Analysis of Enron's Email Communication Network

Parham Oghabi

Abstract

Identifying key nodes in a social network is essential for various real applications and is of huge interest to governments around the world. One such application is to assist in legal case investigations by focusing on interrogating the key nodes for efficiency and to extract the most useful information. My motivation stems from the challenge of being able to locate key nodes, who are either guilty or who can be useful interrogation subjects. In this project, I have constructed various directed weighted networks from the non-anonymized email communications of Enron Corporation to be able to identify such nodes. Enron was a company which went bankrupt in 2001 due to its fraudulent actions and this led to the investigation and conviction of its executives and employees by the Federal Energy Regulatory Commission. I demonstrate the potential of quantitative network analysis by using various centrality measures and other metrics in achieving state of the art results in locating key nodes in order to assist legal investigations and interrogations.

Table of Contents

Introduction……1

Background and Motivation……1

Project Goal……2

Literature Survey……2

Identifying a Criminal's Network of Trust……2

Exploratory Data Analysis of Enron Emails……4

Analyzing Key Mediators through Temporal Centrality Metrics……5

Methodology……6

Dataset and Pre-processing……6

Algorithms and Tools……8

Results……9

Complete Email Network……9

Enron Email Network……13

150 Employee Email Network……15

Discussion and Improvements……16

Conclusion……17

References……19

1. Introduction

This report summarizes the motivation, design, implementation, and quantitative evaluation of an email network analysis as part of my COMPGW01 coursework. It also introduces and critically discusses 3 relevant research literatures. The report concludes with a discussion of improvements and future work.

A. Background and Motivation

Being able to identify key nodes in a network is of huge interest to government bodies in various applications. One such application is to assist different government bodies in legal case investigations of corporations by focusing on interrogating key people for efficiency and to extract the most useful information. Key nodes can be people who are either guilty or who can be useful interrogation subjects. This quantitative network analysis of social network data can be used to speed up the investigation, instead of manually analyzing each person/node, and can also be used as evidence in court. Tracing ties and evidence from such a large network to identify suspects is very difficult to perform manually due to the large number of nodes and the complex relationships between them.

In this project, the email communications of Enron Corporation was used as the dataset. Enron was an American energy supplier company and one of the world's largest electricity and natural gas provider. In the 1990s, Enron opened subsidiaries in other countries and shifted its focus to being an investment firm in order to make more profit off the margins of trading energy. By December 2001, Enron had filed for bankruptcy and it was revealed that the corporation performed fraudulent accounting activities by creating complex offshore partnerships.

The Federal Energy Regulatory Commission investigated the issue and many of Enron's executives and employees were convicted and handed long prison sentences for money laundering. As part of the investigation, the complete email communications and inboxes of 150 employees in Enron were subpoenaed by the U.S. government and was later released to the public. This dataset was used for my project and its pre-processing and details are described in the "Methodology" section.

B. Project Goal

My goal is to perform a quantitative email network analysis of Enron's email dataset and demonstrate its potential in identifying key nodes in a social email network. These key nodes can be either guilty people or useful interrogation subjects, which can be used to assist in legal case investigations of corporations by government bodies.

This project shows the state of the art results achieved in locating the key nodes and the results are evaluated by comparing the nodes to the people who were eventually convicted and sentenced at Enron since the list of convicted people was released by the U.S. government and by various news agencies.

2. Literature Survey

This section introduces and critically discusses 3 relevant research literatures.

A. Identifying a Criminal's Network of Trust [1]

i. Literature Summary

This paper attempts to identify criminals' ties and connections within a network by applying a new algorithm on sub-networks of the original Enron email dataset that I have used as well. Out of all the original Enron emails, only the ones which included 1 or 2 blind carbon copy (BCC) recipient emails were used (ie. the email dataset was reduced substantially). This was done because the 'to' and 'cc' recipients are visible to everyone who receives that email, but 'bcc' recipients remain hidden. This creates a network of trust between the sender and the BCCed recipients. Groups of criminals who are linked in the trust network are then isolated and shortest path concepts (pseudo-code in the paper) along with centrality metrics are used to group and identify the criminal's connections in the isolated group.

In the paper, the node with the highest eigenvector centrality is called "Most Influential" (MI) and the node with the highest betweenness is the "Middleman" (MM). Essentially, the shorter the path between then criminal and MI, the less isolated the criminal and the closer the criminal is to the MM, the higher the probability that the criminal uses the MM to influence other nodes.

Two networks were constructed, one called the 1-BCC (emails in which 1 recipient was

BCCed) and 2-BCC (emails in which 2 recipients were BCCed). An edge exists from node A to B if A sent an email which had B as the BCC recipient. Two investigative scenarios were then described:

- In the first case, the investigator knows the key criminals and wishes to identify its connections and ties. Therefore, each known criminal's subnetwork is then analyzed to find its connections.

- In the second case, the author assumes the investigation has just started and that no criminals are known and the goal is to find a sparse subgraph containing some suspicious nodes to start an investigation. In the paper, since Enron's crime was money laundering, they assumed the starting criminals to be financial managers and started from there.

ii. Critical Analysis

The paper was interesting and their approach to find the criminal and its connections through the blind carbon copies was creative and unique. They have also provided full details in order to re-create their experiments. However, there are a number number of problems that I thought of when I read this paper.

