INDEX

1.0 GENERAL INFORMATION

1.1 Problem Statement:

To analyse the Enron dataset, build a hierarchy structure based on the email conversations and predict the involvement of individuals in the fraudulent activities of the Enron Corp.

1.2 About Enron Corp.:

Enron was formed in 1985 following a merger between Houston Natural Gas Co. and Omaha-based InterNorth Inc. Following the merger, Kenneth Lay, who had been the chief executive officer (CEO) of Houston Natural Gas, became Enron's CEO and chairman, and quickly rebranded Enron into an energy trader and supplier. Deregulation of the energy markets allowed companies to place bets on future prices, and Enron was poised to take advantage.

The era's regulatory environment also allowed Enron to flourish. At the end of the 1990s, the dot-com bubble was in full swing, and the Nasdaq hit 5,000. Revolutionary internet stocks were being valued at preposterous levels and consequently, most investors and regulators simply accepted spiking share prices as the new normal.

Enron participated by creating Enron Online (EOL), an electronic trading website that focused on commodities in Oct. 1999. Enron was the counterparty to every transaction on EOL; it was either the buyer or the seller. To entice participants and trading partners, Enron offered up its reputation, credit, and expertise in the energy sector. Enron was praised for its expansions and ambitious projects and named "America's Most Innovative Company" by *Fortune* for six consecutive years between 1996 and 2001.

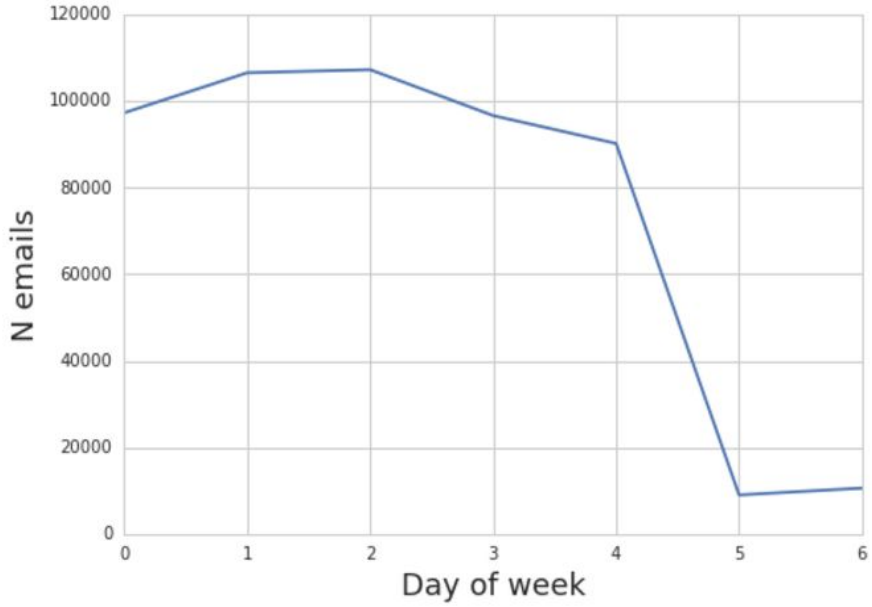
By mid-2000, EOL was executing nearly $350 billion in trades. At the outset of the bursting of the dot-com bubble, Enron decided to build high-speed

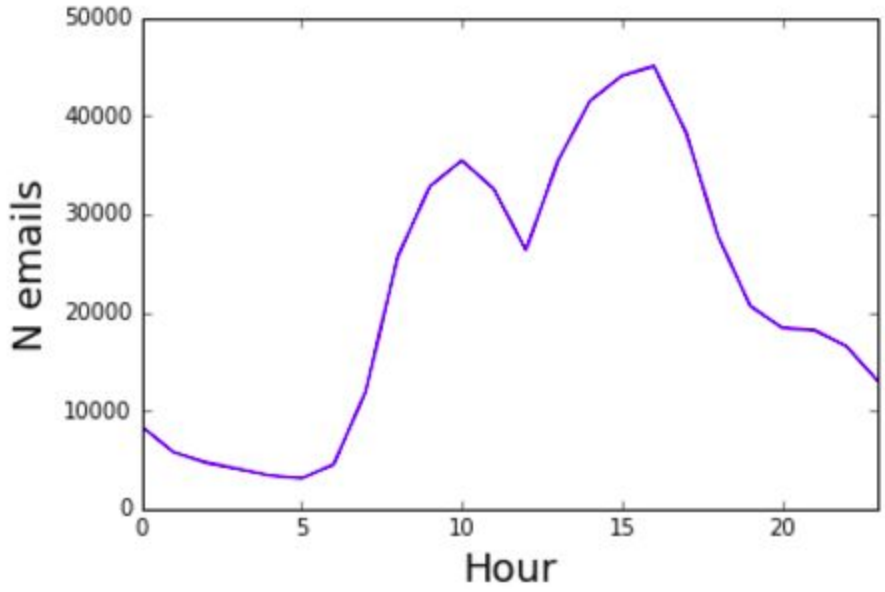
broadband telecom networks. Hundreds of millions of dollars were spent on this project, but the company ended up realizing almost no return.

When the recession began to hit in 2000, Enron had significant exposure to the most volatile parts of the market. As a result, many trusting investors and creditors found themselves on the losing end of a vanishing market cap. Hence cashing out their stocks and when not much investment was available the stocks took a dive and Enron had to file for bankruptcy which created suspicion.

