INTA-6450 Course Project Final Paper
Discovering Wrongdoing from Enron Email Corpus

Ziyi Song

1. Introduction

The Enron scandal is one of the most infamous business frauds in American history involving the once energy giant Enron company. During the event, the Enron company intentionally engaged in fraudulent accounting practices to mislead investors about the performance of the company. After the fraud actions were discovered by the authorities and disclosed to the public, Enron's stock plummeted and the company eventually filed for bankruptcy.

Enron Corporation is a company that used to be a giant in the energy sector headquartered in Houston, Texas. The company was credited for its rapid growth and its innovative approaches on energy markets. Since 1985, Enron quickly spread into businesses such as natural gas, electricity, and broadband services, then grew into one of the major companies in the United States. This sudden success as a company turned Enron into an example of corporate excellence that earned the widest respect among investors and the public.

However, this rapid expansion was based on deceptive accounting practices and immoral corporate strategies.(Barrionuevo, A., 2006) These practices included inflating profits, hiding debts, and misleading investors about the company's true financial health. As a result, while Enron appeared successful on the surface, it was actually facing severe financial issues that were kept hidden from the public. Eventually, these dishonest actions were uncovered, leading to one of the largest corporate scandals in history.

After Enron filed bankruptcy protection, FERC hired Aspen Systems to collect and save data like Email and Oracle database from Enron. After the investigation, the email datasets were released to the public for historical research and academic purposes. These datasets are valued as one of the few publicly available mass collections of real email, which have also been used for research on natural language processing and machine learning. (Wikipedia contributors, 2024).

This paper aims to discover Enron Employees' fraudulent activities from the Enron email Corpus by using a series analytical methods. including natural language processing (NLP), data filtering, network analysis and k-means analysis.

2. Definition of Wrongdoing

For our purpose of analysis, we define the wrongdoing as the manipulation of financial reports, either by hiding/concealing debts & liabilities or inflating earnings & revenue, to mislead the market and investors about Enron's actual financial condition. We believe this wrongdoing is the critical activity that eventually led to the entire Enron scandal. Our goal is to identify the most obvious email evidence (targeting top 5) from the whole available Enron email corpus dataset which will reveal signs of Enron's intentional participation in the defined wrongdoing activities.

3.Literature Review

In preparation of the analysis, the first and foremost task is to identify the news and reports related to our definition of wrongdoing. For this purpose, a brief research with regard to the Enron scandal has been conducted. By reviewing the news and reports, we can focus on those related to our defined wrongdoing and get some inspiration of the financial keywords to kick off the initial screening analysis.

Several news articles, institution reports and research papers have been reviewed. A few of the news and papers discuss the wrongdoing of Enron that is highly related with the wrongdoing defined in this paper. The excerpts that contains information which provides an insight on where the analysis should start are listed as follows:

> *"On paper Chewco appeared independent; control was shared by Enron and outside investors in an arrangement that would permit Enron to keep some of its energy projects and debts off its books". "Chewco was an early example of the Byzantine investment structures that were Fastow's specialty. Its roots go back to 1993, when Enron formed the Jedi partnership with the giant California Public Employees' Retirement System (Calpers) to invest in natural gas projects."* (Peter Behr, "How Chewco Brought Down an Empire," The Washington Post, February 3 2002);
>
> *"... if anything, differentiates Enron's questionable use of off-balance-sheet special purpose entities, or SPEs…" (Steven L. Schwarcz, "Enron and the Use and Abuse of Special Purpose Entities in Corporate Structures". Duke Law Scholarship Repository, Cited by 212, 2002);*
>
> *"The accounting treatment accorded the 'stock for note exchanges' that took place when three of the Raptor SPEs (Talon, Timberwolf and Bobcat) were formed also violated GAAP". "Application of ARB 51 to SPEs is problematic in many cases because the parties involved in an SPE may not control its activities through voting equity interests."* (Anthony Catanach Jr., "Enron: A Financial Reporting Failure". Villanova University Law Library Digital Repository, 48 Vill. L. Rev. 1057 (2003);
>
> *"One of the first actions Jeffrey Skilling took when joining Enron was to seek approval from the SEC (Securities and Exchange Commission) to use "mark-to-market" accounting. Enron used mark-to-market accounting to book revenue from long-term contracts, including energy supply agreements, as soon as the contracts were signed, rather than waiting for the revenue to be earned over time."* (Matthew Briggson, "Enron: An Accounting Scandal That Changed Everything", Encoursa, May 2nd, 2023)

Based on review of the above reference articles, it is understood that Enron's manipulation of the financial documents mainly involves the following activity:

- Use of complex financial structures, such as Special Purpose Entities (SPEs) and off-balance-sheet transactions, to hide debt and inflate earnings.

- Partnership with Chewco with a purpose to help Enron keep certain debt off its balance sheet, which is considered as a key fraudulent accounting practice due to the fact that Enron concealed its control over Chewco and its financial dealings.
- Involvement of CALPERS, a large institutional investor, in the approving of Enron's arrangement of fraudulent accounting activity.

Accordingly, a list of keywords that are found to be strongly related to the wrongdoing activities has been summarized as follows for initial filtering evaluation:

"Chewco, special purpose entity, SPE, off-balance-sheet, partnership, accounting treatment, control, approval, CALPERS"

4. Data Handling

4.1 Email initial filtering strategy

The process of identifying relevant emails in the email dataset to discover Enron's wrongdoing involves an iterative and systematic approach. Since the given dataset has a tremendous amount of emails, narrowing it down according to relevance of communication is a necessary step. Initially, the filter only involves specific keywords and phrases known to be associated with fraudulent activities. These keywords were selected based on the incident's background that described in news and articles stated above, as well as general financial knowledge. This initial filtering strategy aimed to capture emails that involve discussing accounting manipulations, balance sheet modifications, violations of government regulations, or generate additional earnings. The initial keywords we've used were listed in section 3.

As the initial queries were executed, we found refining the keyword list was necessary to improve the precision of the results. This set of keywords did a great job filtering out emails that have relativity with financial activities. However, there exists a large portion of irrelevant emails during the initial exploration, such as emails regarding the news of Enron in 2001, PR emails regarding negative comments, students' direct emails to CEO asking for a job, a large amount of duplication of a single email due to forwarding between employees. Therefore, an elastic search query was constructed using a combination of "must" and "should" clauses. "must" ensured the retrieved emails contained essential keywords related to financial manipulation, and "should" included additional terms which could be the indicator of sensitive discussions. A "must not" was also added to omit irrelevant communications. This 3 layer approach enhanced the relevance of the search results by not only focusing on direct indicators, but also on the context where such sensitive discussions might occur. We've considered such sensitive discussions may not happen in the email explicitly, most of them are implicit and needed to be discovered with clues.

The final version of keywords:

Minimum match 4: off balance sheet, Special Purpose Entity, Special Purpose Vehicle, SPE, hide debt, inflate earnings, additional earnings, creative accounting, generate earnings, earnings management, revenue recognition, mark-to-market, conceal liabilities, financial manipulation, balance sheet, income statement, earnings forecast, balance sheet change, financial statement, pro forma earnings, deferred revenue, deferred expenses

Minimum match 1: confidential, urgent matter, strategy meeting, financial review, need approval, Finance Department, Accounting Department, Legal Department, Risk Management, Compliance Office

Exclude: personal, holiday, lunch, jobs, opening, WSJ, Seminar, meeting agenda, social event, team building, your consideration, TEXAS JOURNAL, Wall Street Journal, forwarded by, forwarded message, forward, fw:, fwd:, ----- forwarded by, please see the forwarded message, forward this email

Along with the enhancement of keywords, the filtering strategy also involved setting specific time duration to focus on periods when Enron was most active in its fraudulent activities. By limiting the elastic search to emails between 1998 and 2000, the search could concentrate on the timeframe that such activities were happening. If we include the emails in 2001, a large portion of them are discussing the news and counter measures against the media and the authorities. During the entire process, it is crucial to monitor and adjust the number of emails. The sample size was narrowed with each iterative step to improve accuracy.

4.2 Social Network Analysis

After keyword filtering on Enron's email collection, we have about 563 rows of datasets, which is much smaller than the original dataset of more than 10,000 rows.

