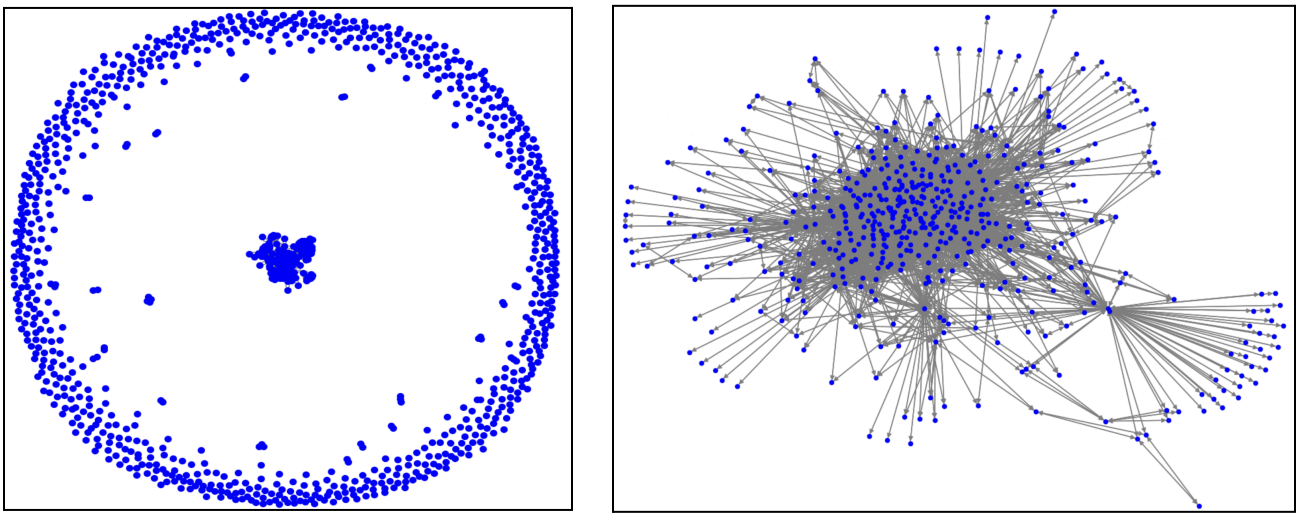

Figure - 1 (a) 500 sampled nodes from the Enron network (no edges shown).
(b) 500 sampled nodes with edges, showing connectivity in the Enron network.

2. Identify Key Influencers through Centrality Analysis

Centrality measures, including in-degree and out-degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality, will be used to identify influential individuals within the network. In-degree and out-degree centrality will highlight individuals who have the most incoming and outgoing communication, respectively. Betweenness centrality will reveal individuals who act as bridges between different parts of the network, facilitating information flow. Closeness centrality will measure how efficiently individuals can spread information, while eigenvector centrality will identify those whose influence is amplified by their connections to other influential nodes.

2.1. In-Out Degree Centrality:

Results: The analysis of in-degree and out-degree centrality shows that certain nodes (individuals) have the highest degree centrality, indicating they held the most direct connections in the network. High in-degree suggests individuals frequently received communications, while high out-degree indicates they often initiated contact with others. These nodes were likely highly active in communication, participating in or managing multiple projects or discussions.

Significance: Nodes with high degree centrality play crucial roles in an organization's daily operations. Acting as connectors, they facilitate cross-departmental exchanges. In Enron's network, these individuals could represent executives, team leads, or administrators who coordinated communications across various parts of the company.