2.0 Preliminary analysis of the Data:

Though not following a specific approach, data was analysed to find patterns and anomalies.

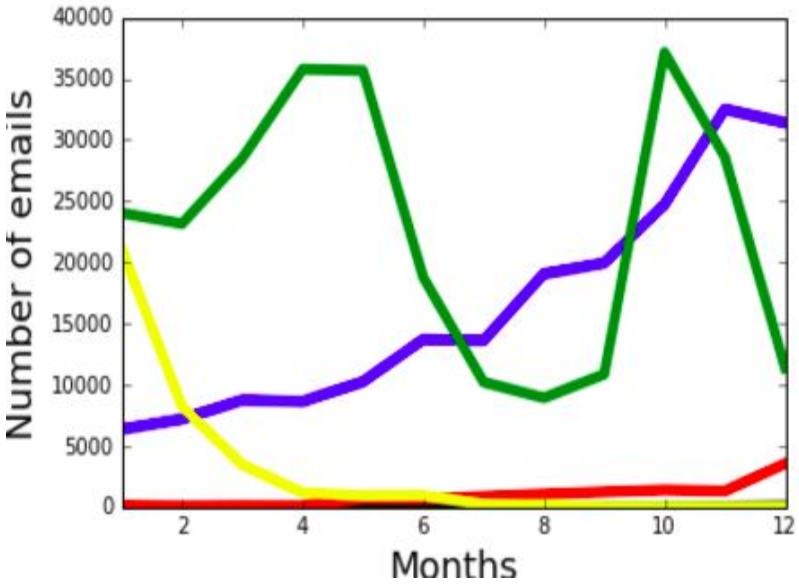




As Enron Corp. was also one of the companies that operated through stocks and hence as a result of which the image above shows a decrease in the number of emails sent during the weekends. Almost for all the weeks the employees behaved in similar fashion.

Also for a day, on an average, the number of emails increased during the early morning hours and a dip during the lunch hours was observed.
Now analysing the overall-monthwise activity of the entire company.

Now trying to analyse the total of email conversations happening in the company.
(Red->1999,Blue->2000,Green->2001,Yellow->2002)



Gradually increasing from the first month of the year 1999, the number of emails suddenly decrease during the august month of 2001, which coincided with the time period of resignation of the then CEO of the corporation(Jeffrey skilling), hence asking for some inspection during that period. (Turns out to be correct, would be seen later in the lda part of the project.)

3.0 Building the hierarchy structure:

The growing number of opportunities and ways people can communicate and exchange information within an organization provide us with a previously unknown way to evaluate company's structure. The data extracted from email services, phone calls, documents co-authoring, and other communication systems or common activities allow to create social networks which contain information about humans interaction and collaboration.
Hence we tried analysing a component of the huge graph developed with the email dataset and coming up with some useful conclusions.

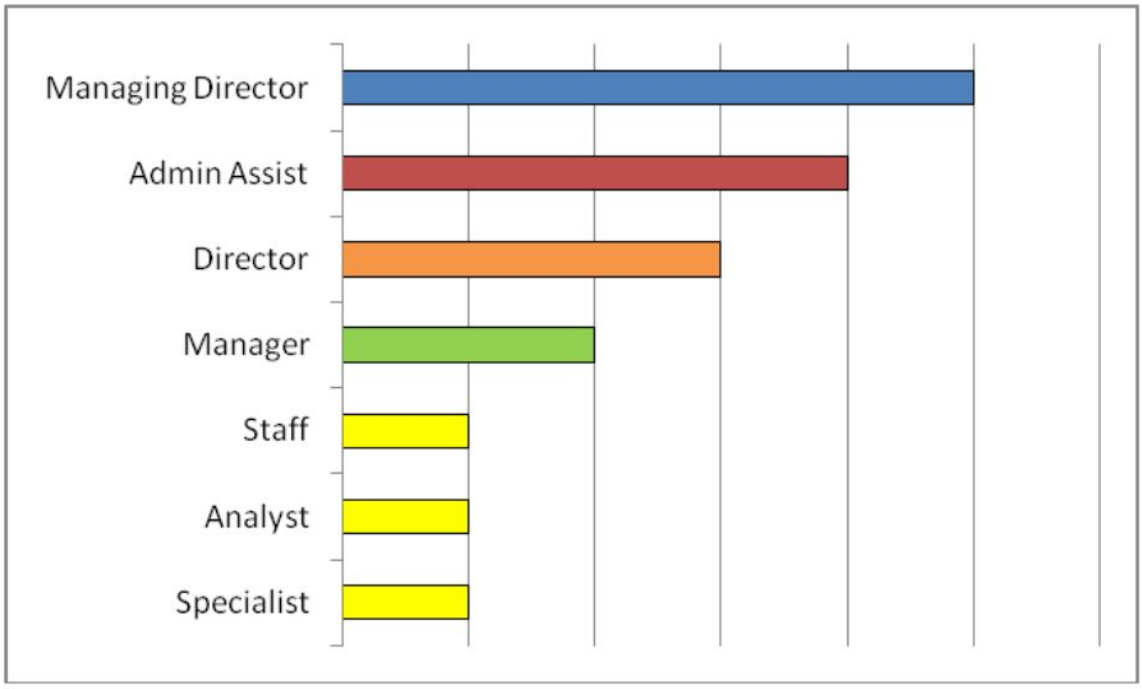


Fig. 1. Job titles hierarchy

An assumption was made that the job ranking looks as presented in Figure, where Analysts, Specialists and Staff are at the same level, following Managing Director, Admin Assistant, Director and Manager.

