Analysis of the Enron Email Network Using Graph-Based Techniques

Michael Rabayda IV
GSAS Data Science
Fordham University
Bronx, NY, USA
mr129@fordham.edu

***Abstract*—This project aims to analyze the Enron email communication network to understand the dynamics between individuals within the organization. By representing the network as a graph, we will apply community detection, and graph analysis to derive insights into organizational roles, communication patterns, and relationship dynamics. The primary goal is to identify key personas and communities within Enron based on roles, communication metrics, and frequency of communication, as well as to explore how information flows throughout the organization. The analysis will be conducted using Python, with the Enron email dataset serving as the foundation for all graph-based insights.**

Keywords—Enron, Communication Analysis, Louvain, Graph, Persona

I. Introduction

The Enron email corpus, derived from the communications of a failed power company in Texas, provides a rich dataset for analyzing corporate interactions. The goal of this project is to use descriptive data mining methods to reveal communication patterns, personas, and relationships within a complex email network. By applying graph-based analysis techniques, such as community detection and graph traversal, we aim to generate hypotheses about how roles and communication metrics influence network structure and information flow. Our approach is intended to be exploratory, focusing on discovering relationships and dynamics within the organization. For example, the identification of Kenneth Lay as a central figure in the communication network suggests that these methods can effectively identify key personas and hierarchical structures. We will evaluate the effectiveness of our descriptive analysis by comparing our findings with other related studies that have used graph theoretical and spectral analysis techniques to understand organizational behavior. By doing so, we hope to demonstrate that graph-based analysis is a useful tool for hypothesizing about corporate interactions, identifying influential individuals, and mapping information pathways.

II. Background

Enron Corporation, once one of the largest energy companies in the United States, rose rapidly in the 1990s due to its aggressive pursuit of energy trading and deregulated energy markets. As the boom years ended, Enron faced increased competition in the energy-trading business, causing profits to shrink rapidly. In response, company executives began relying on dubious accounting practices, including "mark-to-market accounting," which allowed the company to project unrealized future gains as current income, thereby masking financial troubles. Additionally, Enron transferred troubled operations to special purpose entities (SPEs) to keep failing assets off the company's balance sheet, misleading investors about the true state of the company. Arthur Andersen, Enron's auditor and consultant, not only approved these misleading accounting practices but also faced scrutiny for shredding documents during investigations.

In February 2001, Jeff Skilling took over as Enron’s CEO, while founder Kenneth Lay remained as chairman. However, Skilling abruptly resigned in August, and Lay resumed the CEO role. During this period, Enron received an anonymous memo from vice president Sherron Watkins, warning about the Fastow partnerships and the potential accounting scandals. In October 2001, Enron announced a significant third-quarter loss of $638 million and a $1.2 billion reduction in shareholder equity, which triggered an investigation by the Securities and Exchange Commission (SEC). As details of the fraud emerged, Enron's stock price plummeted, dropping from a high of $90 per share to under $1. Enron's failed attempt to be acquired by Dynegy in November 2001 further accelerated its decline, ultimately leading to its bankruptcy filing on December 2, 2001.

Following Enron's collapse, many executives were charged with conspiracy and fraud, including CEO Jeffrey Skilling and CFO Andrew Fastow. Lay died before his sentencing, while Skilling and Fastow received prison sentences for their roles in the scandal. Arthur Andersen was also indicted and eventually lost its license to practice accounting. The Enron scandal highlighted significant shortcomings in financial regulation, prompting the enactment of the Sarbanes-Oxley Act in 2002, which aimed to increase corporate transparency and accountability. Enron's downfall ultimately resulted in significant losses for shareholders and employees, with many losing their life savings and retirement funds [5][6].

III. Experiment Methodology

The experiments conducted in this study utilized the Enron email dataset to explore the structure and dynamics of corporate communication through graph-based techniques. The methodology involved multiple stages, including graph creation, node filtering, centrality analysis, and community detection. These methods were selected to uncover communication patterns, identify key individuals, and evaluate the organizational structure within the dataset.

We constructed a directed graph where each node represented an individual and each edge represented an email sent from one individual to another. Metadata such as sender and recipient fields were extracted from the email dataset to define these relationships. To ensure efficiency and clarity in analysis, the dataset was preprocessed to remove rows with missing or incomplete data.

In-degree and out-degree centrality metrics were computed to identify the most influential nodes in terms of receiving and sending emails. High in-degree nodes indicated individuals who were central recipients of communication, whereas high out-degree nodes highlighted prolific email senders. These metrics were essential for detecting key figures and understanding the flow of information within the organization.