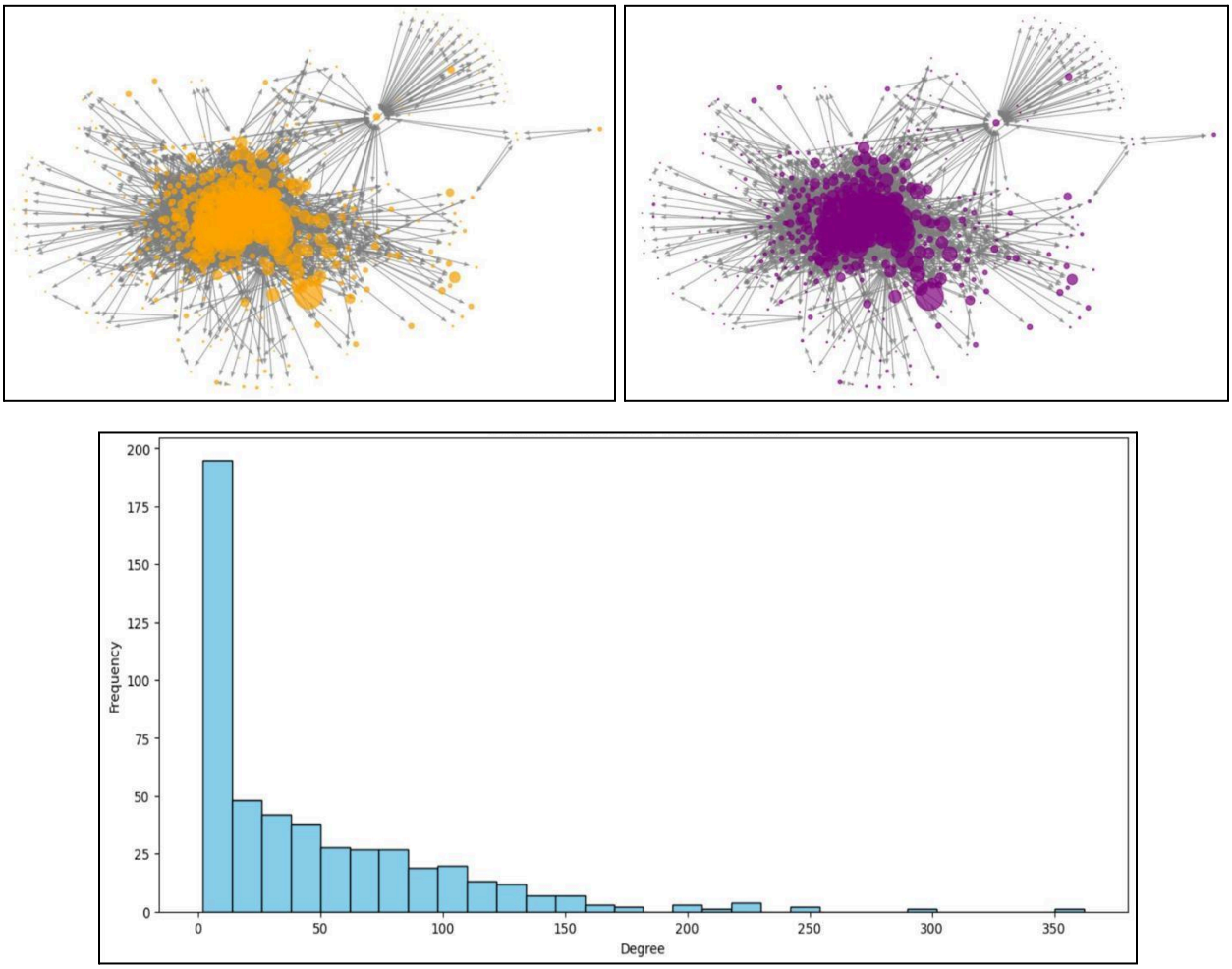

Figure - 2.1 (a) In-Degree Centrality Distribution
(b) Out-Degree Centrality Distribution
(c) Degree Distribution

2.2. Betweenness Centrality:

Results: Graphs highlight individuals with high betweenness centrality, representing those who act as crucial bridges in the network, facilitating communication between disparate groups. These individuals can be thought of as gatekeepers who influence the flow of information and ensure that information reaches parts of the network that are not directly connected.

Significance: High betweenness centrality often points to strategic roles such as project managers or department heads. These individuals may not be the most connected (high degree centrality), but their positions allow them to control information flow and coordinate between groups. In Enron, such roles might have included legal advisors or senior managers involved in cross-departmental coordination.

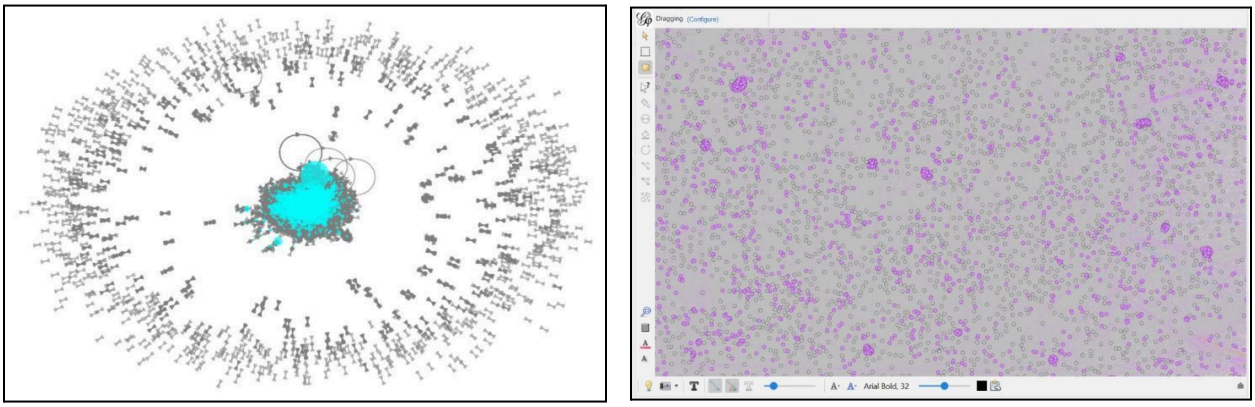

Figure - 2.2 (a) Betweenness Centrality (Jupyter Notebook)
(b) Betweenness Centrality (Gephi)

2.3. Closeness Centrality:

Results: Individuals with the highest closeness centrality are identified, showing those who can reach others in the network with the fewest number of steps. High closeness centrality means an individual is well-positioned to spread information quickly throughout the network.

Significance: High closeness centrality nodes are often efficient communicators who can rapidly disseminate critical information. They play essential roles in urgent or crisis situations where fast information flow is necessary. In the context of Enron, such nodes could be individuals responsible for compliance and rapid dissemination of policy updates.

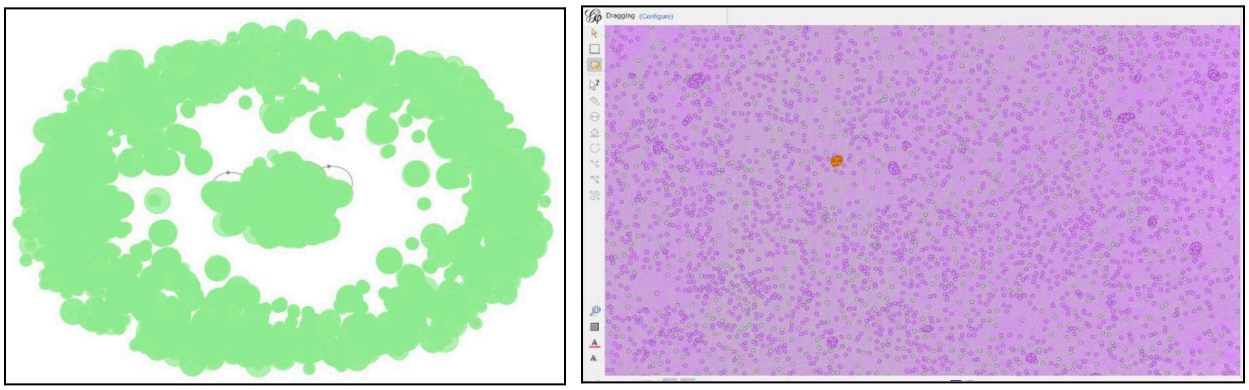

Figure - 2.3 (a) Closeness Centrality (Jupyter Notebook)
(b) Closeness Centrality (Gephi)

2.4. Eigenvector Centrality:

Results: The analysis shows nodes with high eigenvector centrality, suggesting that these individuals are influential not only because of their direct connections but also because they are connected to other influential people.

Significance: This measure identifies individuals whose influence is magnified by the strength of their connections. In Enron's network, these nodes could include high-ranking executives or key decision-makers who interacted frequently with other influential individuals. Such individuals likely had substantial sway in decision-making and strategic planning.