Firstly, they have reduced the dataset substantially from 1,887,305 emails to 21,000 emails (the 1-BCC and 2-BCC emails). That means they are only using 1% of the communications which indicates that the probability that they might miss a criminal is high. Indeed in Enron's case, they missed out Kenneth Lay, who was Enron's first CEO and was also convicted in reality.

Secondly, the use of unweighted edges means that they are considering multiple emails between senders and BCCed recipients as only one email. Using the edges helps in identifying the criminals.

Thirdly, their first investigative scenario (initially knowing the criminals) might include an invalid assumption as in most cases the criminals are not known in advance. In addition to that, they reported that their algorithm did not work well on the second case (initially not knowing the criminals) and that they were not able to figure out the important criminals.

In general, their approach narrows down the dataset and will miss out on criminals in larger datasets. In an investigation scenario, one would rather have more suspects and criminals to investigate than having fewer ones.

B. Exploratory Data Analysis of Enron Emails [2]

i. Literature Summary

This paper attempts to identify if top Enron employees had incriminating evidence in their emails and see if any unusual patterns exist in the months leading to Enron's bankruptcy. The author has approached this from an NLP perspective and only utilizes the "sent" folder of each employee and ignores the other folders due to the large size of the dataset. Topic modelling with Latent Dirichlet Allocation (LDA) and sentiment analysis was used.

LDA with 4 topics was performed on the dataset, specifically on each emails body only, excluding the headers. The top 10 words in each topic was recorded. Topic 1 contained meeting related words, topic 2 contained business terms and legalities, topic 3 contained words relevant to Enron's core business like "gas" and "power", and topic 4 contained casual meeting related terms. The distribution of each topic was then analyzed over time with a monthly frequency. Meeting related topics increase and business related topics decrease as the graph approached the time period close to Enron's bankruptcy and scandal. This makes sense as executives were holding meetings trying to hide Enron's debt and business was decreasing.

For the sentiment analysis portion, unsupervised learning was used to tag the emails with a sentiment score since the emails are unlabelled. The email body is split into sentences and fed to various subjective, pattern-based, and polarity classifiers based on a rule-based pipeline. There is a spike in the sentiment score (positive) during the period when Enron was booming and the stock was at an all-time high.

ii. Critical Analysis

It was interesting to see the problem of identifying key scandalous nodes from a different perspective like NLP. One would expect that NLP processing of the emails would be more accurate in detecting key nodes compared to social network analysis since the contents of the emails are being examined but this wasn't the case in this paper.

Even though the topic modelling with Latent Dirichlet Allocation was pretty simple, it

worked relatively well. I think the sentiment analysis did not work as well because the designed classifier was not very well and not enough data was used to train it (only the "sent" folder of each user was used). In order to see the true power of NLP, better data pre-processing and neural pre-trained models should be used to classify them.

With the results in this paper, one cannot rely certainly on the models to reveal incriminating evidence for each node.

C. Analyzing Key Mediators through Temporal Centrality Metrics [3]

i. Literature Summary

This paper identifies important individuals in the Enron corporate email dataset using various centrality metrics in a dynamic representation of the email network topology. Their aim is to show that their temporal analysis of networks and their novel temporal centrality metrics provide a more accurate identification of important people compared to traditional state-of-the-art static network representations and methods.

Rather than traditional static and aggregated network topologies to represent the email dataset, the authors chose a dynamic topology which is an ordered sequence of snapshots of the network topology as it changes over time. The state of the network topology at each time instant is calculated by aggregating all the edges that appear inside a certain time window. In our email network case, window sized of 24 hours were used. This means that for each snapshot, the graph has an undirected edge between two individuals if an email was exchanged between them in that 24 hour window. The temporal graph was constructed of 24-hour ordered graph snapshots from 1999 to 2002.

Temporal betweenness and closeness centrality was then calculated for each node using temporal shortest paths. Temporal betweenness centrality of node i is the fraction of temporal shortest paths that pass through node i. This can be thought of as individuals who who act as key mediators between the most communication paths over time. On the other hand, temporal closeness centrality is defined as the measure of the geodesic distance between two nodes. This is the temporal shortest path between nodes, which is a measure of how early a source node can delivery a message to all other nodes.

After ranking the nodes by temporal betweenness and closeness, top trading executives were exclusively ranked as the top nodes. Comparing the results with media reports, there

is a correlation between these traders and the top two bonuses received.

ii. Critical Analysis

Using temporal networks along with temporal betweenness and closeness centrality measures was a novel idea and based on the results reported in their paper, this was one of the most accurate results I have read while researching the topic of detecting important individuals in the Enron email dataset.

In the paper, the authors used a window size of 24 hours but it would be interesting to see results with window sizes of 48 and 72 hours since not all email replies fall within that 24 hour window. In addition to that, they also used an undirected graph. It would make more sense to use a directed-weighted temporal graph where the weight is the number of emails sent from node A to node B in the window time period. Also, it seems as though the author only used half of the email dataset (250,000 emails) instead of the whole 500,000 emails. Using the whole dataset might have revealed more important nodes.

The results stated that temporal centrality metrics consistently showed energy traders as the most important nodes in Enron whereas static centrality metrics showed secretaries and managing directors. In order to verify their results, I tried to roughly re-create their experiment but I could not access their dataset since it was inaccessible. If their results are correct and verified, then it would be interesting to see temporal centrality metrics applied in other areas as well.