We decided to perform Social Network Analysis on the 563 rows of data. We imported the networkx package and used ‘sender’ and ‘recipients’ as the nodes, with the lowest frequency set to 1. We obtained 575 nodes and 412 edges as a result. (Part of Code as follows)

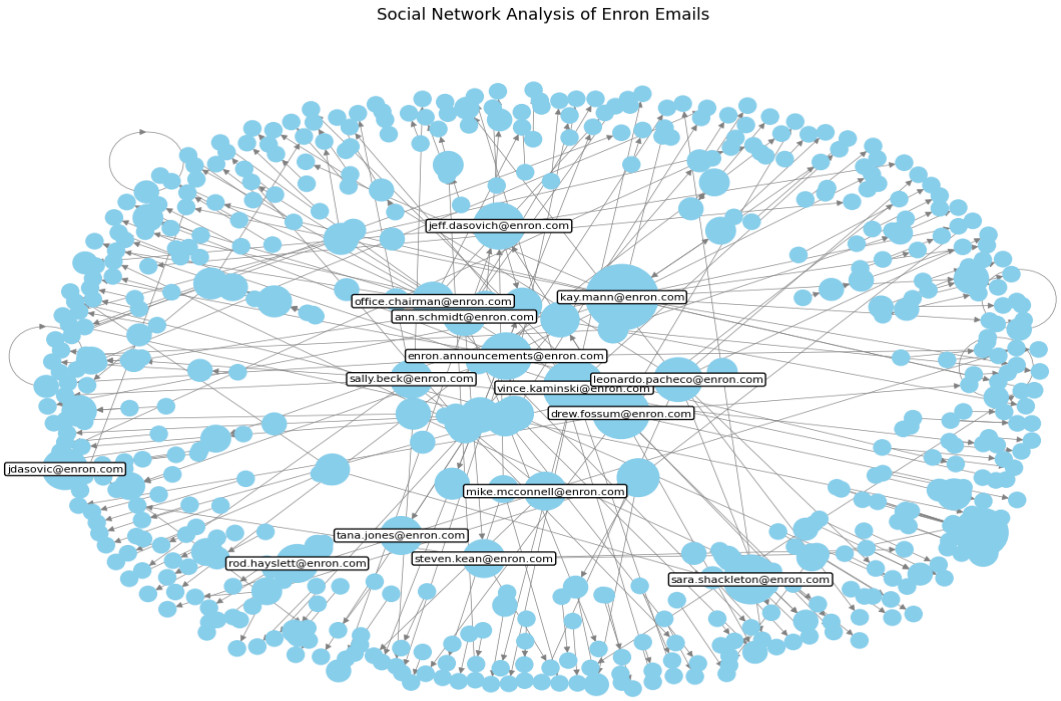
```
Found 563 messages matching the query.
Returned (563, 7) messages.
```

Social Network Analysis

In [32]:
```
import networkx as nx
import matplotlib.pyplot as plt
from collections import Counter
import warnings
warnings.filterwarnings("ignore")

# Create a directed graph
G = nx.DiGraph()

edge_counter = Counter()
for index, row in df.iterrows():
    sender = row['sender']
    recipients = row['recipients']

    if isinstance(recipients, list):
        for recipient in recipients:
            edge_counter[(sender, recipient)] += 1
    else:
        edge_counter[(sender, recipients)] += 1


min_frequency = 1
filtered_edges = [(sender, recipient) for (sender, recipient), count in edge_counter.items() if count >= min_frequency]


G.add_edges_from(filtered_edges)

print("Number of nodes:", G.number_of_nodes())
print("Number of edges:", G.number_of_edges())
```

```
Number of nodes: 575
Number of edges: 421
```

We also calculated the degree centrality score for each node, that is, for ‘senders’ and ‘recipients.’ The higher the score, the more important the role he/she played in this incident. We believe them with higher degree centrality would be our target emails for identifying the top 5 emails related to ‘financial fraud.’ Then, we also plotted the network images to make the social network relationships clearer. The larger the node, the higher the influence he/she had.



We finally selected the top 15 people in the network analysis, as shown below:

> ['kay.mann@enron.com', 'vince.kaminski@enron.com', 'drew.fossum@enron.com', 'sara.shackleton@enron.com', 'jeff.dasovich@enron.com', 'enron.announcements@enron.com', 'leonardo.pacheco@enron.com', 'jdasovic@enron.com', 'rod.hayslett@enron.com', 'sally.beck@enron.com', 'ann.schmidt@enron.com', 'mike.mcconnell@enron.com', 'steven.kean@enron.com', 'tana.jones@enron.com', 'office.chairman@enron.com']

We then analyzed each email address. Some accounts, such as ‘enron.announcements’ and ‘office.chairman,’ were identified as official rather than personal, so we excluded them from the analysis. This process reduced the focus to 13 influential individuals, all of whom likely played significant roles in the incidents. These individuals are also important figures within the company. We’ve attached part of the list detailing their positions, which we sourced online ("Enron Employees," n.d.). The individuals on this list appear to hold important roles within the organization.

Username	Name	Position	Additional Roles
kay.mann	Kay Mann	Employee	
sally.beck	Sally Beck	Employee	Chief Operating Officer
vince.kaminski	Vince Kaminski	Manager	Risk Management Head
drew.fossum	Drew Fossum	Vice President	
jeff.dasovich	Jeff Dasovich	Employee	Government Relation Executive
rod.hayslett	Rod Hayslett	Vice President	Chief Financial Officer and Treasurer
vince.kaminski	Vince Kaminski	Manager	Risk Management Head
drew.fossum	Drew Fossum	Vice President	
steven.kean	Steven Kean	Vice President	Vice President & Chief of Staff

After finalizing the top 13 individuals, we filtered the 563 emails to include only those sent or received by these 13 individuals. This process successfully reduced the email dataset by more than half, leaving 210 rows..

4.3 NLP analysis

Even after the dataset was reduced to 210 rows, we found the volume of text still high given the great deal of content added with each email. We introduced the “nltk” library to reduce the text without eliminating any of the information needed. Based on what we learned in class, we applied “nltk” to remove stop words that significantly reduced the amount of unnecessary content.

We printed the 10 most common words to review the result of the filtered text. The frequency of words such as "enron," "cc," "pm," "new," and "time" were quite high, which indicated that they still introduced noise. Thus, we extended the stop words library with these terms in an attempt to reduce further noise.

Lastly, cleaned text was inserted into a new column named cleaned_text, thus preparing for subsequent word cloud generations after the K-means cluster Analysis.

Some sample of our code to preprocess text is as follows:

```python
def preprocess_text(text, remove_numbers=True, additional_stopwords=None):
    if not isinstance(text, str):
        return ''

    text = text.lower()

    text = re.sub(f'[{re.escape(string.punctuation)}]', '', text)

    if remove_numbers:
        text = re.sub(r'\d+', '', text)

    stop_words = set(stopwords.words('english'))
    stop_words.update({'enron', 'cc', 'subject', 'pm', 'company', 'said', 'also', 'would', 'new','time'})
    if additional_stopwords:
        stop_words.update(additional_stopwords)

    words = text.split()
    words = [word for word in words if word not in stop_words]

    text = ' '.join(words)
    text = re.sub(r'\s+', ' ', text).strip()

    return text
```

4.3 Cluster Analysis

To further analyze the column that was preprocessed with stop word removal, we did a full K-means clustering analysis. The goal of this was to segment similar emails into clusters so as to have a better insight into their patterns and characteristics.

TF-IDF Transformation

First, we used the TfidfV package. We changed the cleaned text into a feature matrix in TF-IDF form. step transforms text information into numerical forms that point to the importance of each word within a document with respect to the entire dataset.