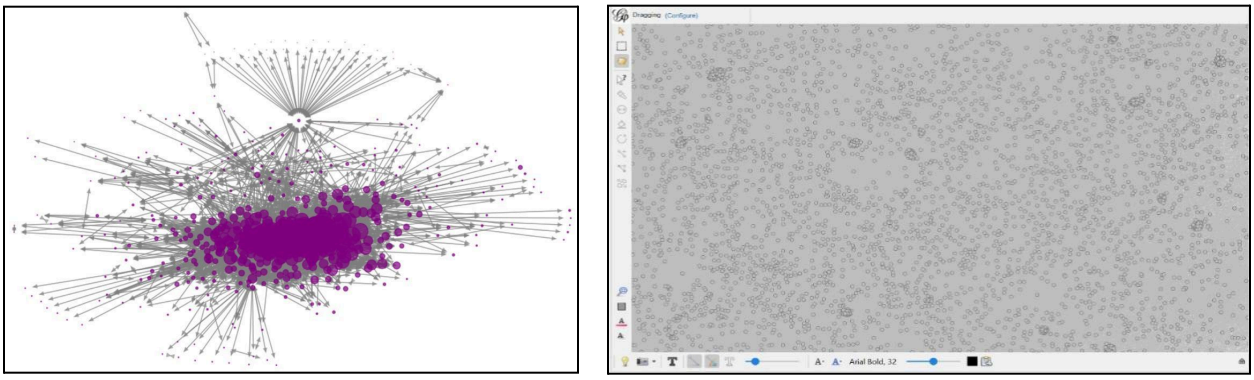

Figure - 2.4 (a) Eigenvector Centrality (Jupyter Notebook)
(b) Eigenvector Centrality (Gephi)

3. PageRank

Results: The PageRank analysis highlights individuals with the highest-ranking scores, identifying nodes that hold significant influence based on the recursive notion of importance. Nodes with high PageRank are not necessarily those with the most connections but those that are connected to other highly-ranked nodes, indicating their central role within influential communication pathways.

Significance: High PageRank suggests individuals who are crucial for maintaining the flow of information, even if they do not have the most direct interactions. These nodes may be strategic players who contribute to or receive information from well-connected individuals. In Enron, such nodes could be senior executives or key advisors who were central in organizational decision-making and had a broad, indirect influence over communications across departments. Their presence often ensures that information reaches important areas within the company and affects significant strategic outcomes.

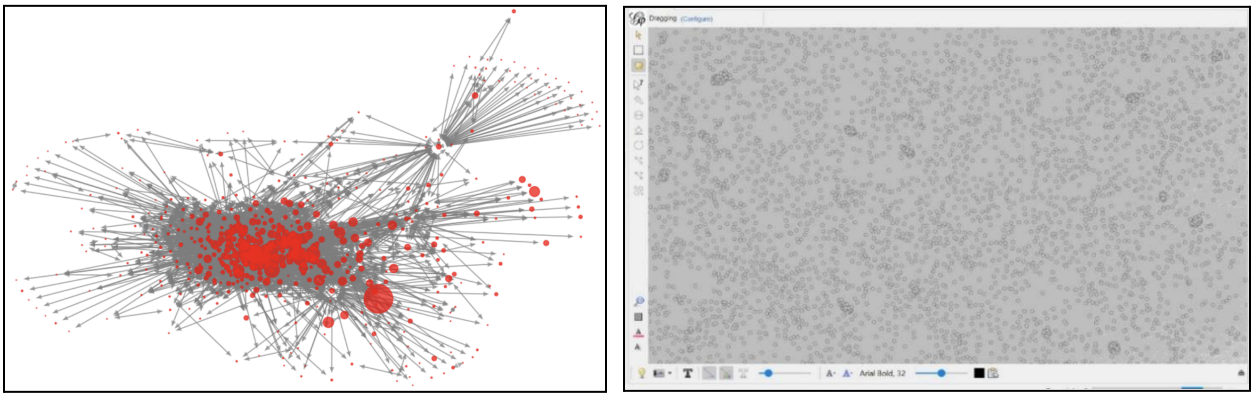

Figure - 3 (a) PageRank (Jupyter Notebook)
(b) PageRank (Gephi)

4. Community Detection Techniques

In the Enron email network, community detection was performed to identify tightly connected groups, which can reveal underlying structures and relationships within the organization. Three algorithms were applied to effectively uncover these communities:

4.1. Greedy Modularity Algorithm

Results: Identified communities with dense internal connections, highlighting groups that likely correspond to functional teams or departments with frequent internal communication.

Significance: Greedy Modularity clusters suggest groups that operate closely, likely working on specific projects or within the same department.

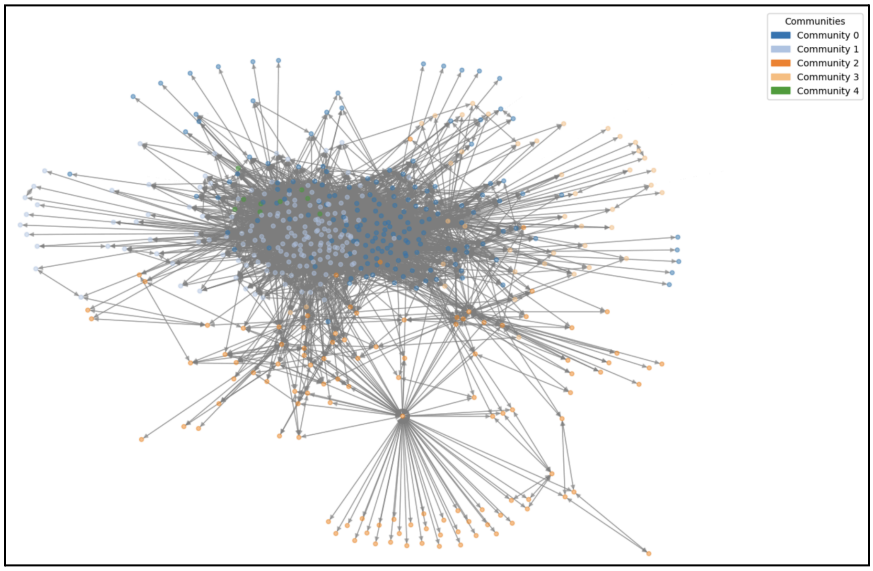

Figure - 4.1 Community Structure Detected Using Greedy Modularity Algorithm

4.2. Girvan-Newman Algorithm

Results: It uncovered clear boundaries between communities by removing high-betweenness edges, effectively separating clusters that may represent different divisions or roles within the organization.

Significance: Communities help to distinguish key divisions, showing where departments or teams are more separated by their communication patterns.