Finding direct relationships was tough though, hence a some information about the data was taken into consideration, as shown in the chart given below.

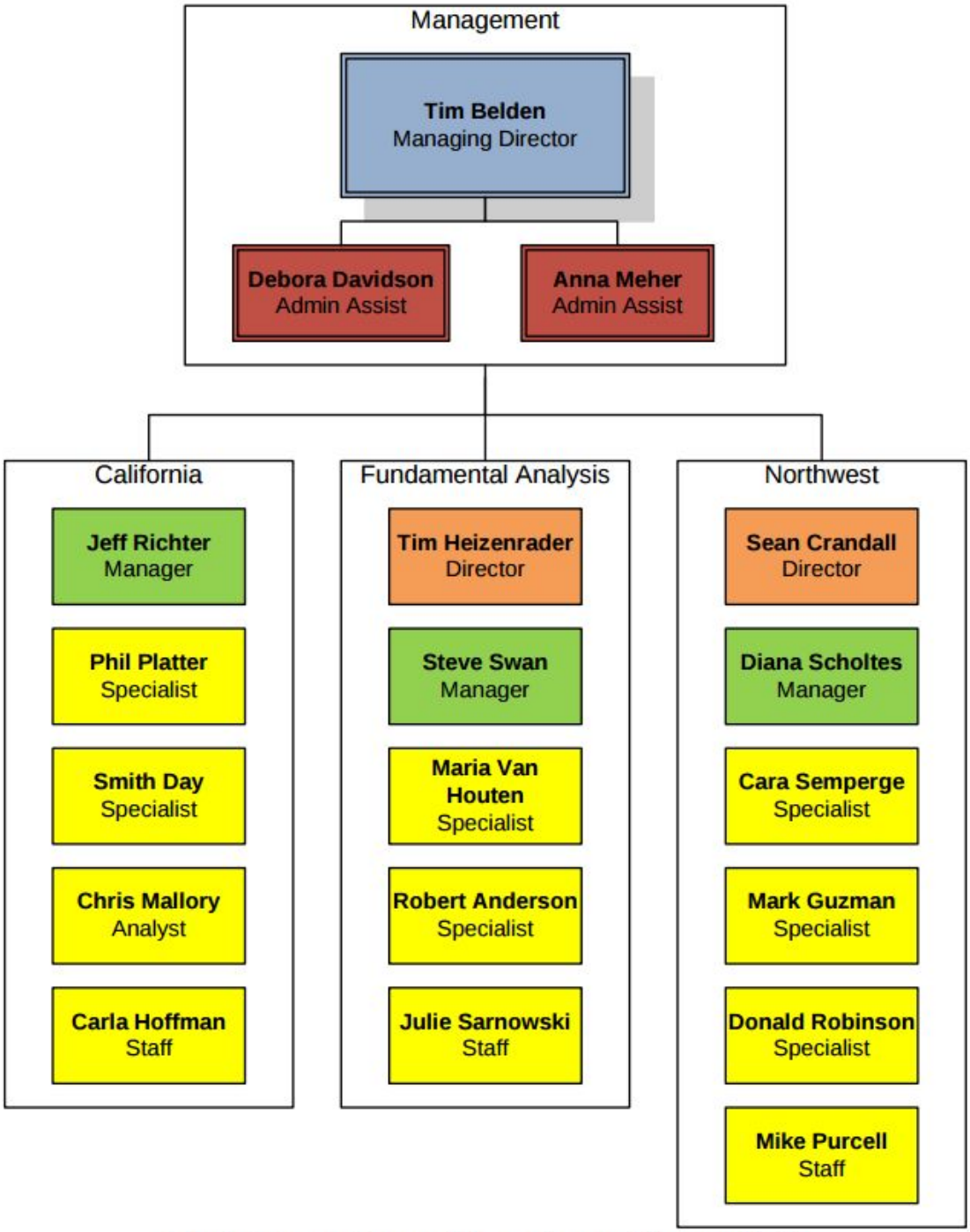


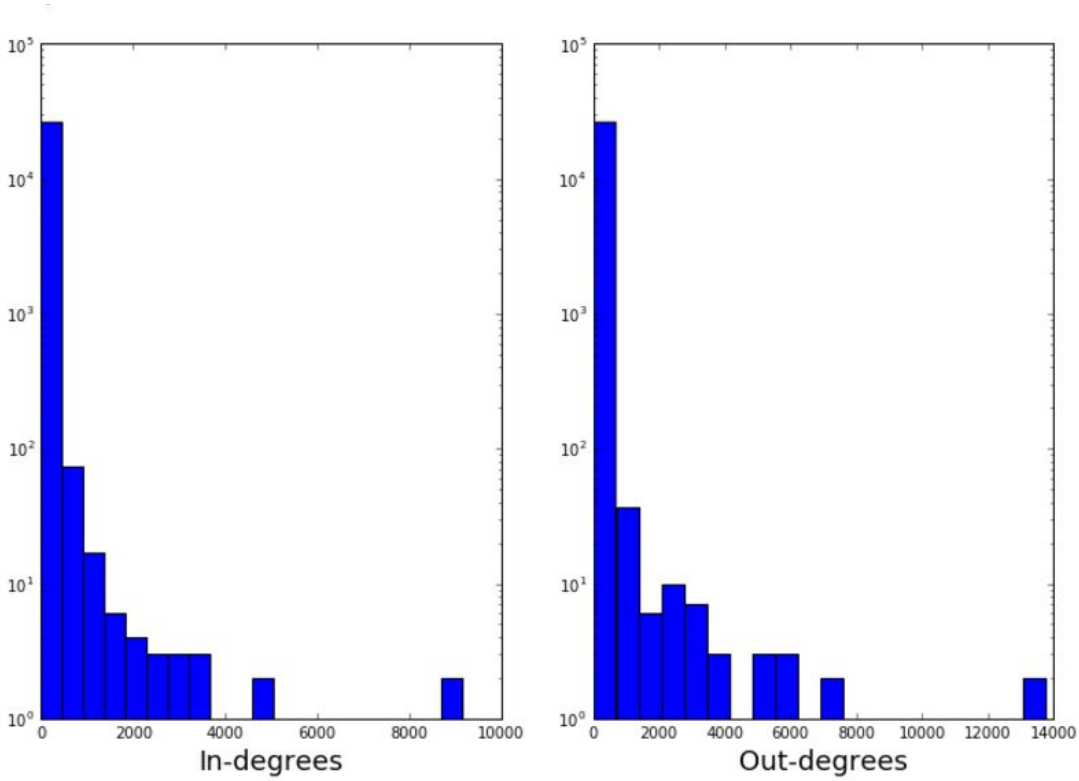
Fig. 2. Part of Enron hierarchy used for analysis

3.2 BUILDING NETWORK GRAPH AND SOME INFERENCES

Every message is in a standard mail format and contains elements such as:

- Message-id;
- Date;
- From;
- To (CC, BCC);
- Subject;
- X-Fields with user-friendly Active Directory names (X-To, X-CC etc.);
- Message body;

And a total of 517,430 messages were present in the dataset.
Constructing a GRAPH for social network analysis(used library networkx):



This figure represents the count of nodes having the incoming edges and outgoing edges lying in the given interval.

3.3 Social score dependence and Computation:

The social network of Enron is extracted from the email dataset. Using all emails in the dataset, one can construct an undirected graph, where vertices represent accounts and edges represent communication between two accounts. Then several measures for each node are applied.
An approach presented was used, where social score is computed from:

a) Emails count – number of email the user has sent and received.
b) Average response time - the time elapsed between a user sending an email and later receiving an email from that same user. An exchange of this nature is only considered a "response" if a received message succeeds a sent message within three business days.
c) Number of cliques – the number of maximal complete subgraphs that the account is contained within.
d) Raw clique score – a score computed using a size of the given account's clique set. Bigger cliques are worth more than smaller ones, importance increases exponentially with size.
e) Weighted clique score – a score computed using the importance of the people in each clique, which is computed strictly from the number of emails and the average response time. f) Centrality Degree - count of the number of ties to other actors in the network.