Finally, we computed the sum of TF-IDF for every email and created a new column called tfidf_score to store this for future use as a numerical summary.

```
vectorizer = TfidfVectorizer(max_df=0.5, min_df=2, stop_words='english')
X = vectorizer.fit_transform(top_emails['cleaned_text'])
tfidf_scores = X.toarray().sum(axis=1)
top_emails['tfidf_score'] = tfidf_scores
```

K-Means Clustering

Since the emails were unlabeled, we decided to run the K-Means algorithm on the TF-IDF matrix based on the clustering technique learned in class. Based on the results of multiple tries, we chose 3 as the best number of clusters (num_clusters). Then, we fit the K-Means model using kmeans.fit(X), which assigned each email to a cluster and stored the cluster labels in a new column, cluster.

```
num_clusters = 3
kmeans = KMeans(n_clusters=num_clusters, random_state=40)
kmeans.fit(X)
```

PCA Dimensionality Reduction

To have better clustering results, we applied PCA to reduce the high-dimensional TF-IDF matrix into 2 dimensions, which could clearly show the 2D scatter of emails for visualization:

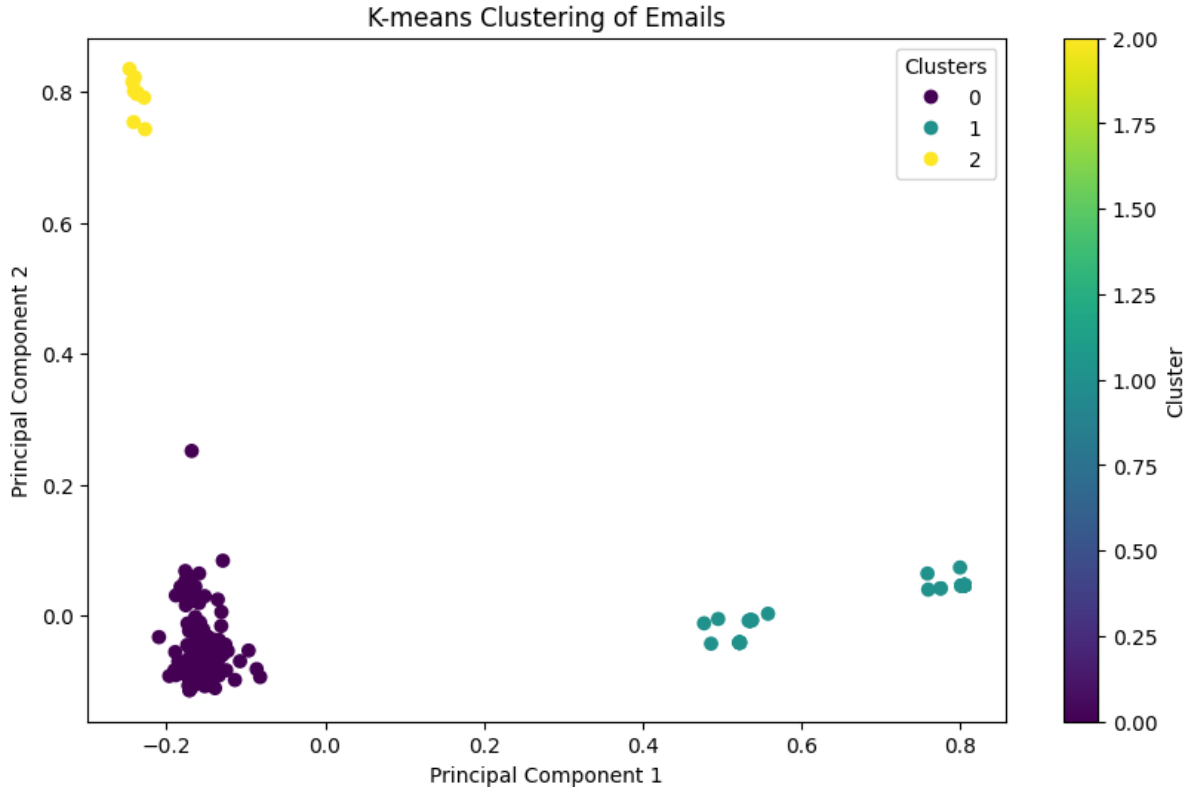
```
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(X.toarray())
```

Cluster Visualization

We then used Matplotlib to create a scatter plot of the reduced data points, which is based on the result of previous PCA and K-means. Each point was colored according to its cluster label, and we added a color bar, title, axis labels, and a legend to make the plot more readable. The scatter plot showed that emails fell into three clear clusters.

```
plt.figure(figsize=(10, 6))
scatter = plt.scatter(reduced_data[:, 0], reduced_data[:, 1], c=top_emails['cluster'], cmap='viridis',
marker='o')
plt.title('K-means Clustering of Emails')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.colorbar(label='Cluster')

handles, labels = scatter.legend_elements()
legend = plt.legend(handles, labels, title="Clusters")
plt.show()
```

The below is the K-means cluster, which are very clear and easy to distinguish.



4.4 Word Cloud and Cluster Selection

After running the K-means algorithm, we got the following distribution of emails in the three clusters:

```
cluster_counts = top_emails['cluster'].value_counts()
# Print the count of emails in each cluster
for cluster_id, count in cluster_counts.items():
    print(f"Cluster {cluster_id}: {count} emails")
-------------------------
Cluster 0: 159 emails
Cluster 1: 41 emails
Cluster 2: 10 emails
```

To further analyze and differentiate these clusters, we decided to focus on the cleaned_text column and visualize the text content within each cluster using word clouds. A word cloud is a visual representation of word frequency; the more frequent words will appear larger. In this way, we could have a quick idea about the key terms in each cluster and find out whether they fit within the topics of financial fraud or not.

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Function to generate word cloud
def generate_wordcloud(text):
    wordcloud = WordCloud(width=800, height=400, background_color='white').generate(text)
    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.show()