3. Methodology

This section introduces the dataset, its pre-processing techniques, and the algorithms used to achieve the results.

A. Dataset and Pre-processing

In this project the dataset used is Enron's email communication network among 150 of its employees, who were mostly senior managers, from 1999 to 2002. [4] This was released to the public by the Federal Energy Regulatory Commission whilst investigating the Enron scandal. There are a total of about 500,000 emails. For each user, the entire state of their custom mailbox and folders were preserved and therefore contain various folders such as "sent", "inbox", and others.

The emails were all in MIME format and an example is shown in the image below.

```
Message-ID: <20702230.1075840284123.JavaMail.evans@thyme>
Date: Fri, 21 Apr 2000 06:23:00 -0700 (PDT)
From: rosalee.fleming@enron.com
To: expense.report@enron.com
Subject: Approval of Cindy Olson's Expense Report
Cc: kayla.ruiz@enron.com
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
Bcc: kayla.ruiz@enron.com
X-From: Rosalee Fleming
X-To: Expense Report
X-cc: Kayla Ruiz
X-bcc:
X-Folder: \Kenneth_Lay_Dec2000\Notes Folders\'sent
X-Origin: LAY-K
X-FileName: klay.nsf

Kenneth Lay has approved Cindy Olson's expense report, which is attached.
```

Figure 1: Screenshot showing the MIME format of the Enron emails.

Pre-processed csv files were already on Stanford's SNAP dataset but they were not used since the node IDs were not matched to the employee emails and therefore my results could not be evaluated with the list of convicted people in the end. Furthermore, it seems as though only a subset of the emails were used by SNAP and a clear description of how they parsed the emails was not mentioned.

As a result, a custom python script was written to loop over all the original directories and sub-directories to parse each email in order to construct an aggregated directed weighted graph. Each unique email was assigned an ID and each edge represents an email being sent from node A to node B while its weight represents the number of emails that was sent from A to B in the 1999 to 2002 time period. Using this script, 3 various email networks were constructed to compare the results extracted from each topology:

- All the emails were used to create a complete network.
- Only emails exchanged between Enron employees, with emails including "@enron.", were used.
- Only emails exchanged between the 150 Enron employees whose email inboxes were subpoenaed were used to build the graph.

Most research papers did not use all the emails in order to narrow down their results and also because working with the whole dataset was very slow since it is computationally expensive but it shows a more complete picture. The python script wrote the edges, the

node IDs, and their respective email addresses to different csv files in order to be exported to Gephi for visualization and metric calculation purposes.

B. Algorithms and Tools

Gephi

Gephi was used to load the edges and nodes csv file for network layout, visualization, and to calculate network metrics. The Force Atlas layout was chosen for small graphs and Openorb layout was chosen for larger graphs with more than 10,000 nodes. The node sizes are proportional to their PageRank value, the colour is based on the community they belong too, and the thickness of the directed outgoing edges is proportional to the number of emails sent.

PageRank

Since the graph is directed, the PageRank algorithm was chosen to try to locate important nodes. The teleportation probability was chosen as 15%. An employee will have a high rank if they receive emails from other high-ranked employees. This might be able to locate the executive directors in the nodes who receive a lot of emails from managers.

Betweenness Centrality

Betweenness centrality of a node is measured using the number of shortest paths between all pairs of nodes in the network that pass through that node. Essentially, it finds the nodes which lie between the most paths of communication. An employee will have a high betweenness centrality if that employee appears consistently in the email proximity of other employees. These nodes can be mediators who fall into middle management and report to senior managers.

Closeness Centrality

Closeness centrality of a node is the inverse of the sum of shortest path lengths between that node and all other nodes. If there exists a pair of nodes that do not have a path between them, then the closeness of those nodes are zero. An employee will have a high closeness centrality if they are able to reach a large portion of the network quickly.

Degree Distribution

The degree distribution of each node was plotted. For each node, the degree was chosen to be the total number of received and sent emails by that node. It is a measure of that node's immediate network. As one would expect, it should follow a power law distribution since there are few top nodes in a corporation which have a high degree.

Community Detection

The Louvain algorithm was used in Gephi for community detection since the graph was very large so the Girvan-Newman algorithm would be very inefficient. This was done to show each community within the corporation and the most important node in each community.

4. Results

As stated in the data pre-processing section, 3 networks were constructed using Enron's email dataset in order to compare the results.

A. Complete Email Network

This is the complete network taking into account all the emails in the Enron dataset. There are a total of 79,750 nodes/emails and 3,130,272 emails sent. This was a very large dataset and computations performed on this graph was very slow. A visualization of the network is shown on the next page using the Openorb layout. The node sizes are proportional to their PageRank value, the colour is based on the community they belong too, and the weight of the directed outgoing edges is proportional to the number of emails sent. Red arrows indicate the important nodes.

Looking at the layout, we can see that their are about 8 dominant communities. The network has a modularity of 0.66 which indicates significant community structure.