g) Clustering coefficient - likelihood that two associates of a node are associates themselves. h) Mean of shortest path length from a specific vertex to all vertices in the graph.
i) Betweenness centrality - reflects the number of people who a person is connecting indirectly through their direct links.
Above metrics are then weighted and normalized to a [0, 100] scale.
Used the functions available in the networkx library to compute the following score.

For each employee in the corporate hierarchy it is possible to find people who are higher or lower in the hierarchy. The Hierarchical Position (HP) is a measure that shows the importance of an employee within a company. For each user ui in a company C there is a sum of hierarchical differences D between ui and every user uj in the company divided by the number of other users. This measure can be computed either globally, where all employees in the company are taken into account (1) or at the branch level B, where only a branch of the hierarchy tree where the user is situated is being investigated (2)

.

$$HP^{global}(u_i) = \frac{\sum_{u_j \in C \wedge u_i \neq u_j} D(u_i, u_j)}{m_C - 1} \quad (1)$$

$$HP^{branch}(u_i) = \frac{\sum_{u_j \in B \wedge u_i \neq u_j} D(u_i, u_j)}{m_B - 1} \quad (2)$$

The hierarchical difference D(x,y) can be calculated in two ways: as a sign function (3) or as a direct difference between assigned hierarchy levels L(u), where L(u)N + , L(u)=1 is the highest level and .L(x)=L(y)+1 when x is one level higher than y.

$$D^{sgn}(x, y) = \begin{cases} 1, & if\ x\ is\ higher\ in\ the\ hierarchy\ than\ y \\ 0, & if\ x\ and\ y\ are\ at\ the\ same\ level\ of\ the\ hierarchy \\ -1, & if\ x\ is\ lower\ in\ the\ hierarchy\ than\ y \end{cases} \quad (3)$$

$$D^{level}(x, y) = L(x) - L(y) \quad (4)$$

3.4 SOCIAL SCORE ANALYSIS :

Now that social scores were calculated for all the employees, the members were arranged according to their score, where the results showed people in managing positions to be higher than others. Though there is lack of information about the hierarchy structure of the company but a similar calculation could be done for the entire graph bringing out people belonging to the similar hierarchy levels.