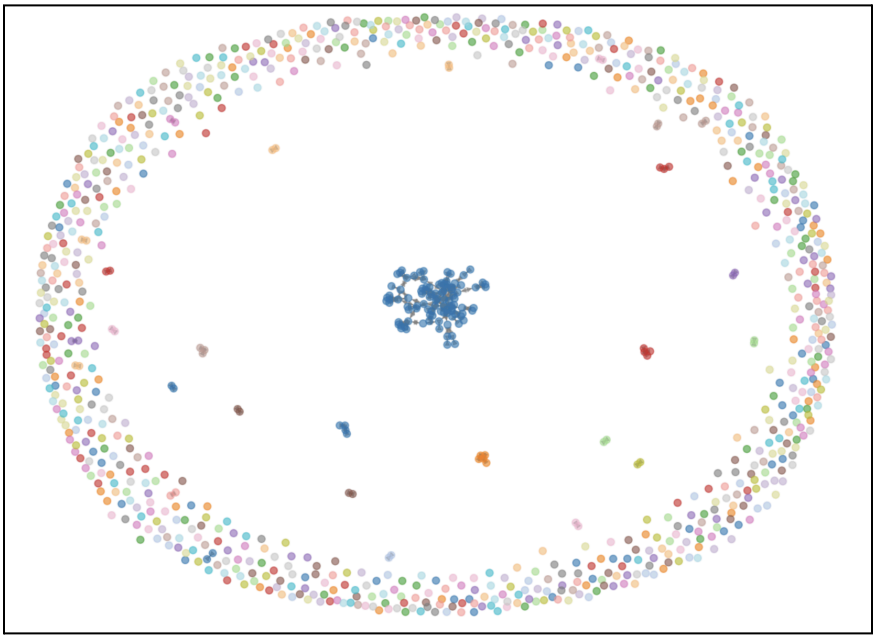

Figure - 4.2 Community Structure Detected Using Girvan-Newman Algorithm

4.3. Louvain Method Algorithm

Results: It revealed a hierarchical structure, detecting both large and nested sub-communities, showing both broad organizational divisions and more specific subgroups within them.

Significance: Louvain Method reveals hierarchical relationships, indicating both broad and niche subgroups, which could represent teams nested within larger departments or collaborative task forces.

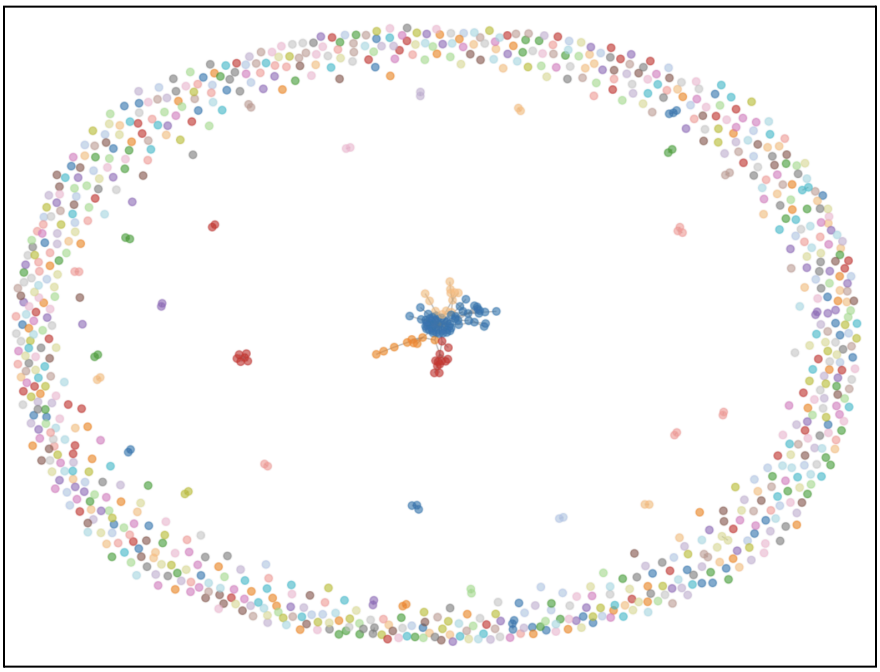

Figure - 4.3 Community Structure Detected Using Louvain Method Algorithm

5. Authority and Hub Analysis

Hub and authority scores are centrality measures that reflect the roles of nodes in the network from different perspectives. In this analysis, we calculated both the Hub and Authority scores for each node to identify key influencers in terms of their connectivity and influence within the network.

5.1. Authority

Score measures how important a node is based on its connections from other highly influential nodes. Nodes with high authority scores are often those who are frequently linked to by other important nodes, indicating their role as trusted or highly referenced sources of information.

Results: The nodes with the highest authority scores (e.g., Node 76, Node 56) are likely those whose communication or actions were frequently referenced or relied upon by other influential individuals in the organization. This suggests that these individuals were critical sources of information or had pivotal roles in decision-making processes.

Significance: High authority nodes are likely to have been individuals whose opinions, actions, or decisions were highly regarded or frequently referenced across the network. They were pivotal in disseminating critical information or setting organizational direction. In the context of Enron, these could have been top executives or department heads whose input shaped the course of events.

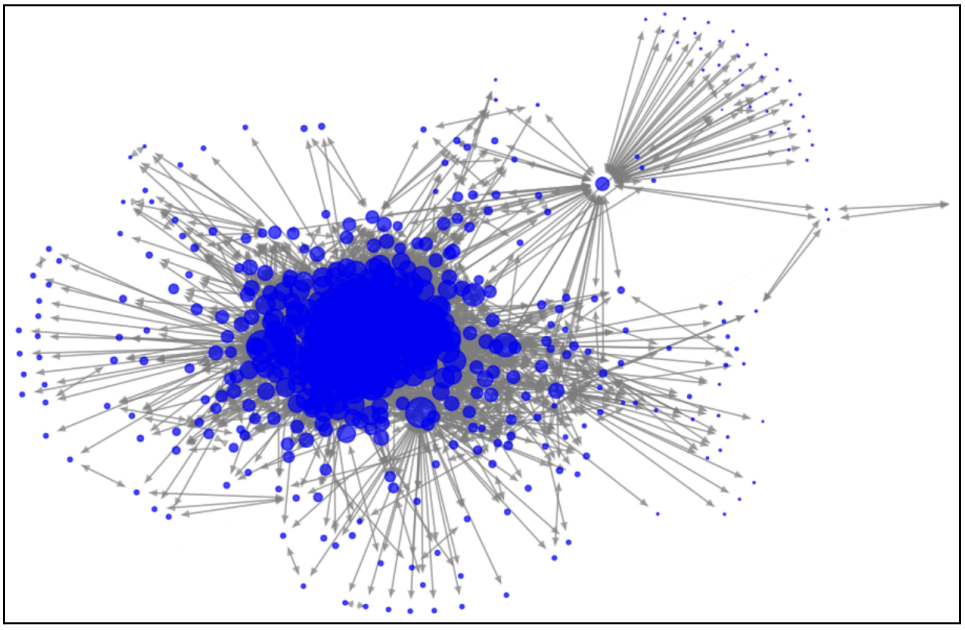

Figure - 5.1 Authority Score in the Enron Email Network

5.2. **Hub** score, on the other hand, identifies nodes that serve as key connectors in the network. Nodes with high hub scores tend to have many outgoing links, connecting to several important nodes or information sources.