To identify clusters of closely connected individuals, we applied the Louvain method for community detection. This method optimizes modularity, a metric that measures the strength of division of a network into communities. Varying the degree threshold of nodes allowed us to explore different configurations of the network. Lower thresholds produced higher modularity scores but resulted in overly dense communities that were less interpretable. Conversely, higher thresholds reduced modularity but yielded more distinct and interpretable clusters. This trade-off between modularity and interpretability was a critical consideration in our methodology.

Filtering out nodes with small degrees helped reduce noise and highlighted the most significant nodes in the graph. By varying the degree threshold, we systematically evaluated how network structure and community detection results changed. This approach allowed us to identify the optimal balance between achieving high modularity and maintaining interpretability of the resulting communities. A threshold of 400 was ultimately chosen, as it produced meaningful clusters that aligned with known organizational structures within Enron while maintaining a reasonable modularity score.

The Louvain method for community detection is a heuristic algorithm designed to identify community structures within a network by optimizing modularity, a measure that quantifies the quality of a particular division of a network into communities. The method is highly efficient, making it suitable for large-scale networks like the Enron email dataset. Its iterative process consists of two main phases that are repeated until no further modularity improvement can be achieved.

In the first phase, each node is initially assigned to its own community. For each node, the algorithm evaluates the modularity gain that would result from moving the node to a neighboring community. This gain is calculated using the weights of edges connecting the node to the target community, the total weights of edges connecting to the node, and the resolution parameter, which influences the size of the resulting communities. If moving the node yields a positive modularity gain, the node is reassigned to the new community. If no gain is achieved, the node remains in its original community. This process continues for all nodes until no further positive modularity changes can be made.

In the second phase, the graph is reduced by collapsing nodes into their respective communities. Each community is treated as a single node, and edges between these new nodes are weighted by the sum of the weights of the edges in the original graph that connected nodes in the corresponding communities. This reduced graph is then subjected to the same process of modularity optimization in subsequent iterations.

By iteratively applying these two phases, the Louvain method constructs increasingly coarser representations of the network, identifying hierarchically nested communities. The process terminates when the modularity gain between iterations falls below a specified threshold, or a predefined maximum number of levels is reached[8].

The Louvain method was implemented using the community_louvain Python library, and the graph analyses were performed using NetworkX. For visualizations, we used a spring layout to plot nodes and edges, assigning community-based color coding to enhance interpretability. Other parameters, such as the number of iterations in the Louvain method, were set to defaults optimized for modularity calculation.

This methodology enabled a thorough exploration of the Enron email network, balancing computational efficiency, clarity, and interpretability. By combining degree filtering, centrality analysis, and community detection, we developed a comprehensive understanding of the network's structure and identified key areas for further investigation.

IV. Results

A. Dataset and Characteristics

The dataset used for this project is the Enron email corpus, which is freely available at https://www.cs.cmu.edu/~enron/ and is widely used for research purposes, especially in the fields of data mining, graph theory, and social network analysis. The dataset contains approximately 500,000 emails, which include various metadata such as sender, recipient, timestamp, and the content of emails. Given the raw nature of the dataset, it was initially organized as a collection of folders for individual employees, with subfolders reflecting personal email management practices. As such, significant cleaning and transformation were necessary to convert the data into a structured format suitable for analysis.

This email corpus is unique in that it provides insight into real corporate communication during a period of crisis, making it invaluable for understanding organizational behavior, communication dynamics, and hierarchical relationships. The features of the dataset include email frequency, role-based metadata, communication content, and temporal attributes—all of which help in understanding how information flowed within Enron before its collapse.

To facilitate analysis, the data had to be processed and standardized. Some of the specific features extracted include:

Email Metadata: Fields such as Message ID, Date, Time, Sender, Recipients, Subject, and Email Body were used to build nodes and edges in the network graph.

Metadata Fields: These fields (e.g., X-From, X-To, X-cc, X-bcc, X-Folder, X-Origin, X-FileName) provide additional routing and storage information, offering insights into how emails were managed and prioritized.

Time-Based Features: Communication frequency was calculated to identify patterns of activity and changes in communication intensity over time.

Relationship Features: Sender-recipient relationships and CC/BCC fields helped identify direct and indirect communication links, which were useful for understanding the network's structure.