	B/W	Clustering	Degree	HR	Social Structurescore
0	376.32	0.041	80	1	74.86
1	276.35	0.046	66	2	65.31
2	262.54	0.048	62	2	61.00
3	143.68	0.040	55	5	61.20
4	80.00	0.500	66	5	52.00
5	21.44	0.300	45	4	52.50
6	45.00	0.600	40	3	44.00

3.5 Conclusions:

- Once modelled the employees into hierarchy orders, analysis regarding the change of behaviour of people belonging to similar level could be analysed.
- A clear demarcation of people belonging to different levels could be seen as the social scores show a jump between consecutive levels.

4.0 TOPIC MODELLING & LATENT DIRICHLET ALLOCATION

In machine learning and natural language processing, a **topic model** is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. Topic modeling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both. Topic models are also referred to as probabilistic topic models, which refers to statistic algorithms for discovering the latent semantic structures of an extensive text body. Topic models can help to organize and offer insights for us to understand large collections of unstructured text bodies.

In this project, we have implemented topic modelling using Latent Dirichlet allocation (LDA), an algorithm that we describe next, using the open source GENSIM implementation.

Our reason to implemented topic modelling on the Enron email dataset is to find out if there is a pattern in the variation of certain topics in the content of the emails in the dataset and to see if these topics could be flags for fraudulent behavior in the company.

In natural language processing, **latent Dirichlet allocation** (**LDA**) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. LDA is an example of a topic model and was first presented as

a graphical model for topic discovery by David Blei, Andrew Ng, and Michael I. Jordan in 2003.

In more detail, LDA represents documents as **mixtures of topics** that outputs words with certain probabilities. It assumes that documents are produced in the following fashion: when writing each document, you

- Decide on the number of words N the document will have (say, according to a Poisson distribution).
- Choose a topic mixture for the document (according to a Dirichlet distribution over a fixed set of K topics).
- Generate each word in the document by:
- ....First picking a topic
- ....Then using the topic to generate the word itself (according to the topic's multinomial distribution).

Assuming this generative model for a collection of documents, LDA then tries to backtrack from the documents to find a set of topics that are likely to have generated the collection.

In this project we have used LDA analysis without verifying the rigorous mathematics that support it and guarantee good results and have rather assumed them to be so.

5.1 PREPROCESSING THE DATA AND INITIAL RESULTS OBTAINED

The preprocessing of the data for LDA involved stripping away the metadata of the emails such as the email addresses, date, time etc. that would have added noise to the analysis. Further the content of the email was retained after tokenization, stop word removal and stemming. Finally the tokens were converted into BOW representation and were presented as inputs to the LDA algorithm.

The parameter for the number of topics was taken to be 4. The top 20 most probable words were extracted from each topic and were assessed to find suitable coherence among them.

However, the results of the LDA suggested that 2 of the 4 lists we got did not have enough coherence among them to be used for analyzing company behavior.
So we labelled them as Noise. The other 2 lists had a definite pattern as can be seen below:

The meeting list:

> Thanks call don week meeting mail day Enron message hope Houston home Vince love night Friday Thursday tomorrow talk weekend

The business list:

> Enron power energy market company business California gas risk Vince Jeff trading management information price meeting markets financial news prices

These lists clearly are very coherent and therefore have been given the labels ‘meeting list’ and ‘business list’. We will use these lists to draw inferences.