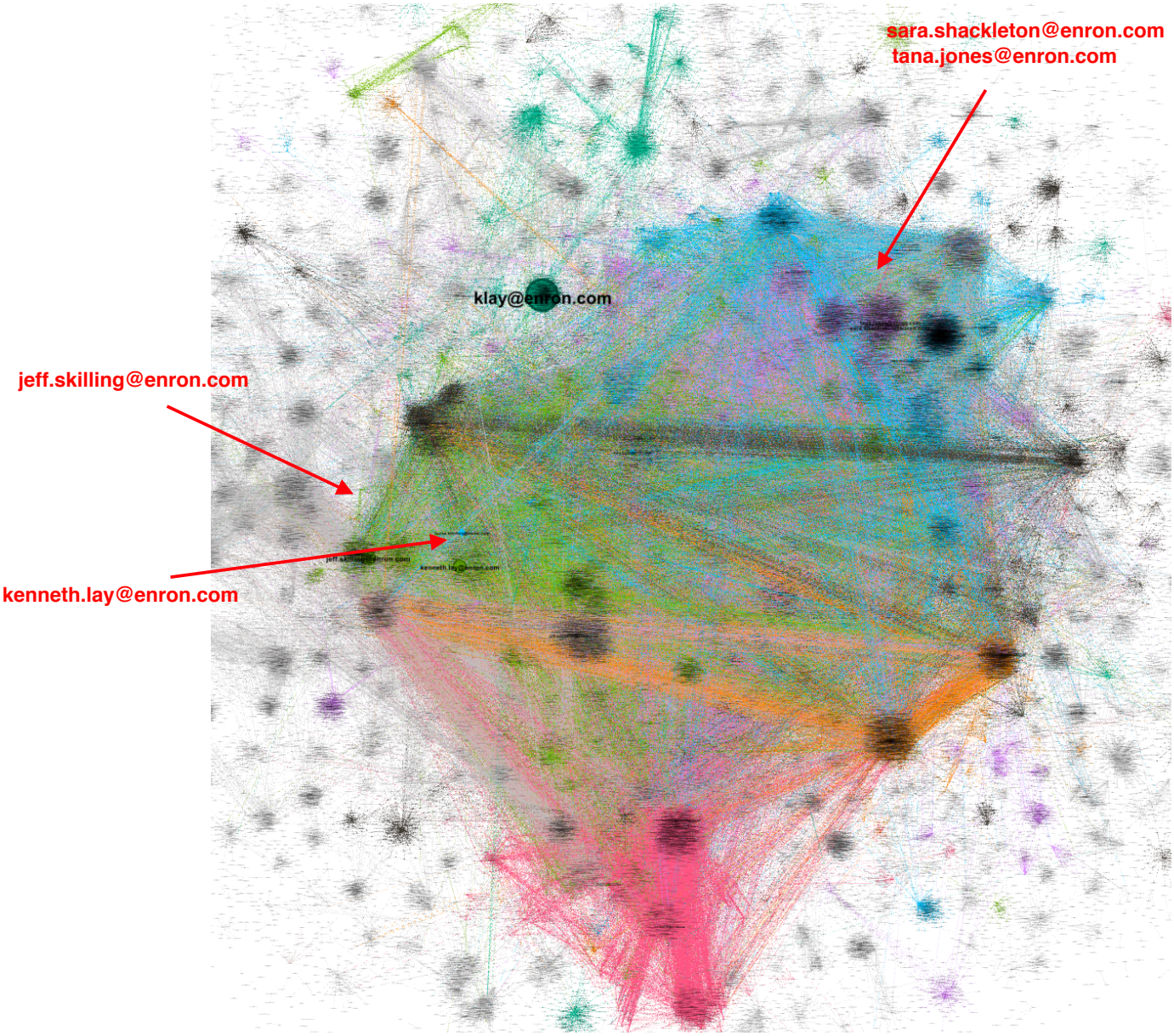

Figure 2: Layout of the complete email network showing the various communities along with the nodes with the highest PageRank values.

In the complete email network visualization, we can see that klay@enron and kenneth.lay@enron.com have a high PageRank value and that they belong to the same person. Let us extract and visualize the email network of Kenneth Lay's inbox to see if we can find anything interesting. This layout is shown on the next page.

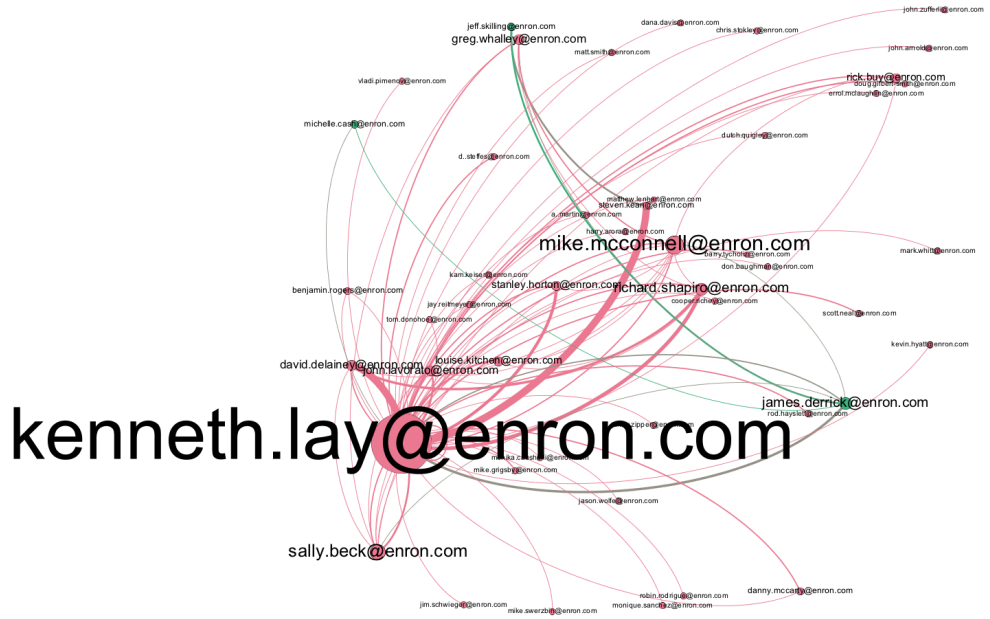

Figure 2: Layout of the email network of Kenneth Lay's inbox.

We can see some interesting communications going on between Kenneth Lay and people like Mike Mcconnell, John Lavorato, Louise Kitchen. Indeed after searching, Kenneth Lay was the former CEO of Enron and was convicted and sentenced to prison for money laundering. In addition to that, many of the names mentioned above are top executives and were also sentenced.

Using excel, I plotted the in-degree distribution and out-degree distribution for the complete email network and we can see that it follows a power law distribution as expected. Kenneth Lay was the node with the highest in-degree. This makes sense since the CEO overseas a corporation's operations and will receive many emails. The power law exponent is about 2.2.

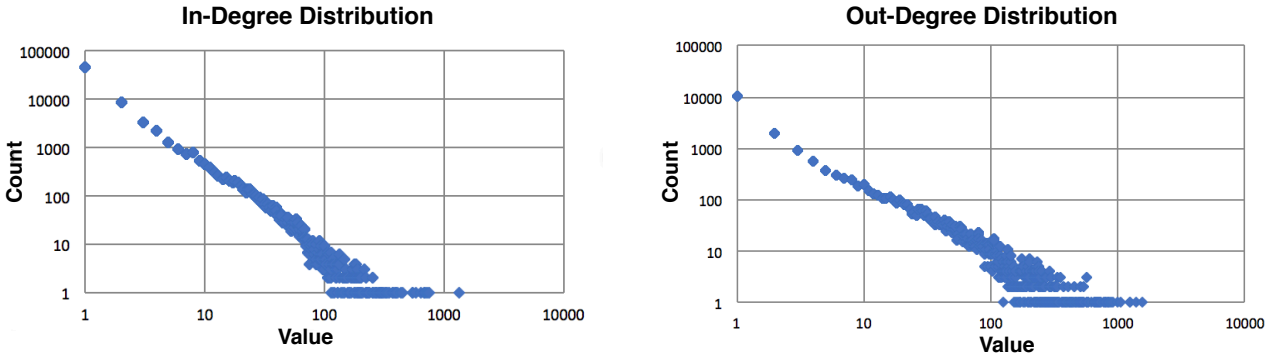

Figure 3: The in-degree and out-degree distributions of the complete email network showing a power law distribution.

If we sort the nodes by their PageRank (eigenvector) and betweenness centralities, we can we can immediately see that many of the top ranked nodes that are in common in both lists were individuals who were convicted for securities fraud.

PageRank

Id	Label
35797	klay@enron.com
2337	jeff.skilling@enron.com
1087	kenneth.lay@enron.com
398	sara.shackleton@enron.com
522	tana.jones@enron.com
560	louise.kitchen@enron.com
150	jeff.dasovich@enron.com
426	gerald.nemec@enron.com
130	mark.taylor@enron.com
770	don.baughman@enron.com
172	sally.beck@enron.com
1230	vince.kaminski@enron.com
571	j.kaminski@enron.com
934	benjamin.rogers@enron.com
2	john.lavorato@enron.com
42120	vkamins@enron.com

Betweenness

Id	Label
150	jeff.dasovich@enr...
1087	kenneth.lay@enro...
172	sally.beck@enron....
2337	jeff.skilling@enron...
426	gerald.nemec@enr...
1230	vince.kaminski@en...
61	outlook.team@enr...
522	tana.jones@enron....
105	susan.scott@enron...
398	sara.shackleton@e...
2	john.lavorato@enr...
560	louise.kitchen@enr...
130	mark.taylor@enro...
571	j.kaminski@enron....
934	benjamin.rogers@...
3842	rosalee.fleming@e...

Figure 4: Table showing Enron's employees ranked by eigenvector and betweenness centralities. The red borders indicate guilty individuals while the green border indicates useful interrogation subjects who were not guilty.

The common names of the important nodes in both sorted lists are outlined with a red or green border. The betweenness centrality measures the number of shortest paths through a node, which can be thought of in a way as the number of emails which pass through that person. People who have a high betweenness centrality can indicate people who were involved in the core of the scandal or had some idea of what was going on, whether guilty or not.

On the other hand, nodes with a high PageRank value are nodes who received a lot of incoming emails and one can assume that these are executive directors or managers of important departments. As we can see in the PageRank list, the top 3 nodes are the individuals who were heavily involved in Enron's scandal. Kenneth Lay was Enron's former CEO who was convicted and Jeffrey Skilling was the chief operating officer of Enron and had complete control of Enron's financial activities and statements. Both were given long-term prison sentences for money laundering.

In the two sorted lists above, the individuals with a red border around them are those who were convicted and found guilty while those with a green border were useful interrogation subjects and were starting to become aware of the fraudulent activities taking place at Enron. For example, Vince Kaminski was a Managing Director at Enron and repeatedly raised strong concerns to the financial practices of Enron's CFO who was hiding Enron's debt.