Results: The nodes with the highest hub scores (e.g., Node 76, Node 56) played a crucial role in connecting various influential individuals, either by initiating communication or facilitating collaboration. These nodes acted as key intermediaries, ensuring information flow across different parts of the network.

Significance: High hub nodes serve as central communication points, connecting various subgroups within the network. Their role as key connectors suggests that they were actively involved in managing or coordinating multiple channels of communication. Individuals could have been administrative leaders, project managers or team leaders responsible for bridging gaps between departments or teams.

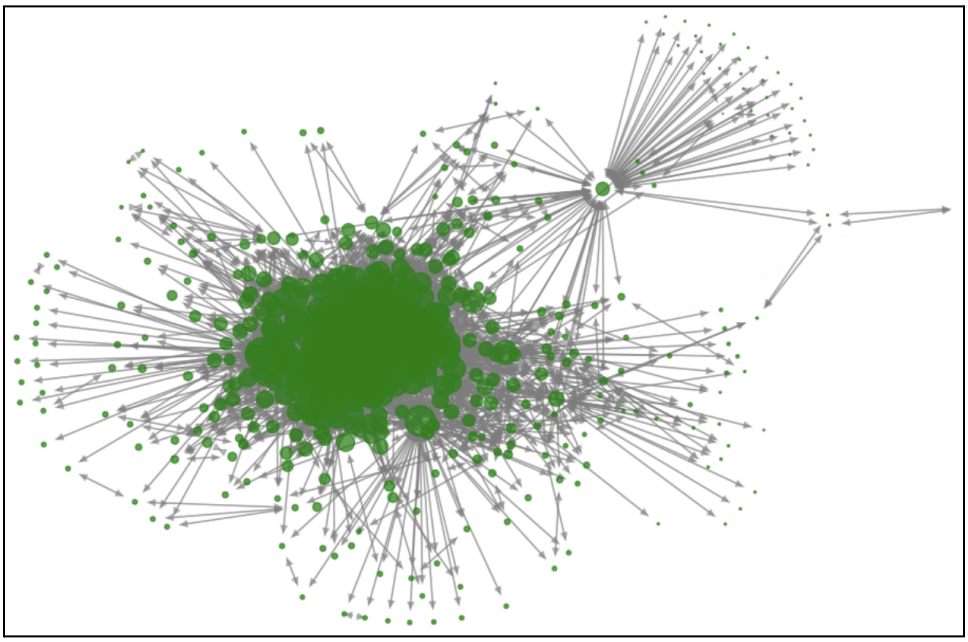

Figure - 5.2 Hub Score in the Enron Email Network

6. Predicting Potential Connections in the Enron Network

Link prediction aims to forecast potential new connections in a network by analyzing patterns in existing relationships. In the Enron email network, we used three different metrics—Jaccard Coefficient, Assortativity Coefficient, and Clustering Coefficient—to predict potential links that could emerge, shedding light on areas where the network may evolve or expand.

6.1. Jaccard Coefficient

This metric measures the similarity between two nodes based on the number of common neighbors they share. A higher Jaccard score between two nodes suggests a stronger likelihood that they will form a link in the future. It helps identify pairs of nodes that, although not directly connected, have a strong connection through mutual contacts.

Results: The Jaccard Coefficient analysis identified several predicted links with a perfect score of 1.0000, indicating that these pairs of nodes share all their neighbors, making them highly likely to form a future connection. This suggests a strong relationship between the nodes, based on their common interactions with other parts of the network. For example, the link between (26725, 3281), with a Jaccard Coefficient of 1.0000, indicates that these nodes are closely connected through shared neighbors, and a new connection between them is expected.

Significance: Analysis highlights potential new connections by identifying pairs of nodes with shared neighbors, suggesting strong relationships that could lead to future interactions. A score of **1.0000** indicates a high likelihood of these nodes forming connections, reinforcing the idea of emerging collaborations or communication channels within the network. This prediction is valuable for understanding the network's evolution and anticipating areas where relationships may strengthen or new communication paths could emerge.

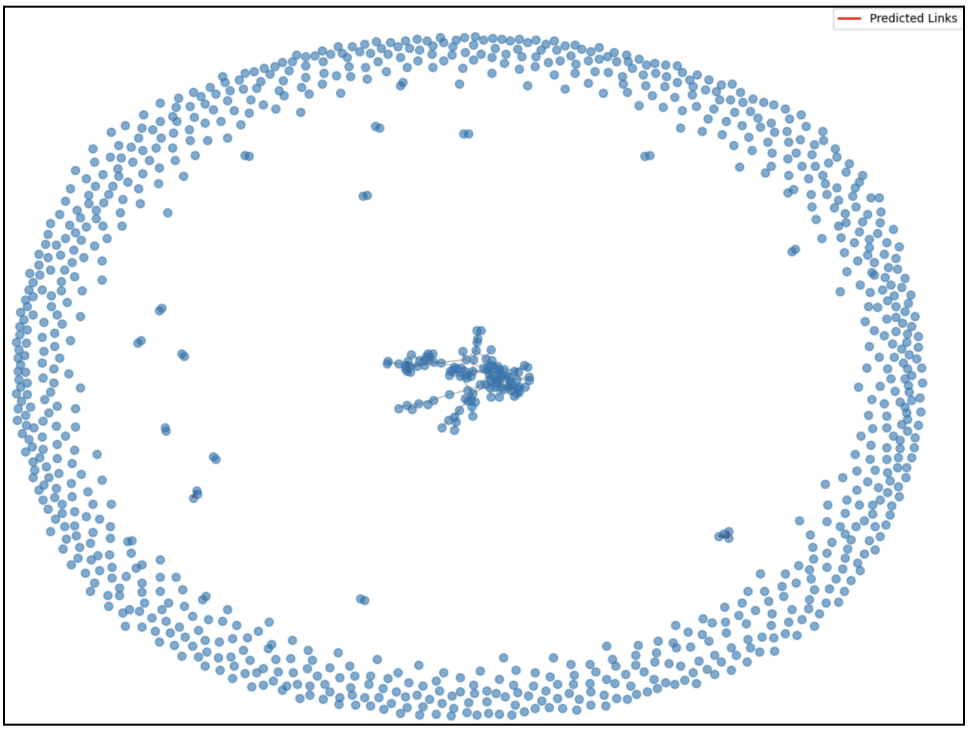

Figure - 6.1 Jaccard Coefficient Link Prediction

6.2. Assortativity Coefficient

Measures the tendency of nodes to connect with others having similar characteristics. A positive value indicates similar nodes connect, while a negative value suggests dissimilar nodes form links.