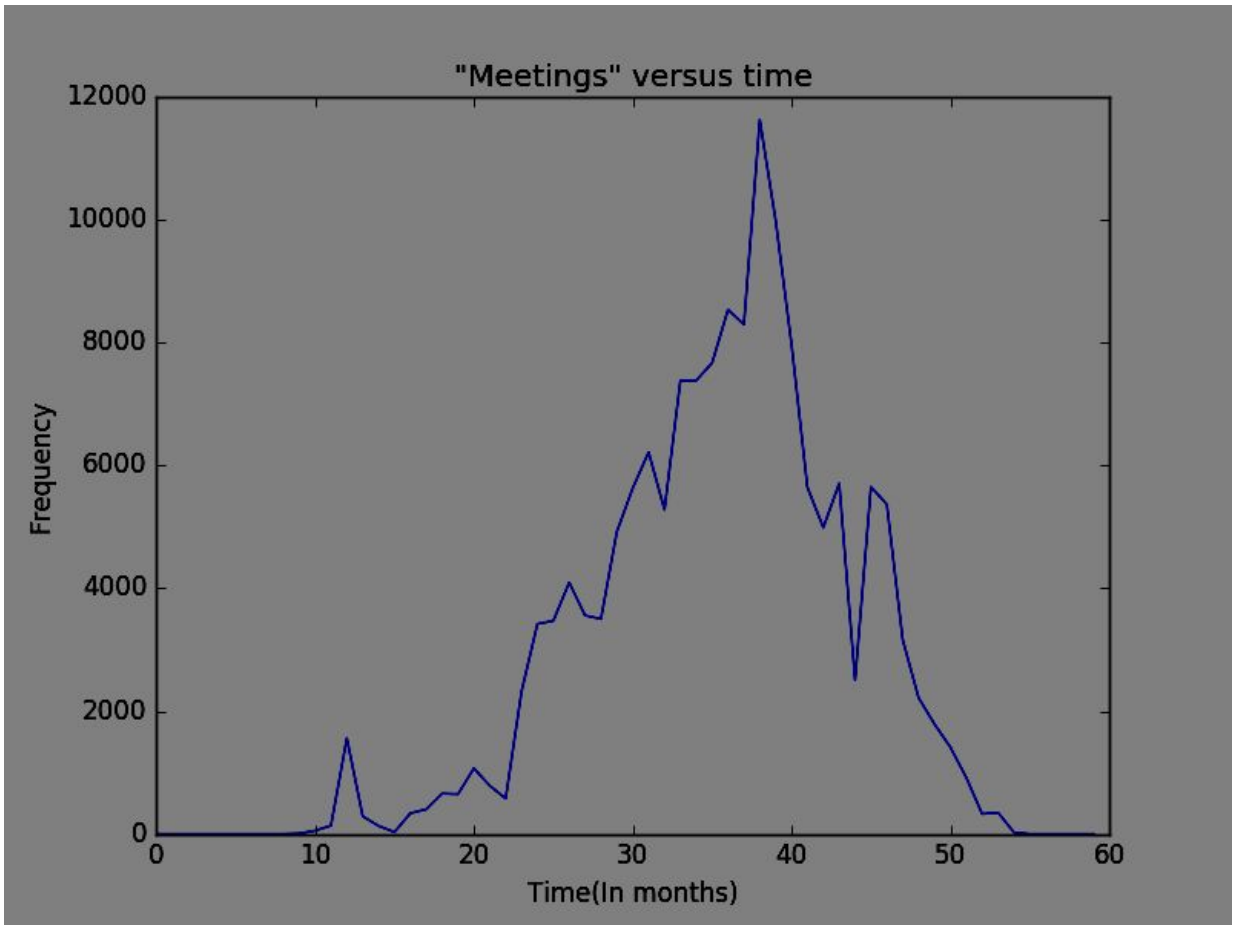
5.2 DISTRIBUTION OF TOPICS WITH TIME

The Enron emails dataset spans from 1998 to 2002 and so, for the purpose of visualization, we have divided them into 60 months. We have then plotted the distribution of the occurrence of the words from the 2 lists that we had generated through LDA analysis against this period of time and tried to look for patterns in these distributions that might serve as indicators of fraudulent behavior.

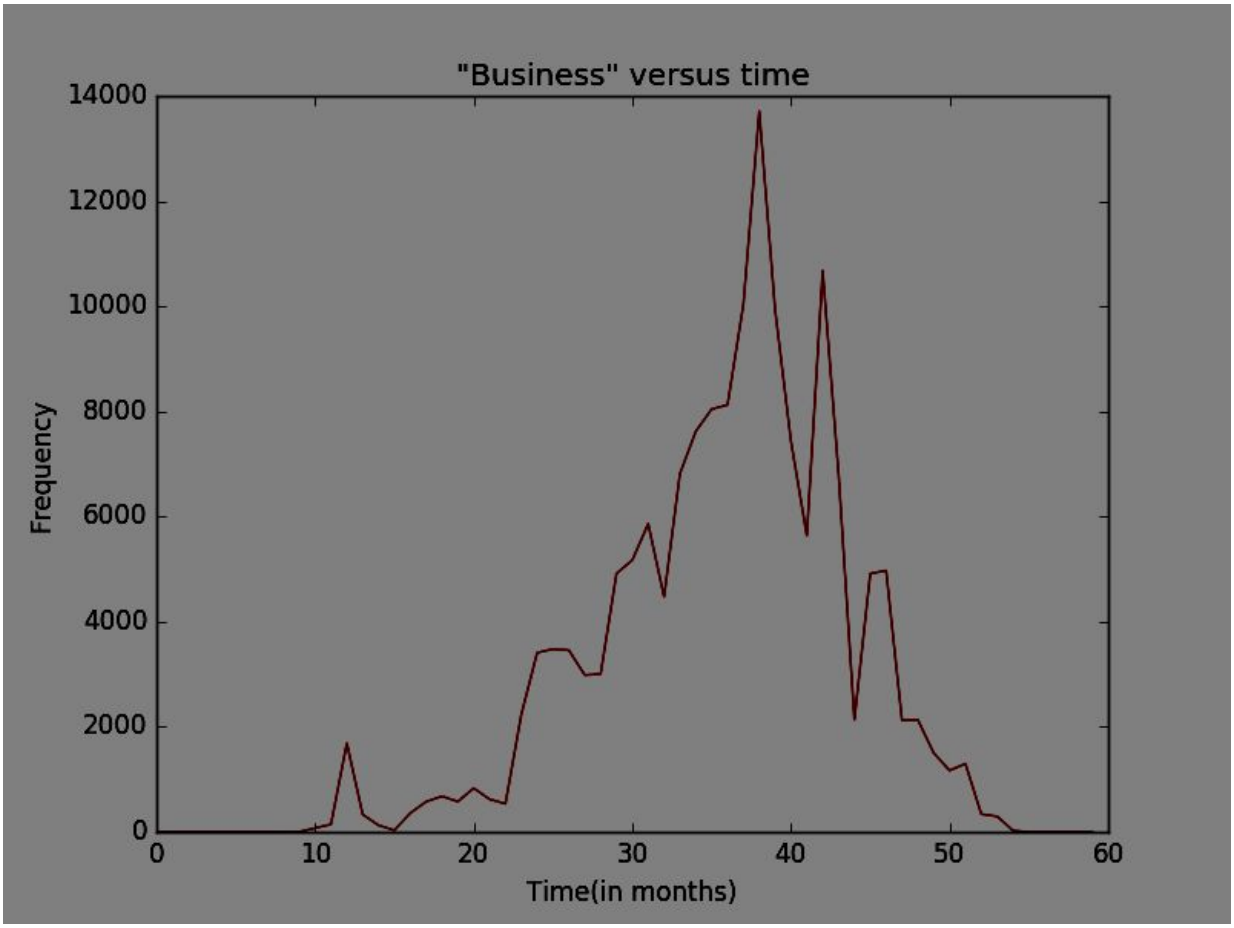
Before we move on to the distributions, it is important that we mention that the fraudulent behavior in Enron corp. was supposed to be at its peak during the 38^{th}-40^{th} month period in our 60 month distribution. We have interpreted our graphs keeping this in mind.