Looking at media reports, we can also evaluate the important nodes found. Below is an article that CNN has published about Enron and the individuals who were awarded top bonuses during the scandal period. It mentions that two of Enron's traders made the top bonuses and if we compare that to the individuals we ranked in our lists above, we can see that we have also detected these important nodes as suspicious.

Top bonuses awarded

John Lavorato: $5 million
Louise Kitchen: $2 million
Jeffrey McMahon: $1.5 million
James Fallon: $1.5 million

Figure 5: A snippet from a CNN article on Enron stating that two of Enron's traders made the top bonuses during the scandal period. [5]

B. Enron Email Network

This is the email network only containing emails with "@enron.". There are a total of 21,104 nodes/emails and 2,515,791 emails sent. Essentially, this was done in order to remove some of the spam emails and the personal emails in the dataset to produce more accurate and clearer results. A visualization of the network is shown below.

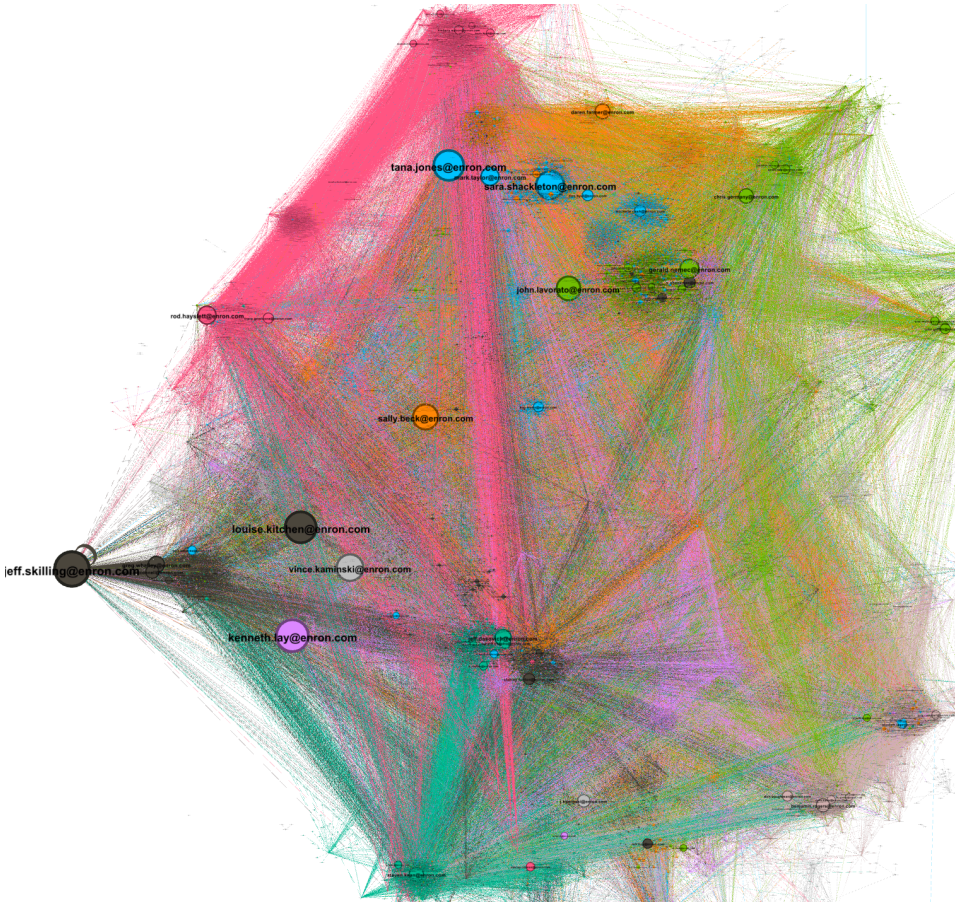

Figure 6: Layout of the Enron email network showing the various communities along with the nodes with the highest PageRank values.

As in the previous layout, node sizes are proportional to their PageRank value, colours are based on communities, and the weights of edges are proportional to the number of emails sent. Compared to the previous layout, we can see that nodes with a higher PageRank value start to emerge more clearly. This is because many of the personal and spam emails are removed since we are only using Enron's domain email addresses. Individuals like Jeff Skilling, Louise Kitchen, Kenneth Lay, and John Lavorato are much more clearer and ranked higher in this image. This reinforces the results that we obtained in the previous network.

The in-degree and out-degree distribution is a power law distribution as before and as expected with a power law exponent of around 2.2.

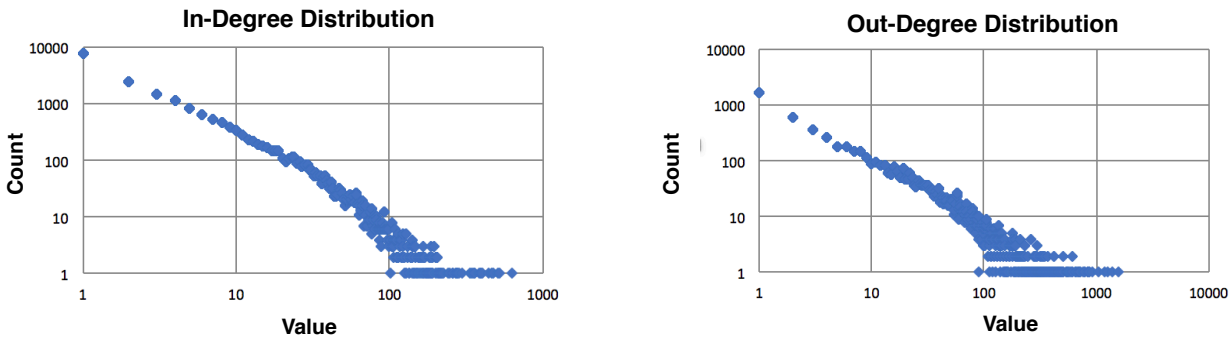

Figure 7: The in-degree and out-degree distributions of the Enron email network showing a power law distribution.

Sorting the nodes in terms of their PageRank value and their betweenness and closeness centralities produces the two lists below.

PageRank

Id	Label
1937	jeff.skilling@enron...
915	kenneth.lay@enro...
464	louise.kitchen@enr..
428	tana.jones@enron...
304	sara.shackleton@e..
997	vince.kaminski@en..
127	sally.beck@enron....
2	john.lavorato@enr...
332	gerald.nemec@enr..
99	mark.taylor@enro...
907	rod.hayslett@enro...
109	jeff.dasovich@enr...
875	greg.whalley@enro..
1101	daren.farmer@enr...
438	chris.germany@en...
117	steven.kean@enro...
798	benjamin.rogers@...
465	mike.mcconnell@e...

Betweenness and Closeness

Id	Label
127	sally.beck@enron....
915	kenneth.lay@enro...
1937	jeff.skilling@enron...
44	outlook.team@enr...
997	vince.kaminski@en...
428	tana.jones@enron....
464	louise.kitchen@enr...
109	jeff.dasovich@enr...
2	john.lavorato@enr...
689	david.forster@enr...
216	shelley.corman@e...
907	rod.hayslett@enro...
304	sara.shackleton@e...
332	gerald.nemec@enr...

Figure 8: Two lists showing Enron's employees ranked by eigenvector (left) and betweenness and closeness centralities (right).

Compared to the previous sorted lists, we can see that some of the noise is removed and the two traders with the top bonuses, Louise Kitchen and John Lavorato, are moved up the list and we can consider them to be more important. We also see a new name appear at the bottom of the PageRank list. Mike Mcconnell was the executive VP of technology at Enron Online which was Enron's subsidiary for trading. Our overall results are still fairly consistent with our initial ones but this network provides more accurate results. In the next section, we will narrow down our dataset even more to only include emails from the 150 employees whose email inboxes were subpoenaed to see if we will get similar results.

C. 150 Employee Email Network

This is the email network only containing emails exchanged between the 150 Enron employees. There are a total of 146 nodes/emails and 69,957 emails sent. A visualization of the network is shown below.

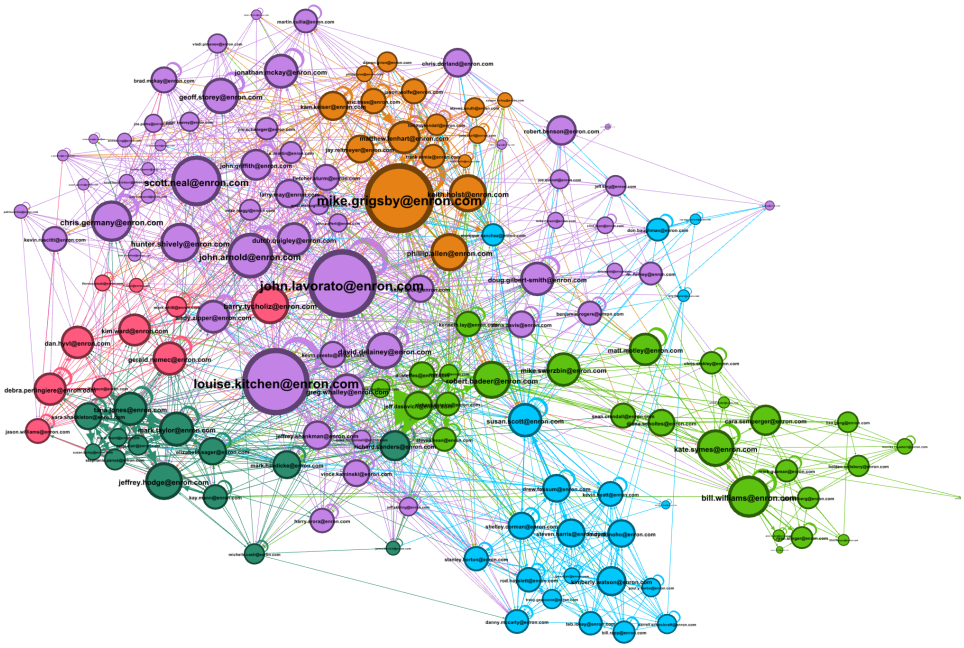

Figure 9: Layout of the 150 email network showing the various communities along with the nodes with the highest PageRank values.

The visualization of this network is the most clear out of the 3 networks created from the Enron email dataset. Applying the Louvain community detection algorithm produces interesting results. Most of the traders fall into the same community, such as Louise Kitchen, John Lavorato, John Arnold, and Scott Neal.