Results: The Degree Assortativity Coefficient of **-0.1108** indicates a slight negative correlation, meaning nodes with dissimilar degrees (e.g., high-degree nodes connected to low-degree nodes) are more likely to connect.

Significance: A negative value suggests the network connects influential nodes with less connected ones, potentially forming bridges between network segments, improving overall connectivity and facilitating information flow.

6.3. Clustering Coefficient

Measures the likelihood that two neighbors of a node are also connected to each other. A high clustering coefficient indicates that the network is highly modular, with groups of tightly connected nodes. Link prediction using this coefficient helps identify potential new edges that could strengthen existing communities or create new subgroups.

Results: The Average Clustering Coefficient of 0.497 indicates that, on average, nodes in the network tend to form tightly-knit groups, with about 50% of their neighbors being connected to each other.

Significance: A moderate clustering coefficient suggests that the network has a balanced structure, where nodes are often part of local clusters, facilitating efficient information flow within groups while maintaining connectivity across the broader network.

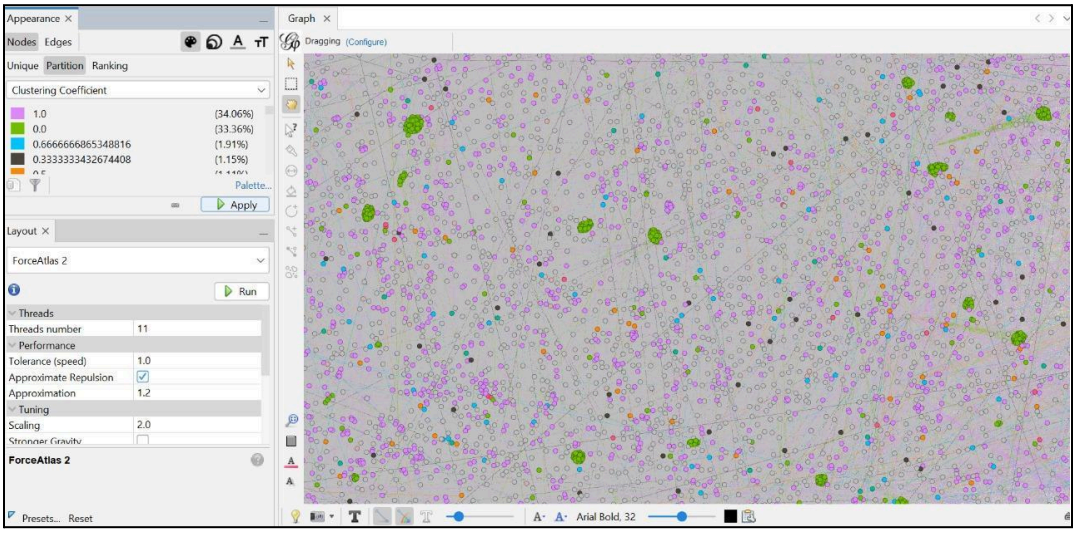

Figure - 6.3 Clustering Coefficient Link Prediction

7. Modularity

Measures how well a network can be divided into distinct communities, where nodes within a community are more connected to each other than to those in other communities. A higher modularity value indicates stronger, well-defined sub-groups within the network.

Results: Modularity analysis measures the network's division into distinct communities, where higher values indicate stronger internal cohesion within sub-groups and weaker connections between them. In the case of Enron, these communities likely represent specific departments or teams with frequent internal communication but limited interaction with other groups.

Significance: High modularity suggests the presence of well-defined sub-groups that operate semi-independently. This structure reflects how Enron's internal teams or departments functioned, highlighting potential barriers to communication between them. Understanding modularity can reveal collaboration patterns, identify information hubs, and inform strategies for improving cross-departmental communication.

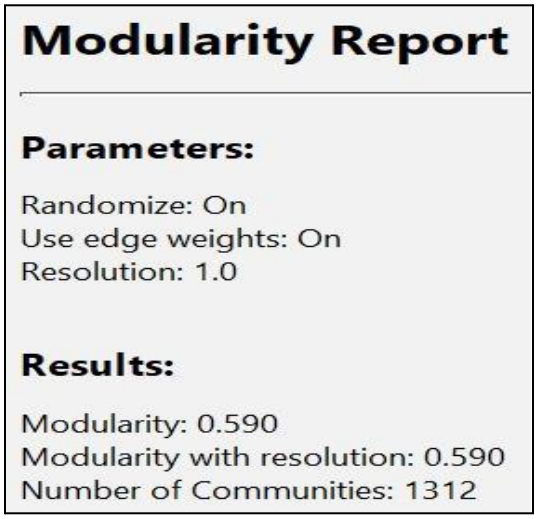

Figure - 7 Modularity Analysis Report

8. K-Core, K-Club, and K-Clan Analysis

K-Core, K-Club, and K-Clan analyses help reveal the most influential subgroups and key players within the organization. By identifying these core actors and their interactions, organizations can gain insights into how information is shared, where bottlenecks may occur, and which individuals or teams drive communication across the network. Understanding these structures is critical for improving collaboration, decision-making, and information flow.

8.1. K-Core Analysis

This method identifies the largest subgraph where each node is connected to at least k other nodes. It helps pinpoint core actors who are highly involved in the network, contributing significantly to the flow of information and decision-making.

Results: The k-core analysis (k=3) identifies a subgraph with **21,309 nodes** and **166,039 edges**, indicating a large, well-connected core of the network. These nodes have at least 3 connections to other nodes within the core, highlighting a group of individuals deeply embedded in the communication flow of the network.