for cluster_id in range(num_clusters):
    cluster_text = ' '.join(top_emails[top_emails['cluster'] == cluster_id]['cleaned_text'])
    print(f"Word Cloud for Cluster {cluster_id}")
    generate_wordcloud(cluster_text)
```

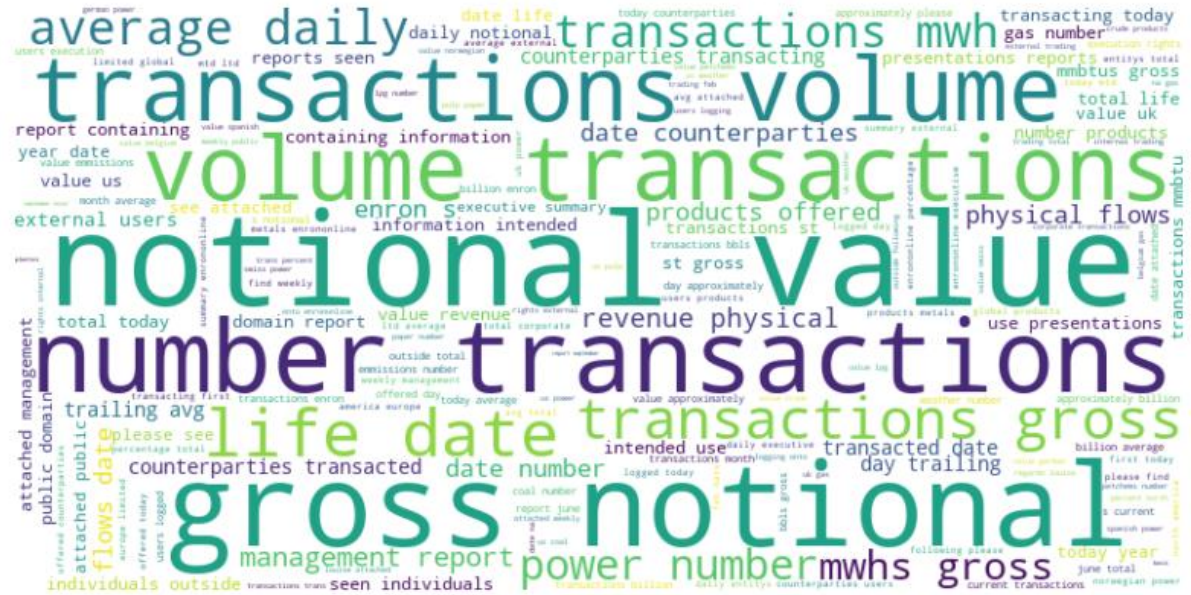
Cluster 0

Key Words: "market," "price," "power," "customer," "business," "service," "energy," "cost," "year."
Analysis: The dominant terms like "market," "price," and "power" indicate that this cluster is closely related to market dynamics, pricing strategies, and energy-related services. Words like "cost", "financial", "assets", "stock", and "earnings" bring into focus discussions about financial aspects. Thus, this cluster may be linked with financial transactions, pricing abnormalities, or potential market manipulations. All these would directly relate it to investigations regarding financial fraud.

Cluster 1

Key Words: "transactions," "volume," "notional," "value," "gross," "number," "mwhs," "average," "life."

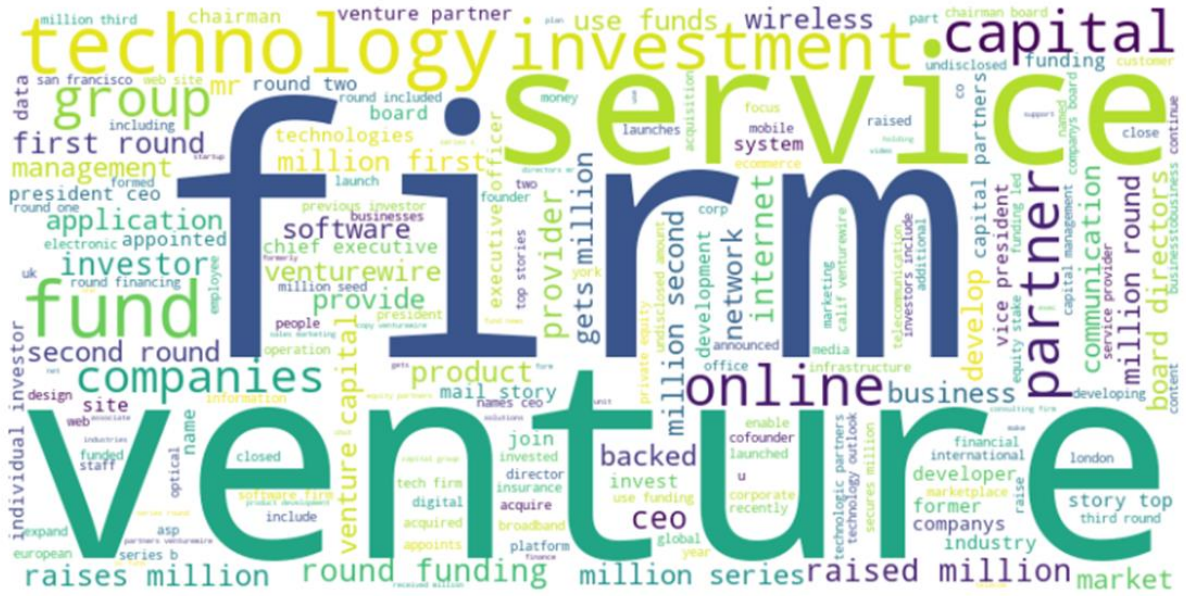
Analysis: This cluster heavily focuses on transaction-specific terms, including "notional value," "gross transactions," and "volume." The frequent appearance of some quantitative terms indicates that this cluster is related with transactional data, trade volumes, or possibly energy market analytics. While it does provide some insight into transactional activity, it's more technical and less indicative keywords of financial fraud than Cluster 0.



Cluster 2

Key Words:"firm," "venture," "service," "technology," "investment," "partner," "capital," "companies."

Analysis: This cluster focuses on corporate-level topics such as venture capital, investment strategies, and partnerships. Terms such as "technology" and "investment" lead one to believe this cluster could pertain more to strategic discussions or corporate investments. This cluster is less likely to relate directly to financial fraud but may hint at the organizational structure or funding sources.



Cluster Selection

After analyzing the three clusters, we decided to focus on Cluster 0, as it closely relates to topics such as "market," "price ".Moreover, smaller terms such as "issue," "financial," "review", "cost", "assets" and "account" further reinforced its relevance to fraud detection.

4.5 Final Analysis and Summary

Final Analysis

To get a deeper analysis of the dataset, we will use the TF-IDF scores calculated above. The TF-IDF score indicates words that are not only frequent in a document but also very distinctive across the dataset. The higher the score, the higher the probability that an email contains important and unique information. This makes high-scoring emails more relevant for investigation because they likely contain specific details related to financial activities or irregularities.

We filtered Cluster 0 for the selection of 52 highest-scoring emails as contributed by their TF-IDF scores. Because the emails that linked to financial fraud may be subtly or ambiguously worded, we expanded our scope of review to 52 emails. Our team consisted of four members, and we divided the 52 emails equally among us; each member was to review 13 emails.