PageRank

Id	Label
6	mike.grigsby@enron.com
1	john.lavorato@enron.com
36	louise.kitchen@enron.com
31	scott.neal@enron.com
32	john.arnold@enron.com
125	bill.williams@enron.com
91	chris.germany@enron.com
14	hunter.shively@enron.com
21	barry.tycholiz@enron.com
0	phillip.allen@enron.com
2	keith.holst@enron.com

Betweenness and Closeness

Id	Label
1	john.lavorato@enron.com
125	bill.williams@enron.com
18	susan.scott@enron.com
36	louise.kitchen@enron.com
6	mike.grigsby@enron.com
28	sally.beck@enron.com
22	jeff.dasovich@enron.com
31	scott.neal@enron.com
33	kim.ward@enron.com
91	chris.germany@enron.com
98	kenneth.lay@enron.com

Figure 10: Table showing Enron's employees ranked by eigenvector in the left list and betweenness and closeness centralities in the right list.

Sorting the nodes in terms of their PageRank value and their betweenness and closeness centralities also rank Louise Kitchen and John Lavorato higher than before. It seems as though the more we refine the dataset, the better the achieved results. However, at the same time other important nodes like Jeffrey Skilling and Greg Whalley start to become less visible.

Looking more into depth, I realized that this is caused due to the following issues. Top executives like Kenneth Lay, the former CEO, had multiple email addresses in both inside and outside of Enron which is why he appears to be ranked lower in these lists because many of his emails weren't considered. Also, Jeffrey Skilling's inbox was not part of the 150 inboxes subpoenaed which is why his email address doesn't appear. Even though his email address appears in the emails, here we are only considering emails exchanged among the 150 employees whose inboxes were subpoenaed.

4. Discussion and Improvements

From the results achieved in the three networks created above, we can confirm that network analysis can successfully identify guilty individuals or useful interrogation subjects with a relatively high accuracy in order to assist government investigations.

However, one exception did occur which was interesting and is worth discussing. Our results did not manage to detect Andrew Fastow at all, who was the Chief Financial Officer of Enron and was convicted as well. One hypothesis is that he was completely aware of the fraudulent accounting activities he was performing from the beginning and therefore he did not communicate via e-mail very often and our network analysis cannot detect him through the metrics discussed in lecture. Neither the aggregate weighted directed graph that we constructed here nor the temporal undirected graph discussed in literature survey 3 managed to detect Andrew Fastow. However, the 1-BCC and 2-BCC email network discussed in literature survey 1 did manage to detect him. This is because in the aggregate and temporal graphs, it depends on

who Andrew Fastow sends or receives emails to whereas in the BCC email network, it only depends on who BCCs Andrew Fastow which is completely out of his control

Combining the results from the aggregate, temporal, and BCC email networks, one can see that the top executives in a corporation form a tight clique to oversea the operations of a corporation and to share any major events going on.

Furthermore, there are some improvements which can enhance the results in order to give more accurate individuals.

Better pre-processing of emails will definitely produce more accurate results. For example, a large number of duplicate emails exist. When individual A sends an email to individual B, that email can be in the "Outbox" folder of A and "Inbox" folder of B. Furthermore, a large number spam emails existed in the dataset. One can use a pre-trained machine learning spam classifier, such as a simple Naive Bayes model or more complex neural networks, to remove the spam messages before constructing the email graph. The two issues mentioned above cause the dataset to be very large and very difficult to work with. For example, the dataset I used to construct the whole email network was 2.7GB. Another issue was that many of the employees have multiple email addresses, both internally inside and outside Enron's domain. It would be much easier to have a complete list of the email addresses used by an employee in order to aggregate their multiple email addresses into one when constructing the email graph. There are many research papers on data cleaning strategies for organizational emails. [6]

Another improvement is to combine network analysis with natural language processing to analyze the body of the emails instead of only the headers. The text included in the emails can definitely show signs during Enron's period of distress and bankruptcy and might even include some content to use as incriminating evidence for the guilty people.

5. Conclusion

The consistent results shown in this paper from the 3 datasets confirm the project goal and the possibility of network analysis in identifying guilty people and useful interrogation subjects in a social email network in order to assist in legal case investigations of corporations by government bodies. The results were evaluated by comparing the nodes with the publicly released list of people who were convicted and sentenced and also with the individuals mentioned in media articles such as CNN.

From a wider perspective, network analysis is of huge interest to the government in many other applications. Another major use of it in recent years has been in the field of terrorism to detect terrorists and their connections and ties. Network analysis can speed up the process of these applications instead of the manual auditors currently used by governments.

References

[1] Magalingam, P., Rao, A., & Davis, S. (2014). Identifying a Criminals Network of Trust. 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems.

[2] (n.d.). Retrieved December 14, 2017, from https://www.stat.berkeley.edu/~aldous/Research/Ugrad/HarishKumarReport.pdf

[3] Tang, J., Musolesi, M., Mascolo, C., Latora, V., & Nicosia, V. (2010). Analysing information flows and key mediators through temporal centrality metrics. Proceedings of the 3rd Workshop on Social Network Systems

[4] (n.d.). Retrieved December 12, 2017, from https://www.cs.cmu.edu/~enron/

[5] Enron paid hefty bonuses before bankruptcy. (n.d.). Retrieved December 18, 2017, from http://edition.cnn.com/2002/LAW/02/09/enron.bonuses/

[6] Zhou, Y.,Goldberg, M., Ismail, M.,Wallace, W. (2014). Strategies for cleaning organizational emails with an application to enron email dataset