B. Data Preparation

Data preparation began by downloading the maildir folder from the Enron dataset repository, which contained only raw email files. The email data was loaded into Pandas, and significant preprocessing was performed to transform it into a structured format suitable for analysis. The data was saved as separate CSV files, scaled based on the number of emails, resulting in three CSV files of varying sizes:

enron_emails_15k.csv (26 MB)

enron_emails_100k.csv (186 MB)

enron_emails_550k.csv (1.1 GB)

This approach allowed for incremental analysis, starting with smaller datasets to create manageable network graphs before scaling up to larger datasets. This ensured that the computational process could be monitored effectively without overwhelming system resources. Additionally, rows containing missing values (e.g., blank rows representing folder structures due to the recursive reading of folders) were filtered out to ensure data quality.

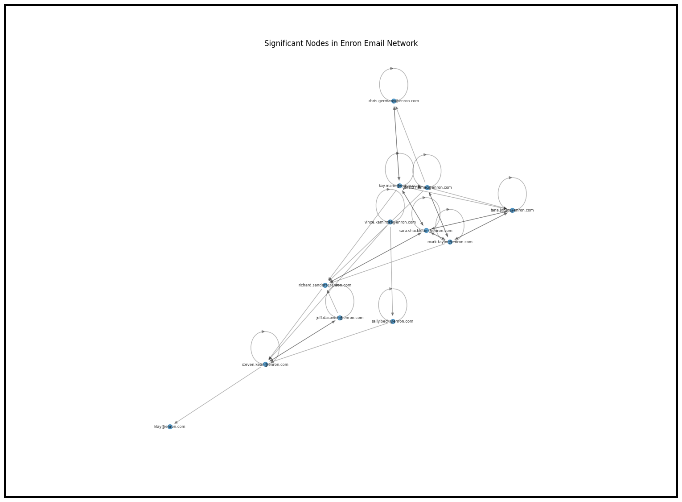
C. Initial Graph Creation & Enron Executive Analysis

Using the pre-processed CSV files, the first graph of the Enron email network was created to analyze communication patterns. The graph creation process involved:

Data: The CSV file enron_emails_550k.csv was loaded into Pandas for processing.

Directed Graph Creation: A directed graph (DiGraph) was constructed using NetworkX, where each node represented an individual, and each directed edge represented an email sent from one individual to another.

This initial graph analysis of the Enron email network revealed several key individuals and notable communication patterns. Kenneth Lay, the former CEO and chairman, emerged as a prominent node in the network, consistent with his role in the company's leadership. Other significant individuals, such as Richard Sanders, Vince Kaminski, and Kay Mann, were also identified as having substantial involvement in internal communications. The presence of self-referential loops in the network indicates instances where individuals sent emails to themselves. This behavior could represent automated system messages, personal notes for organization, or drafts being saved and reviewed. The dense interconnections observed among key figures suggest close communication ties, which may reflect collaborative efforts, hierarchical reporting structures, or attempts to manage the escalating issues within Enron. Understanding these communication dynamics provides valuable insight into the chain of decision-making and the flow of information during Enron's operational crisis.

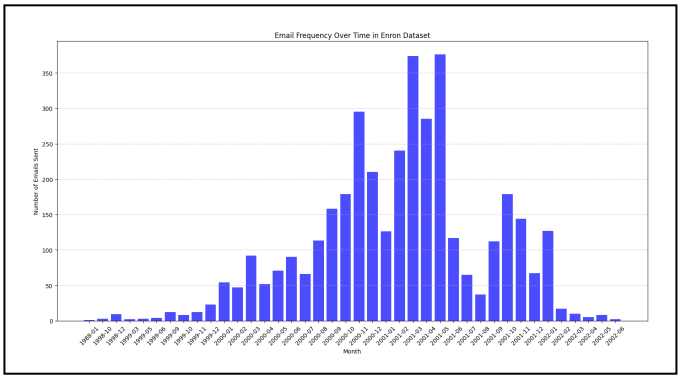


1. Significant Addresses Identified (*Significant Node in Enron Email Network*)

To further validate the findings of the initial graph analysis, we referred to the publicly available Enron executive board website to manually match several known executive names to email addresses within the dataset[7]. This manual cross-referencing allowed us to confirm the presence of key executives, such as Kenneth Lay and other influential individuals, within the communication network. By associating these names with specific email nodes, we were able to determine whether our algorithms successfully identified significant communication patterns involving high-ranking personnel. This step also provided additional context for interpreting the network's structure, as the involvement of these individuals often aligned with periods of high activity or critical organizational changes.