```
top_52_emails = target_cluster.sort_values(by='tfidf_score', ascending=True).head(52)
```

We identified 4 emails strongly related to financial fraud during the first round of review. In order not to miss any of the top emails, we extended the review to the next high-scoring email and identified 1 additional email linked to financial fraud.

In sum up, from the 53 emails reviewed, 5 emails were identified as directly related to financial fraud. We printed the final result below.

> Email ID: 196894, Subject: 2001 Goals and Objectives, Cluster: 0
> Email ID: 52514, Subject: Re: Q2 for Patten Case, Cluster: 0
> Email ID: 66884, Subject: NYMEX F/S, Cluster: 0
> Email ID: 234695, Subject: Incorporating comments from Rick Causey, Cluster: 0
> Email ID: 248242, Subject: Potential Write-offs, Cluster: 0

In the next paragraphs, we will conduct a detailed analysis of these 5 emails, examining the content and context to better understand their connection to financial fraud.

Summary

In general, we think we did a pretty good job in detecting emails related to financial fraud. The following is a wrap-up of how we did this.

We started with stopword removal and extraction of meaningful terms using “nltk”, by which we were able to extract the relevant keywords and cleaned the text. Then we did Social Network Analysis; thus, we identified influential persons and their associated e-mails for further investigation.

We also used TF-IDF scores to quantitatively evaluate the relevance of each email. K-means clustering algorithm groups similar emails, while PCA is giving a clear visualization of those clusters. A word cloud analysis enabled us to get an overview of the terms standing out in each cluster. And we selected the best cluster according to the word cloud analysis.

Last, we changed to the manual check of 53 emails, which gave us 5 emails with high relevance for financial fraud.

5. Analysis of Wrongdoing Identified in Emails

5.1 Email-1

This email mentioned inaccuracies in the financial statement and improper accounting measures explicitly. It mentions there are some improper inclusion and exclusion that affected the earnings by $192MM. It also highlights inaccuracies in the Hyperion legal structure, which reflects the actual ownership of Enron. This email mentioned a company called Enron Gas Liquids, Inc. is excluded. This company was a subsidiary of Enron Corporation, involved in processing and marketing of liquefied natural gas.

Our secondary research shows Enron did not provide standalone financial data for many subsidiaries, this company is one of them. The exclusion of this company may be due to the reason that this subsidiary had a major deficit, excluding that may lead to a better number on

Enron’s financial conditions. The Securities Act of 1933, also called the "truth in securities" law prohibits fraudulent financial reporting, which means adjusting financial data to present a false financial condition is not lawful.

Identified Email Content #1:

"RE: NYMEX F/S
At: 2000-06-30 14:49:00+00:00
From: darin.talley@enron.com
To: tana.jones@enron.com mark.jones@enron.com julie.gartner@enron.com
mark.ng@enron.com faith.killen@enron.com johnson.leo@enron.com
gilda.bartz@enron.com
CC: georgeanne.hodges@enron.com mark.frank@enron.com
wes.colwell@enron.com trey.hardy@enron.com patricia.anderson@enron.com
BCC: georgeanne.hodges@enron.com mark.frank@enron.com
wes.colwell@enron.com trey.hardy@enron.com patricia.anderson@enron.com
Body:

ENA BA&R is currently in the process of preparing the 12/31/99 financial statements to be filed with the NYMEX by July 21st in order to keep our trading seat. These financial statements reflect all of the North America trading operations. ECTEG1 is ENA's legal rollup within Hyperion and should reflect all of the North American trading (i.e., price risk management) activity. Therefore, Consolidated Reporting, with the assistance of Trading Accounting and Tax, prepared the financial statements with ECTEG1 activity. These financial statements are complete and ready for BA&R management review and reflect $722MM in earnings.

However, as of 6/29, we were notified by the Tax department that three companies were included in ECTEG1 as of 12/31/99 that were no longer legally owned by ENA. In addition, there was one company that became ENA-owned during 1999. The companies are as follows:

Improperly Included:

- ECM Treasury Consolidated (Company 969)
- ECT Retail Services (Company 1Q6)
- ECT Power Marketing, Inc. - Retail (Company 890)

Improperly Excluded:

- Enron Gas Liquids, Inc. (Company 104)

I have corresponded with Tana Jones from the Legal department, Johnson Leo in Corporate Reporting, and Julie Gartner from the Tax department to determine if the financial statements that we prepared reflected accurate and complete North America trading activity. These discussions confirmed that the Hyperion legal structure was not accurately reflecting ENA's actual legal ownership. In addition, the three companies have $192MM in trading activity income that would be removed from the bottom line. This activity should probably be reported to the NYMEX (regardless of legal structure) as it represents price risk management activity.

The primary issue is the accuracy of the financial statements. We are prepared to have to restate the compiled financials as soon as we are notified where we can retrieve accurate ledger information. The other main issue is timing. We need to file these financial statements no later than July 21st. Before that can happen, they need to be reviewed by several levels of management in several departments.

The Legal department is currently reviewing the NYMEX requirements in order to determine the proper activity to report. As soon as that is determined, we can proceed with data retrieval.

Please contact me if you have any questions, concerns, or comments."

5.2 Email-2

This email discussed some adjustments in accounting methods to reflect a desired financial condition. It shows an intention of modifying revenue recognition and deferred profits or liabilities. "Does the balance sheet have to change a little?" is a very straight-forward sentence to ask for manipulation of the balance sheet. Although the sender mentioned he has no right to make changes, suggesting it as an executive level person is suspicious enough.

The email explicitly mentioned changing revenue recognition methods to manipulate earnings and suggested introducing deferred profits or liabilities to offset changes that reflect the true financial situation. The SEC (Securities and Exchange Commission) requires firms should not manipulate their financial statements to mislead investors or the authorities. This email violated these regulations and attempted to improve the company's image and financial condition through a modified balance sheet.

Identified Email Content #2:

"RE: Re: Q2 for Patten Case
At: 2000-09-17 20:18:00+00:00
From: jeff.dasovich@enron.com
To: kkupiecki@arpartners.com
CC: nan
BCC: nan
Body:
Hey, nice spreadsheet. Two minor questions:

1. **Provision for Taxes:**
   Isn't the provision for taxes on the income statement actually the taxes on the "income" they made from using the sales method (equal to 46%), rather than what they will actually pay the IRS under the installment method? I think the notes show how, instead of paying the $4.1M based on their recognized income, they pay some itty-bitty amount based on installment.
   If so, I think they actually get a 46% tax break on the $1 million and change that they lose using a cash basis. Still a loss, just not as big. Anyway, I'm not sure if I'm thinking straight on this, but that's how I read the numbers.

2. **Balance Sheet Adjustment:**
   Does the balance sheet have to change a little? For example, does shareholder equity change since the income that goes to retained earnings is now a loss, rather than a gain?
   Also, if revenue is recognized on a cash basis and is now much smaller, there needs to be another liability to equal out the decrease in revenues with the still large notes receivables on the asset side (as you note in the answer, the notes receivables stay the same). Seems like they might need a liability like "deferred profits" or some such thing, such that the ["deferred profits" + revenues (recognized on cash basis)] = notes receivables.

Anyway, I may not have this right, but I thought I'd bring it up to see what you think.

Best,
Jeff"

5.3 Email-3

This email suggested accelerating expenses to a prior fiscal year to utilize non-recurring earnings. The second paragraph in the email asked "identify potential write-offs in 2000", also it suggested moving the "DOT user fee" and payment to Mobil from 2001 to 2000 to relieve the negative impact on financial conditions in the year 2000. Those actions violated GAAP (Generally Accepted Accounting Principles) which requires the financial statement to be complete and accurate. After Enron's incident, the Sarbanes-Oxley Act of 2002 (SOX) was settled to prevent such actions. Section 404 of this act requires that company management and external auditors evaluate internal controls to ensure the accuracy and completeness of financial statements, and this financial statement should reflect the actual financial well-being of a company.

Those manipulations show Enron attempted to improve the net income on paper by disguising some financial or payment obligations to present a healthier financial state for 2000. At the end of the email, the sender says "we could probably accelerate them into 2000 without too much visibility". This shows they were aware this is a violation of regulations and they don't want such actions to be discovered.

Identified Email Content #3:

"RE: Potential Write-offs
At: 2000-08-15 08:06:00+00:00
From: bob.chandler@enron.com
To: rod.hayslett@enron.com
CC: harry.walters@enron.com elaine.concklin@enron.com
steve.kleb@enron.com allen.joe@enron.com aurora.dimacali@enron.com
james.centilli@enron.com
BCC: harry.walters@enron.com elaine.concklin@enron.com
steve.kleb@enron.com allen.joe@enron.com aurora.dimacali@enron.com
james.centilli@enron.com
Body:

This is a follow-up on our efforts to identify potential write-offs in 2000 to utilize the possible availability of non-recurring earnings at the Corporate level.

Attached are spreadsheets for the 2nd quarter noteholders' reports balance sheets, including detailed breakdowns of balance sheet items. We reviewed these schedules for possible additional write-off candidates. This review confirmed that the big items had already been identified (regulatory assets and south-end impairment reserve).

However, reviewing the spreadsheets reminded us that we have:

- **DOT safety user fees** due in December (~$1.1MM), and
- A **$1.5MM annual payment to Mobil** due by 1/1/2001.

While both of these items would more appropriately be charged to expense in 2001, we could probably accelerate them into 2000 without too much visibility."

5.4 Email-4

This email involves language adjustments in accounting terms and earnings reporting, and can be perceived as an attempt to present financial outcomes in a specific way to hide details on the negative side. The wording adjustments including editing or deleting specific phrases like "on accounting issues and reporting earnings" could result in a healthy presentation on the financial well-being which may have the purpose of bringing less attention to accounting measurements. Rick Causey is the chief accounting officer at Enron, this conversation involving him may indicate these adjustments were planned and executed at company level to meet specific financial targets.

Identified Email Content #4:

"RE: Incorporating comments from Rick Causey
At: 2000-08-16 16:29:00+00:00
From: sally.beck@enron.com
To: cathy.phillips@enron.com
CC: nan
BCC: nan
Body:

Cathy,

Slight wording change from Rick Causey: Add the words at the beginning of the third sentence in *italics* and drop the phrase "on accounting issues and reporting earnings" at the end of that third sentence. Thanks. Clear as mud?

—Sally

Brent Price will be joining Enron Global Markets as Vice President of Operations and Chief Accounting Officer. He will report to the EGM Office of the Chairman and to Sally Beck, Vice President of Global Risk Management Operations. *In his role as Chief Accounting Officer, he will also work*

closely with Rick Causey, Executive Vice President and Chief Accounting Officer for Enron Corp. Reporting to Brent in his new position will be Sheila Glover, business controller for Global Financial Products; Todd Hall, business controller for Weather; and Scott Earnest, business controller for Global Products and Coal. In addition, Tom Myers has joined Brent's management team as Director of Accounting.

Brent and his team are responsible for all accounting, risk reporting, and trading operations for all the businesses within EGM."

5.5 Email 5

This email explicitly mentioned "creative and cost saving ideas" and "generate additional earnings", as well as monetize assets. This suggests Enron is positively seeking complicated accounting measures to improve the company's financial performance. That is even stated as a goal of 2001. It also mentioned "leveraging cost with Enron Compression Services or other off balance sheet structures", indicating the company was seeking to reduce the liability on the book through some off balance sheet financing. This violated some laws and regulations mentioned above such as GAAP and SEC because those financial structures may not be fully disclosed and will mislead investors and authorities.

"Arthur Andersen experts" mentioned in email is an external auditing organisation that may indicate Enron is seeking accounting treatments to legalize their financial operations. Later news shows Arthur Andersen was convicted of obstruction of justice in the Enron scandal.

Identified Email Content #5:

"RE: 2001 Goals and Objectives
At: 2000-12-12 17:32:00+00:00
From: james.centilli@enron.com
To: rod.hayslett@enron.com
CC: dave.waymire@enron.com
BCC: dave.waymire@enron.com
Body:

1. Develop creative and cost saving ideas to generate additional earnings.
2. Coordinate administrative and analytical support for the monetization of a significant level of non-strategic assets. Include in the analysis all appropriate contacts within ET&S, Arthur Andersen experts and corporate accounting contacts.
3. Perform economic analysis of projects and asset sales to provide the management team a complete risk assess evaluation on all projects. Provide financial support to Marketing in developing financing structures and evaluation process for structured products. Update evaluation as factors change, and provide information to appropriate contacts within ET&S.

Other Ideas for goals and objectives:

```
Developing Economic Analysis utilizing the Revenue Management information.

Northern's South End Assets - Depreciation Study, Enhancing Value of
Assets, monetizing non-strategic assets.

TW's Expansion - leveraging cost with Enron Compression Services or other
off balance sheet structures.

Northern's North End Power Generation - structured deals support”
```

6. Current Limitations (unfixable issues) and Further Considerations

We have successfully uncovered five emails directly related to the wrongdoing defined earlier. The primary approach utilized email screening based on a predefined set of keywords. The keyword set has gone through several iterations to identify the most relevant emails associated with the defined wrongdoing.

However, we have to admit that there are some inevitable limitations (unfixable issues) in the current approach. The keyword-based filtering approach inherently limits the analysis to emails containing pre-defined terms. Emails discussing wrongdoing indirectly, using vague language or euphemisms, may go undetected. This limitation stems from the inability to anticipate all possible ways wrongdoing might be communicated, and no refinement of keywords can completely eliminate this issue.

Moreover, the current approaches might struggle to handle the volume or complexity efficiently. Several rounds of manual reviews have been conducted to refine keyword sets. Hence, if more data need to be included in the analysis in the future, the extensibility of current approaches would face some difficulties of being applied to additional datasets.

In addition to the methodology employed, several approaches could be leveraged to deepen the understanding of the dataset and uncover additional insights:

1. The identified emails could be cross-referenced with external sources to validate the findings and provide a broader context. External information may include Enron’s historical financial statements and legal or regulatory filings related to the investigation, enhancing the credibility and depth of the analysis.

2. The employed elastic search queries and clustering focus on explicit textual patterns. These approaches could be extended, to some extent, to perform deeper semantic investigation to uncover wrongdoings that were communicated in vague language and help with the scalability of the approaches.

3. A user interface could be developed to facilitate results analysis. Identified top emails can be printed out in a readable format to improve manual screening efficiency.

7. Conclusions

The project focuses on uncovering evidence from Enron email corpus to prove the existence of the wrongdoing the researchers defined. The primary wrongdoing is the manipulation of

financial reports, either by hiding/concealing debts & liabilities or inflating earnings & revenue, to mislead the market and investors about Enron's actual financial condition. The team starts with a keyword-based screening methodology with an initial set of keywords. Iterative manual reviews were involved in the process to refine the keyword sets to eventually point to five emails directly related to the wrongdoing.

Several other approaches, such as social network analysis, natural language processing, and K-mean clustering, are leveraged to assist with determining the final candidates.

The approaches are highly successful which can enable researchers to quickly investigate a dataset containing hundreds and thousands of data points to filter out important information.

In [26]:

```
import requests
import pandas
from dateutil import parser
import json
host = 'http://18.188.56.207:9200/'
requests.get(host + '_cat/indices/enron').content
```

Out[26]:

```
b'yellow open enron lVq0is2BTCmgDk2kFyZHTQ 1 1 251735 73233 1.4gb 1.4gb\n'
```

In [27]:

```
def elasticsearch_results_to_df(results):
    '''
    A function that will take the results of a requests.get
    call to Elasticsearch and return a pandas.DataFrame object
    with the results
    '''
    hits = results.json()['hits']['hits']
    data = pandas.DataFrame([i['_source'] for i in hits], index = [i['_id'] for i in hits])
    data['date'] = data['date'].apply(parser.parse)
    return(data)

def print_df_row(row):
    '''
    A function that will take a row of the data frame and print it out
    '''
    print('_____')
    print('RE: %s' % row.get('subject',''))
    print('At: %s' % row.get('date',''))
    print('From: %s' % row.get('sender',''))
    print('To: %s' % row.get('recipients',''))
    print('CC: %s' % row.get('cc',''))
    print('BCC: %s' % row.get('bcc',''))
    print('Body:\n%s' % row.get('text',''))
    print('_____')
```

In [28]:

```
# df.columns
# Index(['date', 'text', 'sender', 'recipients', 'subject', 'cc', 'bcc'], dtype='object')
```

Match Financial Keywords

In [29]:

```
# Define a focused Elasticsearch query to retrieve the top 20 most relevant emails
doc = {
    "query": {
        "bool": {
            "must": [
                {
                    "bool": {
                        "should": [
                            {"match": {"text": "off balance sheet"}},
```

```
                {"match": {"text": "Special Purpose Entity"}},
                {"match": {"text": "Special Purpose Vehicle"}},
                {"match": {"text": "SPE"}},
                {"match": {"text": "hide debt"}},
                {"match": {"text": "inflate earnings"}},
                {"match": {"text": "additional earnings"}},
                {"match": {"text": "creative accounting"}},
                {"match": {"text": "generate earnings"}},
                {"match": {"text": "earnings management"}},
                {"match": {"text": "revenue recognition"}},
                {"match": {"text": "mark-to-market"}},
                {"match": {"text": "conceal liabilities"}},
                {"match_phrase": {"text": "financial manipulation"}},
                {"match_phrase": {"text": "balance sheet"}},
                {"match_phrase": {"text": "income statement"}},
                {"match_phrase": {"text": "revenue recognition"}},
                {"match_phrase": {"text": "earnings forecast"}},
                {"match_phrase": {"text": "balance sheet change"}},
                {"match_phrase": {"text": "financial statement"}},
                {"match": {"text": "pro forma earnings"}},
                {"match": {"text": "deferred revenue"}},
                {"match_phrase": {"text": "deferred expenses"}}
            ],
            "minimum_should_match": 4
        }
    },
    {
        "bool": {
            "should": [
                {"match": {"text": "confidential"}},
                {"match": {"text": "urgent matter"}},
                {"match": {"text": "strategy meeting"}},
                {"match": {"text": "financial review"}},
                {"match": {"text": "need approval"}},
                {"match": {"text": "Finance Department"}},
                {"match": {"text": "Accounting Department"}},
                {"match": {"text": "Legal Department"}},
                {"match": {"text": "Risk Management"}},
                {"match": {"text": "Compliance Office"}}
            ],
            "minimum_should_match": 1
        }
    }
],
"filter": [
    {
        "range": {
            "date": {
                "gte": "1998-01-01",
                "lte": "2000-12-31"      #This filter is to avoid returning news.
            }
        }
```

```
            }
        ],
        "must_not": [
            {"match": {"text": "personal"}},
            {"match": {"text": "holiday"}},
            {"match": {"text": "lunch"}},
            {"match": {"text": "jobs"}},
            {"match": {"text": "opening"}},
            {"match": {"text": "WSJ"}},
            {"match": {"text": "Seminar"}},
            {"match": {"text": "meeting agenda"}},
            {"match_phrase": {"text": "social event"}},
            {"match_phrase": {"text": "team building"}},
            {"match_phrase": {"text": "your consideration"}},
            {"match_phrase": {"text": "TEXAS JOURNAL"}},
            {"match_phrase": {"text": "Wall Street Journal"}},
            {"match_phrase": {"text": "forwarded by"}},
            {"match_phrase": {"text": "forwarded message"}},
            {"match": {"text": "forward"}},
            {"match": {"text": "fw:"}},
            {"match": {"text": "fwd:"}},
            {"match_phrase": {"text": "----- forwarded by"}},
            {"match_phrase": {"text": "please see the forwarded message"}},
            {"match_phrase": {"text": "forward this email"}}
        ]
      }
    },
    "from": 0,
    "size": 10000,  # Retrieve top 20 results
}
```

In [30]:

```
r = requests.get(
    host + 'enron/_search',
    data=json.dumps(doc),
    headers={'Content-Type': 'application/json'}
)
```

In [31]:

```
r.raise_for_status()

total_hits = r.json()['hits']['total']
if isinstance(total_hits, dict):
    total_matches= total_hits.get('value', 0)
else:
    total_matches= total_hits
print(f"Found {total_matches} messages matching the queray.")
df = elasticsearch_results_to_df(r)
print(f"Returned {df.shape} messages.")
```

```
Found 563 messages matching the queray.
Returned (563, 7) messages.
```

Social Network Analysis

In [32]:

```
import networkx as nx
import matplotlib.pyplot as plt
from collections import Counter
import warnings
warnings.filterwarnings("ignore")

# Create a directed graph
G = nx.DiGraph()

edge_counter = Counter()
for index, row in df.iterrows():
    sender = row['sender']
    recipients = row['recipients']

    if isinstance(recipients, list):
        for recipient in recipients:
            edge_counter[(sender, recipient)] += 1
    else:
        edge_counter[(sender, recipients)] += 1


min_frequency = 1
filtered_edges = [(sender, recipient) for (sender, recipient), count in edge_counter.items() if count >=
min_frequency]


G.add_edges_from(filtered_edges)

print("Number of nodes:", G.number_of_nodes())
print("Number of edges:", G.number_of_edges())
```

```
Number of nodes: 575
Number of edges: 421
```

In [33]:

```
degree_centrality = nx.degree_centrality(G)
top_15_nodes = sorted([node for node in degree_centrality if node == node],
key=degree_centrality.get, reverse=True)[:15]
print("Top 15 most connected nodes:", top_15_nodes)
```

```
Top 15 most connected nodes: ['kay.mann@enron.com', 'vince.kaminski@enron.com',
'drew.fossum@enron.com', 'sara.shackleton@enron.com', 'jeff.dasovich@enron.com',
'enron.announcements@enron.com', 'leonardo.pacheco@enron.com', 'jdasovic@enron.com',
'rod.hayslett@enron.com', 'sally.beck@enron.com', 'ann.schmidt@enron.com',
'mike.mcconnell@enron.com', 'steven.kean@enron.com', 'tana.jones@enron.com',
'office.chairman@enron.com']
```

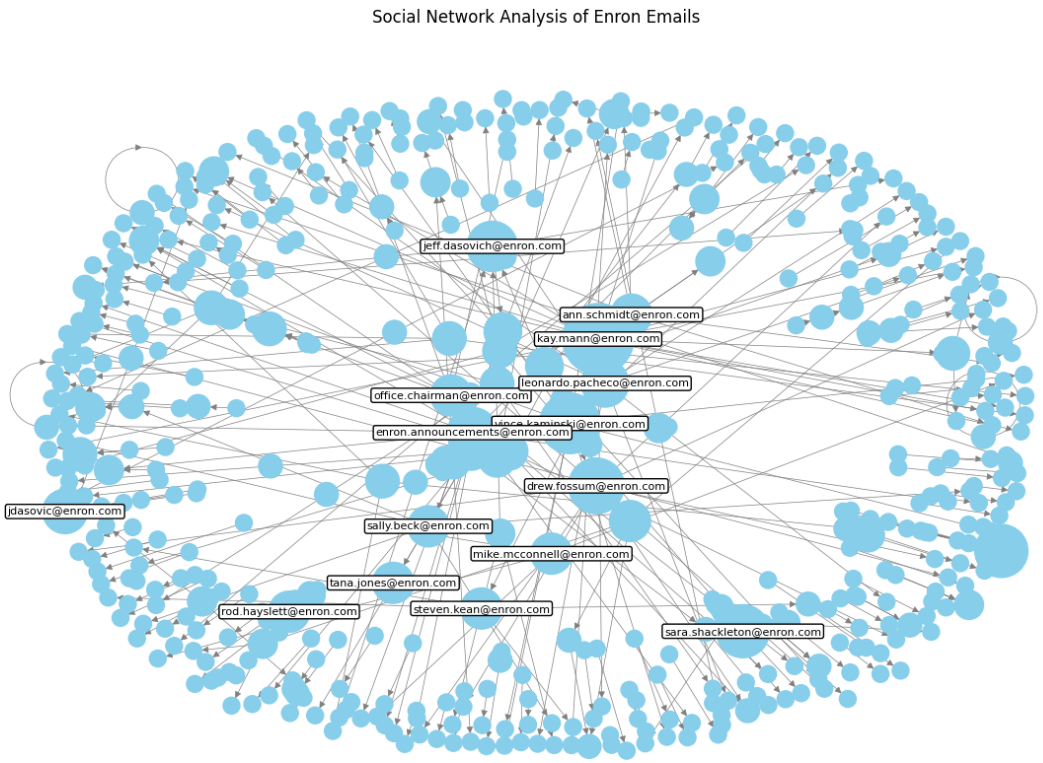
In [34]:

```
top_15_df = pandas.DataFrame({
```

```
    'Node':  top_15_nodes,
    'Degree Centrality': [degree_centrality[node] for node in top_15_nodes]
})
```

In [35]:

```
plt.figure(figsize=(12, 8))
pos = nx.spring_layout(G, seed=42)
node_sizes = [G.degree(node) * 150 for node in G.nodes()]
nx.draw(G, pos, with_labels=False, node_size=node_sizes, node_color='skyblue', edge_color='gray',
font_size=8, width=0.5)
for node in G.nodes():
    if node in top_15_nodes:
        x, y = pos[node]
        plt.text(x, y, node, fontsize=8, ha='center', va='center', bbox=dict(facecolor='white',
edgecolor='black', boxstyle='round,pad=0.2'))


plt.title('Social Network Analysis of Enron Emails')
plt.show()
```



Filter the emails from Top 15 Person

In [36]:

```
top_15_nodes
```

Out[36]:

```
['kay.mann@enron.com',
```

```
 'vince.kaminski@enron.com',
 'drew.fossum@enron.com',
 'sara.shackleton@enron.com',
 'jeff.dasovich@enron.com',
 'enron.announcements@enron.com',
 'leonardo.pacheco@enron.com',
 'jdasovic@enron.com',
 'rod.hayslett@enron.com',
 'sally.beck@enron.com',
 'ann.schmidt@enron.com',
 'mike.mcconnell@enron.com',
 'steven.kean@enron.com',
 'tana.jones@enron.com',
 'office.chairman@enron.com']
```

In [37]:

```
top_15_nodes.remove("enron.announcements@enron.com")
top_15_nodes.remove("office.chairman@enron.com")
top_nodes = top_15_nodes
```

In [38]:

```
top_emails = df[
    (df['sender'].isin(top_nodes)) |
    (df['recipients'].apply(lambda x: isinstance(x, str) and any(r in x.split() for r in top_nodes)))
]

print("Number of emails from or to top 13 most connected nodes:", top_emails.shape[0])
```

```
Number of emails from or to top 13 most connected nodes: 210
```

NLTK analysis to remove stopwords

In [39]:

```
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
import re
import string
from sklearn.feature_extraction.text import TfidfVectorizer

def preprocess_text(text, remove_numbers=True, additional_stopwords=None):
    if not isinstance(text, str):
        return ''

    text = text.lower()

    text = re.sub(f'[{re.escape(string.punctuation)}]', '', text)

    if remove_numbers:
        text = re.sub(r'\d+', '', text)

    stop_words = set(stopwords.words('english'))
```

```
    stop_words.update({'enron', 'cc', 'subject', 'pm', 'company', 'said', 'also', 'would', 'new','time'})
    if additional_stopwords:
        stop_words.update(additional_stopwords)

    words = text.split()
    words = [word for word in words if word not in stop_words]

    text = ' '.join(words)
    text = re.sub(r'\s+', ' ', text).strip()

    return text
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\leixu\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [40]:

```
top_emails['cleaned_text'] = top_emails['text'].apply(preprocess_text)
```

In [41]:

```
all_text = ' '.join(top_emails['cleaned_text'])
```

In [42]:

```
words = all_text.split()
word_counts = Counter(words)
most_common_words = word_counts.most_common(10)
print("Most common words in the filtered emails:")
for word, count in most_common_words:
    print(f"{word}: {count}")
```

```
Most common words in the filtered emails:
million: 779
transactions: 520
firm: 514
round: 472
capital: 414
partners: 357
services: 349
number: 329
power: 325
value: 320
```

K-means Analysis

In [43]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
```

In [44]:

```
vectorizer = TfidfVectorizer(max_df=0.5, min_df=2, stop_words='english')
X = vectorizer.fit_transform(top_emails['cleaned_text'])
tfidf_scores = X.toarray().sum(axis=1)
```

In [45]:

```
top_emails['tfidf_score'] = tfidf_scores
```

In [46]:

```
# top_emails.head(3)
```

In [47]:

```
num_clusters = 3
kmeans = KMeans(n_clusters=num_clusters, random_state=40)
kmeans.fit(X)
```

Out[47]:

```
KMeans
KMeans(n_clusters=3, random_state=40)
```

In [48]:

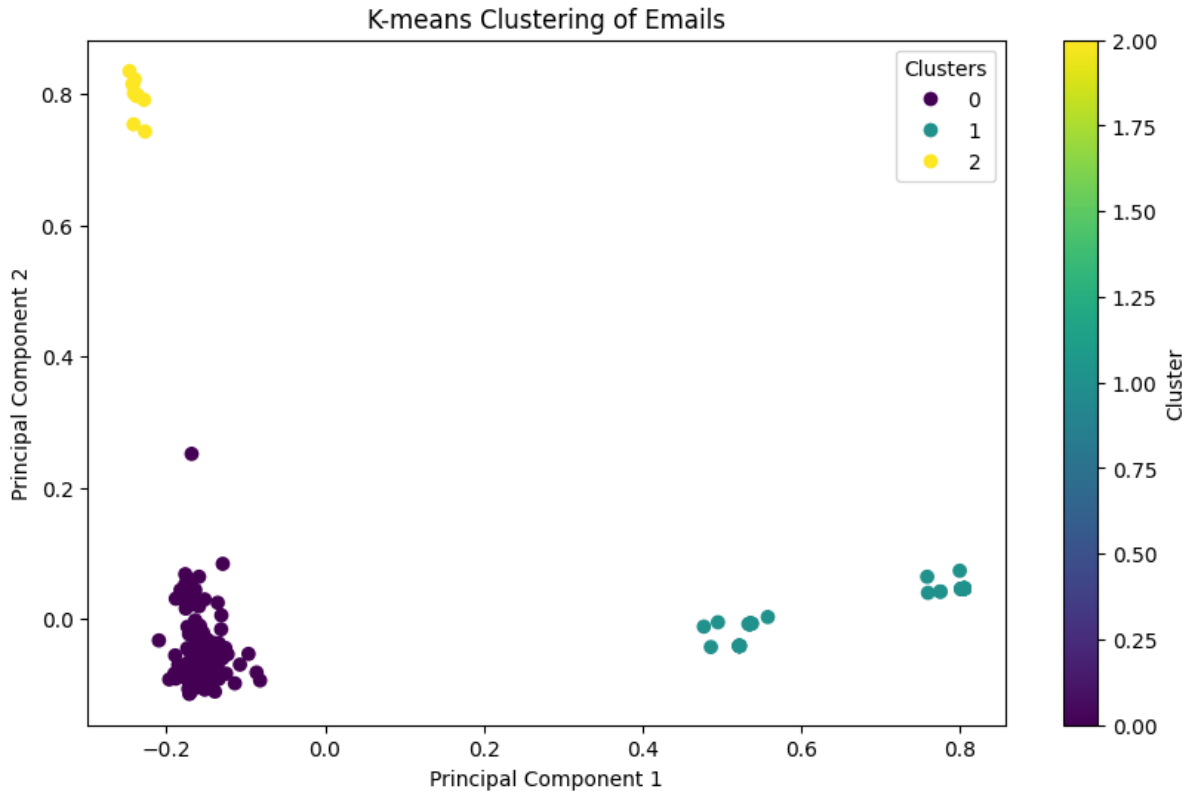
```
top_emails['cluster'] = kmeans.labels_
```

In [49]:

```
pca = PCA(n_components=2)
reduced_data = pca.fit_transform(X.toarray())
```

In [50]:

```
plt.figure(figsize=(10, 6))
scatter = plt.scatter(reduced_data[:, 0], reduced_data[:, 1], c=top_emails['cluster'], cmap='viridis',
marker='o')
plt.title('K-means Clustering of Emails')
plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.colorbar(label='Cluster')

# Add a legend
handles, labels = scatter.legend_elements()
legend = plt.legend(handles, labels, title="Clusters")
plt.show()
```



Word Cloud Analysis

In [51]:

```
cluster_counts = top_emails['cluster'].value_counts()

# Print the count of emails in each cluster
for cluster_id, count in cluster_counts.items():
    print(f"Cluster {cluster_id}: {count} emails")
```

```
Cluster 0: 159 emails
Cluster 1: 41 emails
Cluster 2: 10 emails
```

In [52]:

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Function to generate word cloud
def generate_wordcloud(text):
    wordcloud = WordCloud(width=800, height=400, background_color='white').generate(text)
    plt.figure(figsize=(10, 5))
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis('off')
    plt.show()
```

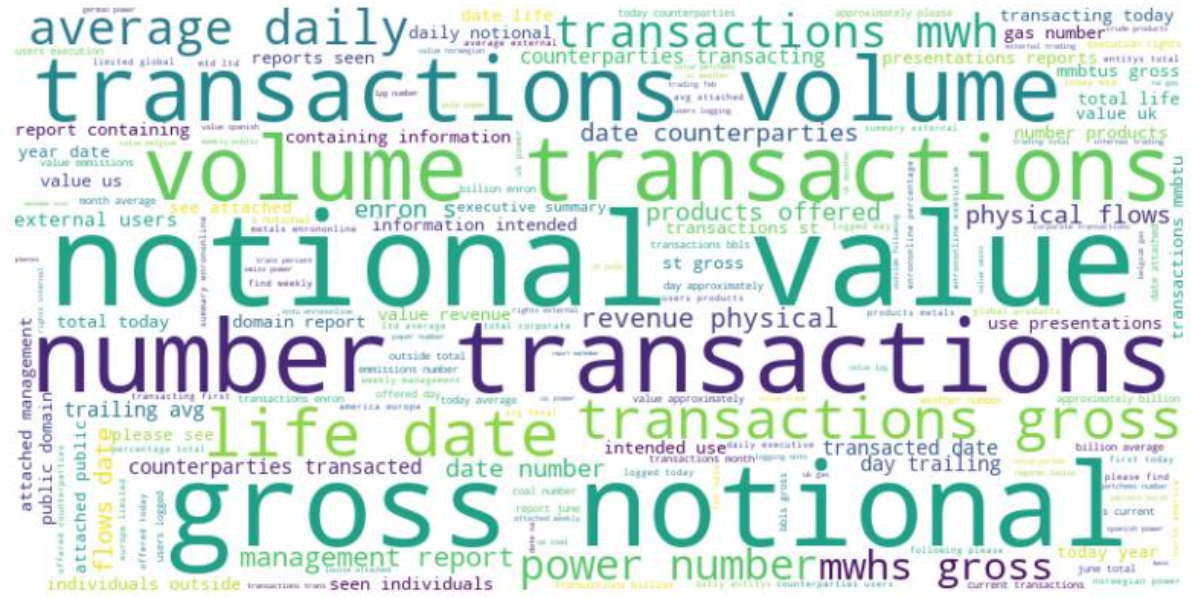
In [53]:

```
for cluster_id in range(num_clusters):
    cluster_text = ' '.join(top_emails[top_emails['cluster'] == cluster_id]['cleaned_text'])
```

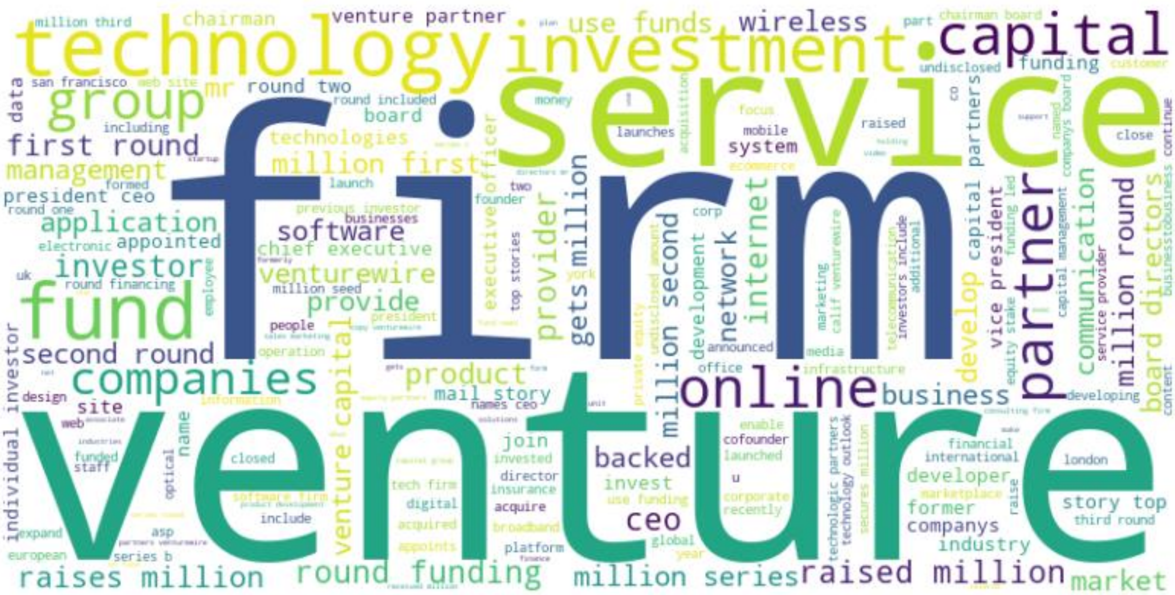
```
print(f"Word Cloud for Cluster {cluster_id}")
generate_wordcloud(cluster_text)
```

Word Cloud for Cluster 0

Word Cloud for Cluster 1



Word Cloud for Cluster 2



In [54]:

```
cluster_counts = top_emails['cluster'].value_counts()

# Print the count of emails in each cluster
for cluster_id, count in cluster_counts.items():
    print(f"Cluster {cluster_id}: {count} emails")
```

Cluster 0: 159 emails
Cluster 1: 41 emails
Cluster 2: 10 emails

In [55]:

```
target_cluster = top_emails[top_emails['cluster'] == 0]
target_cluster_2 = top_emails[top_emails['cluster'] == 2]
target_cluster_0 = top_emails[top_emails['cluster'] == 0]
```

In [58]:

```
top_52_emails = target_cluster.sort_values(by='tfidf_score', ascending=True).head(5)
```