Significance: The k-core subgraph reveals the core group of highly connected individuals within the organization. These individuals are crucial for the network's cohesion and information flow. The presence of a large k-core suggests a robust central structure, where communication is dense, and these nodes likely play pivotal roles in decision-making and influencing others.

8.2. K-Club Analysis

This technique focuses on groups of nodes that have a high degree of mutual connectivity, but with a specific threshold (k). It identifies tightly-knit subgroups within the larger network, indicating individuals who share strong relationships and influence within these sub-clusters.

Results: The approximate k-club analysis shows that the ego networks for nodes 0, 2, and 8 each consist of **632 nodes** and **6,857 edges**, demonstrating strong local connections within these ego networks. These nodes are at the center of tightly connected subgroups, indicating that they hold influential positions within their local communities.

Significance: K-clubs identify smaller, tightly-knit groups within the larger network. The strong connectivity within these ego networks suggests that individuals within these groups likely serve as central communicators or decision-makers. Understanding these communities can provide insights into the micro-dynamics of information flow within specific segments of the organization.

8.3. K-Clan Analysis

A more relaxed version of the K-Club, K-Clan detects groups of nodes with a minimum number of connections but allows for some flexibility in the connectivity threshold. This method reveals clusters with significant internal cohesion, helping to identify subgroups that may operate semi-independently but still play a role in broader communication dynamics.

Results: The k-clan analysis on a sample reveals three distinct communities:

- **Community 1:** 15 nodes, 16 edges
- **Community 2:** 10 nodes, 9 edges
- **Community 3:** 10 nodes, 9 edges

These communities are characterized by fewer nodes and edges, suggesting they are smaller, loosely connected subgroups within the broader network.

Significance: K-clans highlight smaller, less tightly connected subgroups within the network. These communities may represent specialized teams or departments that, while not as interconnected as the k-core, still play a role in maintaining communication and collaboration.

9. Simulate Information Spread using Diffusion Models

Simulating information spread through the Independent Cascade Model (ICM) and Linear Threshold Model (LTM) offers valuable insights into how information may propagate across the Enron email network under varying conditions and diffusion mechanisms.

9.1. The Independent Cascade Model (ICM)

ICM identifies key influencers who initiate information cascades. Nodes are categorized as Susceptible, Infected, or Recovered, showing the progression of information spread. High-centrality nodes act as major spreaders, while bottlenecks—nodes linking different groups—control the flow of information. Identifying these influencers and bottlenecks helps understand how information moves through the network and where communication might be stalled.

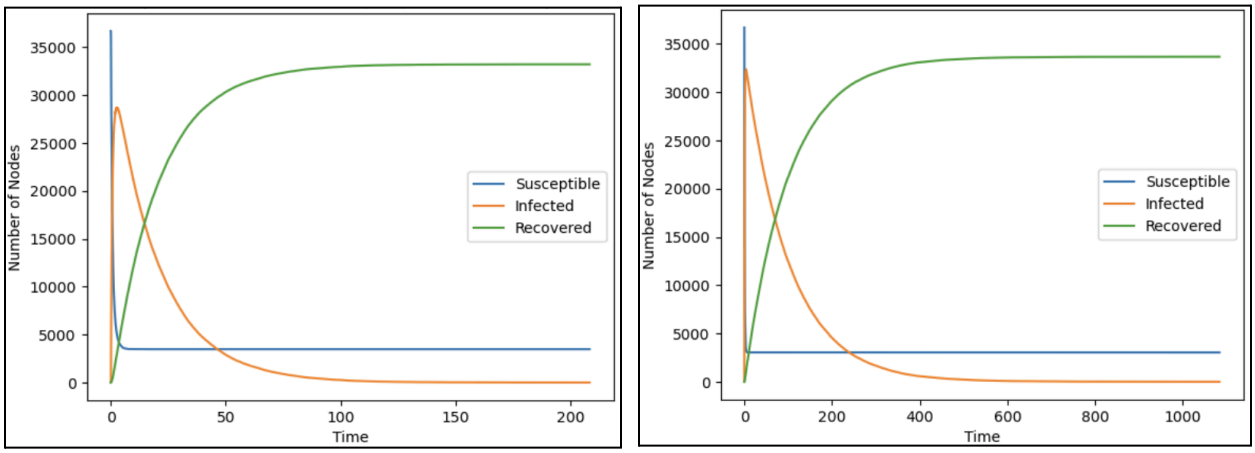

Figure - 9.1 (a) Information Spread using the Independent Cascade Model (ICM)
(b) Information Spread using degree centrality of the Independent Cascade Model

9.2. The Linear Threshold Model (LTM)

LTM models peer influence, where nodes adopt information only when influenced by enough neighbors. It identifies critical thresholds and the minimum number of adopters needed to trigger widespread diffusion. LTM highlights individuals who are more resistant to influence and reveals how information circulates within subgroups, showing how peer pressure drives communication.

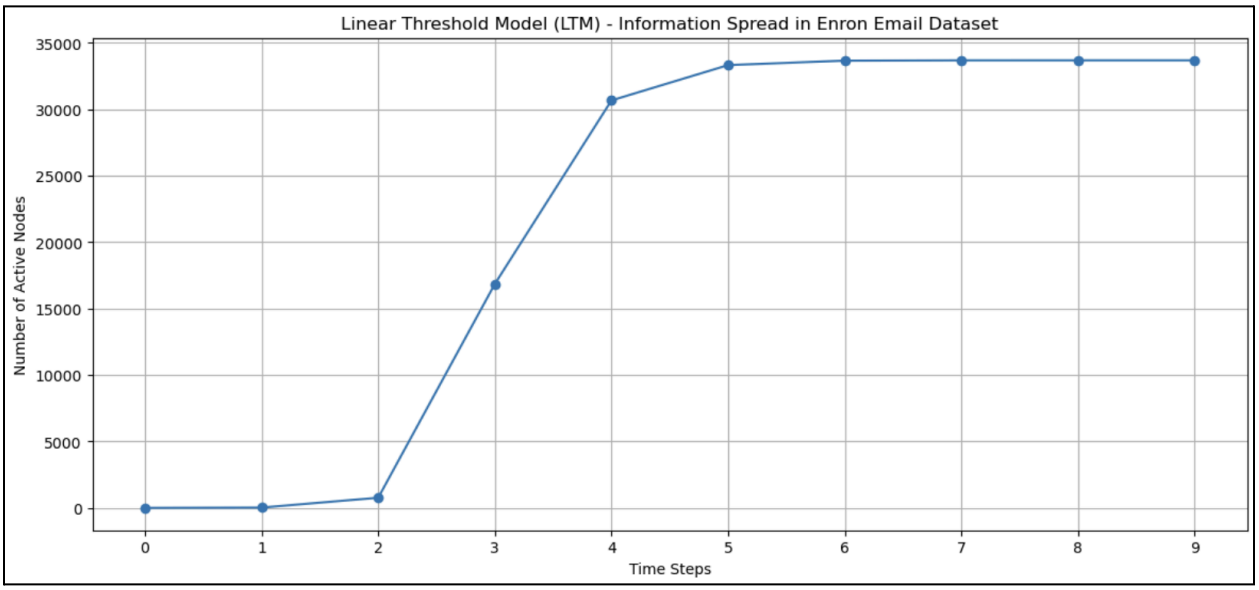

Figure - 9.1 Information Spread using the Linear Threshold Model (LTM)

9.3. Comparison to Observations

The simulated diffusion rates show a strong alignment with real-world observations, such as the timeline of information spread during the Enron crisis. The observed cascade included **36,692** active nodes, while the simulated cascade captured **33,729** active nodes. This high alignment is further supported by a precision of **1.0000**, meaning that every node predicted as active by the model was indeed active in the real network. The recall of **0.9192** indicates that the model successfully captured about **91.92%** of the actual active nodes, with a small portion (around 8%) not captured. The F1 score of **0.9579** reflects this balance of high precision and recall, showing that the Linear Threshold Model effectively replicated the real spread of information.

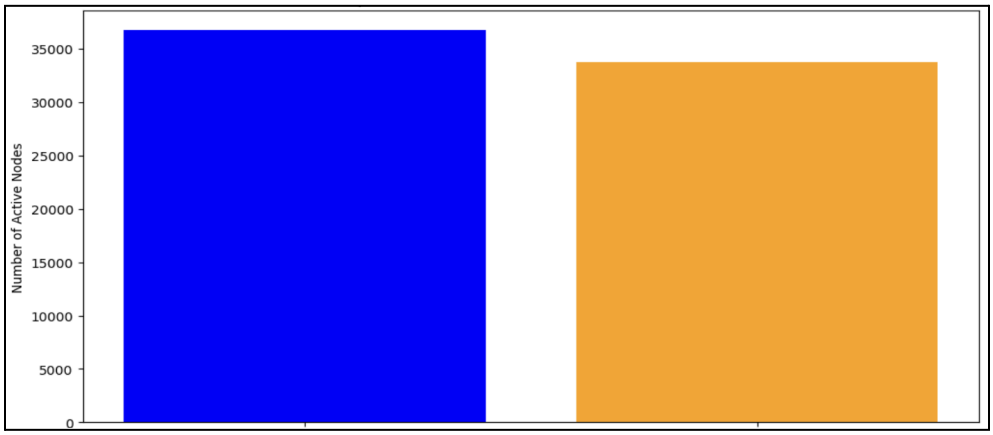

Figure - 9.3 Comparison of Observed V/S Simulated Active Node
(a) Observed Active Nodes [Blue]
(b) Simulated Active Node [Orange]

Enron Network Analysis: Real-World Insights and Implications

1. Key Influencers and Communication Patterns

Network analysis identified central figures within Enron who likely held influential roles, mirroring real-world executives and decision-makers. High centrality scores suggest these nodes were pivotal in cross-departmental communication and information flow. The observed patterns align with documented connections, indicating that key nodes reinforced operational clusters and influenced decision-making. This insight underscores the value of network analysis for identifying core influencers and assessing communication efficiency in complex corporate environments.

2. Crisis Communication Dynamics

The network's structure may have influenced how quickly information about internal crises spread within Enron. The presence of well-connected nodes, or bottlenecks, suggests they played a role in managing or controlling the flow of sensitive information, which could have affected transparency and response times. This insight emphasizes how understanding network communication patterns can be essential for evaluating a company's preparedness and responsiveness to crises, particularly regarding transparency and information sharing.

Future Trends in Network Analysis: AI and Big Data

1. AI and Machine Learning for Network Efficiency

AI and machine learning are becoming transformative tools for studying information spread across networks. They provide powerful methods for predicting how information will diffuse and identifying influential nodes within complex datasets. For instance, neural networks and machine learning models can enhance recommendation systems, allowing for more personalized and efficient content delivery. In network analysis, this means companies can predict the most effective ways to disseminate information, thus optimizing engagement and minimizing delays.

2. Big Data for Strategic Communication

Big data offers a wealth of insights into user behaviors, preferences, and communication pathways within organizations. By analyzing these interactions, companies can better understand which channels are most effective, allowing for a tailored approach to information diffusion. For companies like Enron, such insights could provide data-driven approaches to enhance communication strategies, reduce informational bottlenecks, and foster greater inter-departmental synergy.

Conclusion

The analysis of the Enron email network sheds light on the dynamics of information flow and the roles of influential individuals within complex organizations. Centrality measures identified key figures and primary communication pathways, underscoring the significance of certain employees in shaping information dissemination. Understanding these patterns is essential for enhancing internal communication efficiency and decision-making processes.

Emerging technologies like AI and machine learning further enable organizations to analyze communication networks with precision. However, their use must balance ethical considerations, including data privacy and fairness, to maintain trust and accountability.

These findings have practical implications for optimizing communication strategies. Recognizing influential nodes and clear pathways allows organizations to improve internal connectivity, streamline processes, and support better decision-making. However, this study's insights are based solely on Enron's dataset. Future research could broaden its scope by exploring other sectors and longitudinal datasets to capture evolving diffusion patterns. Incorporating sentiment analysis and qualitative data could further enrich our understanding of information reception and responses across networks, enabling more adaptable and effective organizational communication strategies.

References

1. Enron Email Dataset. *Direct access to the dataset for social network analysis.*

2. Zhou, C., Lyu, M.R., & King, I. (2017). "A Survey on Information Diffusion in Online Social Networks: Models and Methods." *Information*, 8(4), 118. Available: https://www.mdpi.com/2078-2489/8/4/118

3. Sun, Y., Lin, M., & Wu, X. (2020). "A Unified Information Diffusion Model for Social Networks." *IEEE Transactions on Knowledge and Data Engineering*, DOI: 10.1109/TKDE.2020.3024218

4. Gephi. (n.d.). *Gephi: An Open Source Network Visualization and Exploration Software.* Available: https://gephi.org/