D. Email Frequency

After analyzing the initial network graph to identify key players and significant relationships, we sought to further understand the temporal dynamics of communication within Enron. Examining the frequency of emails over time provided insight into periods of heightened activity, which could indicate critical moments of decision-making, coordination, or crisis response within the company. By understanding when email traffic spiked, we could correlate these peaks with known events or internal shifts, thereby highlighting times of increased pressure or organizational shifts that may warrant deeper investigation. Next, we took a closer look at email frequency over time to identify specific periods of high communication traffic, which could provide valuable context



2. Periods of High Communication Traffic(*Email Frequency Over Time in Enron Dataset*)

for understanding the roles and influence of key individuals during pivotal moments in Enron's collapse.

The email frequency graph illustrates the number of emails sent over time within the Enron dataset. The peak email activity occurred between late 2000 and early 2001, reflecting a surge in internal communication during what was likely a turbulent period for the company. This timeframe coincides with Enron's financial troubles and the unfolding of various crises, including scrutiny over their accounting practices and the eventual collapse. The significant spike in email volume may indicate an increased need for coordination, crisis management, or damage control among employees and executives. After early 2001, the frequency of emails began to decline, aligning with Enron's downward spiral and eventual bankruptcy filing in December 2001.

E. Community Detection

The community detection analysis revealed distinct clusters of individuals within the Enron email network, highlighting significant communication patterns and structures. Below are the key findings from the analysis:

<u>Top 10 Nodes by In-Degree</u>

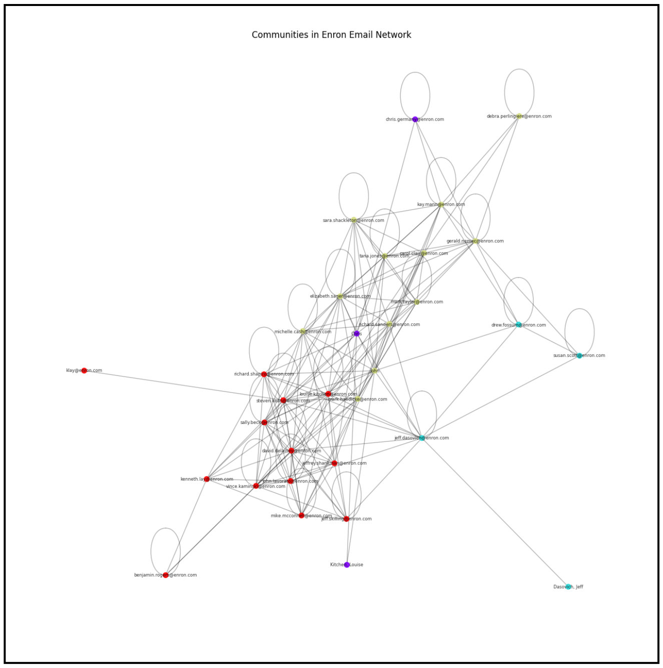
These nodes represent individuals who **received** the most emails within the network:

klay@enron.com	(1293)
kenneth.lay@enron.com	(613)
jeff.skilling@enron.com	(596)
John	(518)
sara.shackleton@enron.com	(467)
louise.kitchen@enron.com	(452)
jeff.dasovich@enron.com	(424)
sally.beck@enron.com	(411)
tana.jones@enron.com	(409)
Chris	(405)

<u>Top 10 Nodes by Out-Degree</u>

These nodes represent individuals who **sent** the most emails within the network:

jeff.dasovich@enron.com	(1252)
sara.shackleton@enron.com	(1112)
vince.kaminski@enron.com	(1014)
tana.jones@enron.com	(905)
kay.mann@enron.com	(860)
mark.taylor@enron.com	(747)
richard.sanders@enron.com	(726)
chris.germany@enron.com	(702)
gerald.nemec@enron.com	(690)
steven.kean@enron.com	(670)



3. Periods of High Communication Traffic(*Email Frequency Over Time in Enron Dataset*)

The Louvain method identified several distinct communities based on a degree threshold of 400. Below are the most notable communities. The modularity score for the community detection was 0.31, reflecting a moderate level of community structure in the filtered network.

Community 3 (13 members):
Key members include Kenneth Lay (klay@enron.com, kenneth.lay@enron.com), Jeff Skilling (jeff.skilling@enron.com), and Vince Kaminski (vince.kaminski@enron.com). This group likely represents central decision-makers and executives within the organization. Other members include Louise Kitchen, Sally Beck, and Steven Kean, all of whom were heavily involved in financial and operational decision-making.

Community 1 (4 members):
This smaller cluster includes Jeff Dasovich (jeff.dasovich@enron.com), Drew Fossum, and Susan Scott, individuals known for their work in regulatory and legal matters. The composition suggests this group may have focused on external relations and compliance.

Community 2 (12 members):
Members such as Tana Jones, Sara Shackleton, and Mark Taylor suggest this community was likely involved in administrative and legal operations, possibly including contract management. The presence of individuals like Kay Mann and Richard Sanders further supports this interpretation.

Community 0 (3 members):
A smaller group consisting of Louise Kitchen, Chris Germany, and Chris. This cluster may represent a highly specific subgroup within the organization, potentially focused on operational or specialized tasks.

(*Self-referential nodes are marked by automated or draft email workflows.)

The community detection analysis using the Louvain method offered valuable insights into the communication structures within the Enron email network. By setting a degree threshold of 400, we filtered out less active nodes, focusing on the most significant communicators. This choice was deliberate, as lower thresholds, such as 100 or 150, resulted in higher modularity scores—up to 0.43 in some cases—but created overly dense communities with too many members. While these configurations might have captured tighter statistical cohesiveness, they made the results less interpretable. By using a threshold of 400, we prioritized practical insights over modularity, achieving a balanced modularity score of 0.31 while highlighting more actionable clusters of communication. This trade-off was crucial in identifying meaningful groupings that align with Enron's organizational structure and functions.

Key nodes in the network, such as Kenneth Lay, Jeff Skilling, and other high-ranking executives, featured prominently in several communities, underscoring the effectiveness of this approach. For instance, Community 3 included notable individuals like Louise Kitchen, Sally Beck, and Vince Kaminski, who were central to Enron's operations and heavily involved in managing the company's financial and trading activities. The presence of both high in-degree and out-degree nodes in this community indicates its role as a hub for decision-making and the flow of critical information. Similarly, Community 1 included figures such as Jeff Dasovich, Drew Fossum, and Susan Scott, who were heavily involved in regulatory and legal matters, suggesting that this cluster focused on compliance, external relations, or legal strategy during the company's crisis.

Another significant cluster, Community 2, consisted of individuals such as Tana Jones, Sara Shackleton, and Mark Taylor. These individuals were known for their roles in handling contracts and sensitive documentation, pointing to this community's role in legal and administrative operations. The clustering of individuals with similar roles in these communities lends credibility to the Louvain algorithm's outputs, which captured the functional segmentation within Enron. Additionally, the detection of self-referential nodes (e.g., Tana Jones and Sally Beck) within these groups provides further insight into email usage patterns, suggesting automated systems, draft management, or other specialized workflows that reflect their roles within the organization.

By manually cross-referencing these detected communities with known executive names from the Enron board, we validated the algorithm's ability to identify meaningful clusters. For example, Kenneth Lay and Jeff Skilling, central figures in Enron's leadership and decision-making processes, were consistently identified within key communities. This alignment between algorithmic outputs and real-world organizational roles demonstrates the practical utility of combining degree thresholds with community detection methods for exploratory network analysis. The insights gained from these clusters provide a clearer picture of Enron's internal structure and how information flowed during pivotal periods.

Conclusion

The descriptive analysis of the Enron email network using graph-based techniques has yielded significant insights into the organizational communication patterns, influential individuals, and structural dynamics of the company during its period of crisis. By employing methods such as frequency analysis, in-degree and out-degree rankings, and Louvain community detection, we uncovered both the macro-level structure of communication flows and the micro-level dynamics of specific individuals and groups. These findings highlight the potential of network-based approaches to inform hypotheses about organizational behavior and decision-making processes.

Key figures such as Kenneth Lay, Jeff Skilling, and other senior executives were consistently identified as central nodes within the network, reflecting their roles as pivotal decision-makers and conduits of information. The analysis of email frequency over time suggested distinct periods of heightened communication, likely corresponding to critical organizational events or crises. These periods could support hypotheses about how internal communication intensifies during times of stress and how leadership coordinates efforts in response to external pressures.

The community detection analysis further enriched our understanding of Enron’s internal structure by identifying distinct clusters of communication that align with known functional teams. For example, Community 3 represented high-level executives and key decision-makers, suggesting a tightly knit group responsible for strategic operations. Meanwhile, other communities, such as Community 1 and Community 2, reflected regulatory, legal, and administrative operations, shedding light on how these departments interacted during the company’s decline. These findings could support hypotheses regarding the roles of different functional groups in managing crises, such as regulatory compliance, legal challenges, and operational adjustments.

Despite the descriptive nature of this analysis, several promising directions for future research emerge. One avenue could involve correlating communication spikes with specific known events, such as SEC investigations or major announcements, to better understand how communication patterns respond to external shocks. Another direction could involve temporal community detection to examine how clusters evolved over time, providing insight into shifts in organizational roles or priorities during different phases of the crisis. Additionally, incorporating email content analysis, such as topic modeling or sentiment analysis, could deepen our understanding of the context and nature of communications within and between communities.

The findings of this study demonstrate that descriptive network analysis is a powerful tool for generating hypotheses about organizational behavior and identifying key areas for deeper investigation. By prioritizing interpretability over purely statistical measures like modularity, this approach has provided actionable insights that align with known facts about Enron's structure and collapse. Future work could build on these results by applying predictive models to test the hypotheses generated or by exploring how similar methods might reveal patterns in other large-scale communication networks.

Related Work

The Enron email dataset has been widely studied in the field of social network analysis due to its rich metadata and unique insights into organizational communication. Researchers have applied various graph-based techniques to uncover structural dynamics, influential nodes, and community patterns within the dataset.

In the study "Analysis of social networks to identify communities and model their evolution", the authors utilize the Enron email corpus to investigate organizational structures using graph theoretical and spectral analysis methods. By constructing an email graph, they analyze properties such as degree distribution, clustering coefficients, and graph compactness. A key contribution of this study is the application of spectral analysis, revealing a rank-2 approximation of the email adjacency matrix. The findings emphasize the importance of robust data preprocessing for consistent results, advocating for standardized benchmarks in social network research [1].

Additional work by Burkhardt et al. demonstrates the scalability of graph processing techniques to understand communication networks, even at the exabyte scale. Their study highlights the practical challenges of processing large-scale communication datasets and the utility of centrality measures in identifying key individuals within corporate networks [2].

Other analyses of the Enron corpus have taken a practical and exploratory approach. For example, "Analyzing the Enron Email Corpus" provides a hands-on exploration of the dataset, using Python to parse and visualize communication patterns. The study underscores the importance of modularity and clustering in identifying functional teams and assessing their interactions within the network [3].

Harish Kumar's undergraduate research report, "Analyzing the Enron Email Dataset," focuses on the statistical properties of the network, including in-degree and out-degree distributions, community detection, and hierarchical clustering. The report provides a comprehensive overview of how various graph algorithms can be applied to detect influential individuals and map information flow in a corporate setting. The findings reinforce the value of email communication graphs in modeling organizational behavior and decision-making processes [4].

References

1. E. Yener and C. Rose, “Analysis of social networks to identify communities and model their evolution,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551. https://www.cs.rpi.edu/~yener/PAPERS/SECURITY/enron.pdf.
2. T. Burkhardt, J. Li, and D. Steffen, “Large-scale graph processing for corporate communication analysis,” in Proceedings of the IEEE High-Performance Extreme Computing Conference (HPEC), 2015. https://ieee-hpec.org/2015/Final_Presentations/hpec_2015_burkhardt.pdf.
3. Python for Engineers, “Analyzing the Enron email corpus,” 2015. [Online]. Available: https://new.pythonforengineers.com/blog/analysing-the-enron-email-corpus/.
4. H. Kumar, “Analyzing the Enron email dataset,” Undergraduate Research Report, University of California, Berkeley, 2015. https://www.stat.berkeley.edu/users/aldous/Research/Ugrad/HarishKumarReport.pdf.
5. P. Bondarenko, “Enron scandal: Downfall and bankruptcy,” Encyclopedia Britannica, 2024. https://www.britannica.com/event/Enron-scandal/Downfall-and-bankruptcy.
6. J. Chen, “Notable Enron executives and their roles in the scandal,” Investopedia, 2023. https://www.investopedia.com/enron-executives-6831970.
7. Enron Corporation, “Executive biographies,” Enron Pressroom. [Online]. Available: https://enroncorp.com/corp/pressroom/bios/.
8. N. NetworkX Developers, “louvain_communities,” NetworkX 3.4.2 Documentation, 2024. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.louvain.louvain_communities.html