5.3 SOME PLOTS

PLOT OF MEETINGS LIST VERSUS TIME



PLOT OF BUSINESS LIST VERSUS TIME

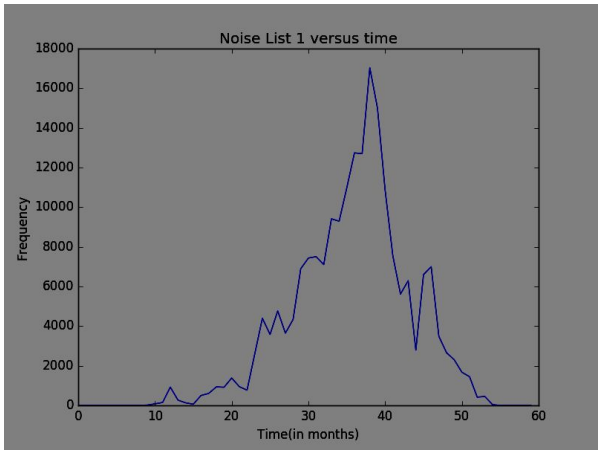


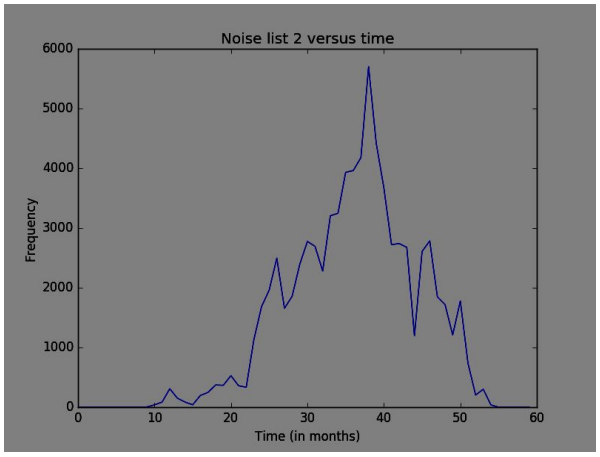
As is apparent from the 2 graphs above, the meetings seemed to have a drastic drop during the suspected time of fraudulent behavior while words such as 'risk' or 'financial' or 'energy' peaked in frequency. This could be looked upon as an indicator for fraudulent activities because of the sudden drop in meetings and the unexpected rise in business related words.

While it is possible that one could say that this “pattern” we have could be coincidental, we have therefore supplemented these graphs with the

graphs of the noise lists we have to show that the behaviour shown by business related words is in fact suspicious and not coincidental.

EVIDENCE THAT BUSINESS LIST'S DISTRIBUTION IS SUSPICIOUS





The fact that these 2 lists behave in identical manner and are similar to the Meetings list indicate that the behavior shown by the Business list is not a coincidence.

5.4 INFERENCE FROM INDIVIDUALS

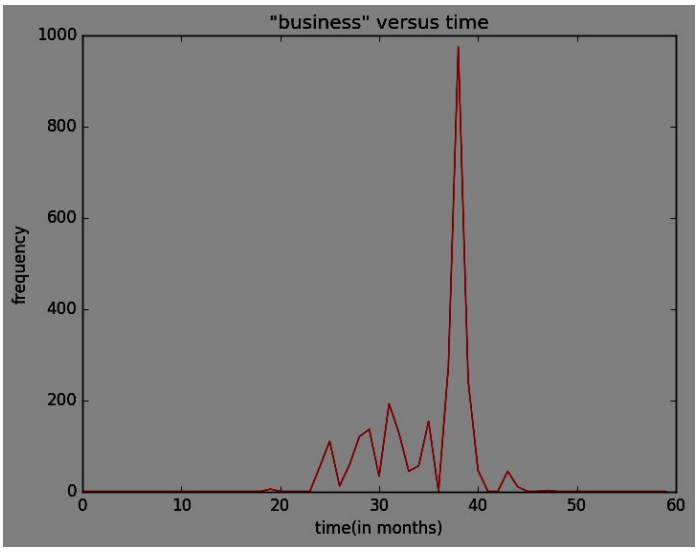
While the above analysis provided us insights to the behaviour of the entire company as a whole, we now looked for the distribution of the frequency of the two lists in each individuals' emails and plotted them against time to categorize/cluster people according to their distributions.

We tried to find clusters in those plots to identify the people who had similar plots which could indicate fraudulent behaviour. We used plots of people we knew to be guilty as references for clustering. Two of those are:

1. KENNETH LAY

   Kenneth Lee "Ken" Lay (April 15, 1942 – July 5, 2006) was an American businessman. He was the CEO and chairman of Enron Corporation. Lay was indicted by a grand jury and was found guilty of 10 counts of securities fraud. Lay died while vacationing, three months before his October 23 sentencing. A preliminary autopsy reported Lay had died of a heart attack caused by coronary artery disease and his conviction was vacated.

   Business Plot for Kenneth Lay:



Points to note in this distribution are that there is a sharp peak right before the 40 month mark when the scandal was very active and extremely dormant after the 40 month mark during the Audit period.

2. JEFFREY SKILLING

Jeffrey Keith "Jeff" Skilling (born November 25, 1953) is the former CEO of the Enron Corporation, headquartered in Houston, Texas. In 2006, he was convicted of federal felony charges relating to Enron's financial collapse and is currently serving 14 years of a 24-year, four-month prison sentence at the Federal Prison Camp (FPC) – Montgomery in Montgomery, Alabama.

Business Plot for Jeff Skilling:

In his case too, the peak happens in the 35-40 month period and goes extremely dormant in the audit period.

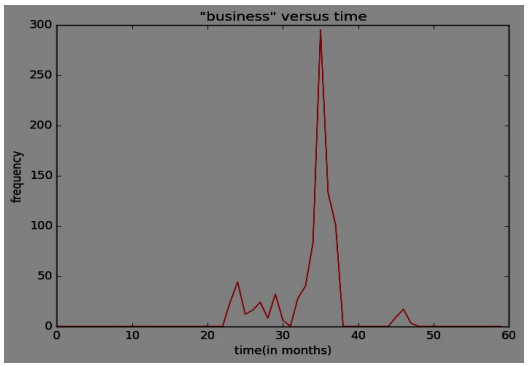
Thus our clustering scheme was to look for people who had identical distributions to those of Kenneth lay and Jeff Skilling as that could indicate possible fraudulent behaviour.

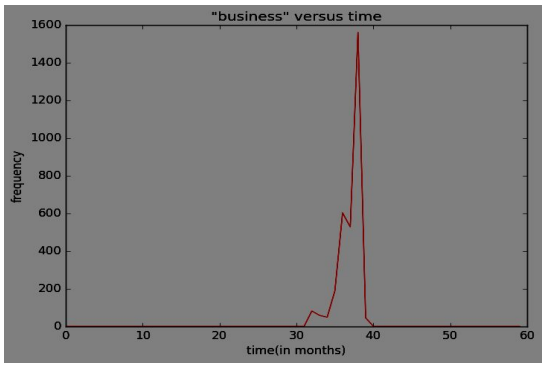
5.5 OBSTACLES FACED IN CLUSTERING

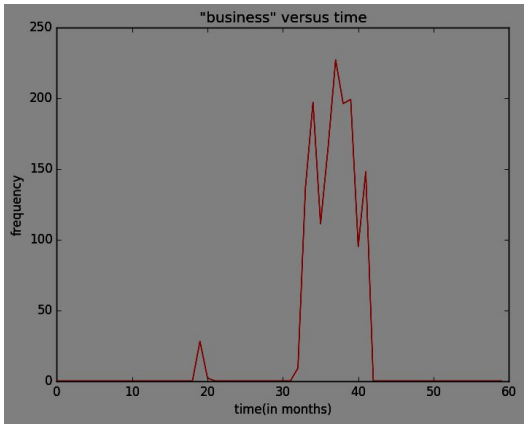
A major obstacle that we faced during clustering was that very few emails are available of the known “main players” in the Enron email dataset which limited us to draw conclusions from the distributions of only 2 people with known fraudulent behaviour. Thus our basis for clustering is not very strong. In fact, most emails are from people who were not in the limelight during the scandal and therefore it is difficult to say whether the conclusions we derive about these people are accurate or not.

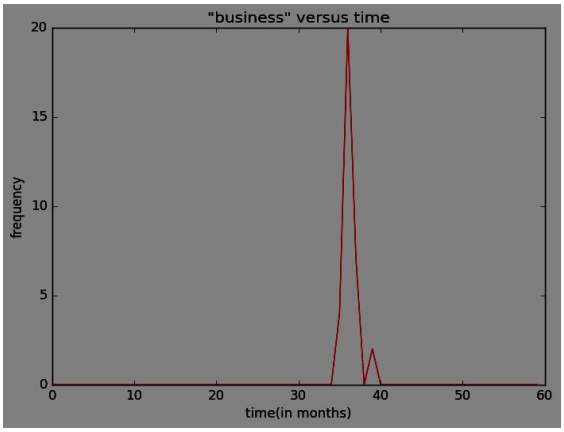
PLOTS SIMILAR TO THOSE OF KENNETH LAY AND JEFFREY SKILLING

We found a few (around 10) graphs which were similar to those of Kenneth Lay and/or Jeff Skilling and we have attached 4 of those here as proofs. Since the 10 people we have clustered have not sent a very large number of emails, our next task would be to go through the content of their emails and actually look for ourselves whether our conclusions about their involvement in the fraudulent behaviour is correct or not.









Because there is little to no information available on the internet about the involvement of these people in the scandal, cross validation of these results will take time because we would do them manually by going through their emails.

5.6 CONCLUSIONS

We have seen that topic modelling could be used to identify periods of fraud on a timescale using plots of frequency versus time. This could be used to identify fraud in other companies. This technique could be used to single out individuals using clustering. Thus, we have created a framework which could be applied to various Natural language processing situation where we need to gather insights about fraudulent activities from textual data.

6.0 References

- https://www.ii.pwr.edu.pl/~brodka/art/2011_IJKSR.pdf
- https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf