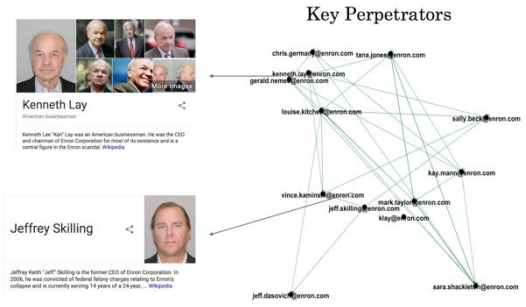
Columbia Data Science Society (CDSS) Hackathon, 2017
Team Name – Noob Network
Case study – Analyzing Enron emails for fraudulent behavior

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. It was alleged of unethical and fraudulent behavior and the story ends with the bankruptcy of one of America's largest corporations. The problem statement here was to mine through the its emails dataset and leverage the insights from the data to some meaningful conclusions.

We implemented a two-fold analysis addressing its two most important questions: *"What are the employees talking about? (NLP)"* and *"Who are they communicating with? (Network Analysis)"*

Network Analysis

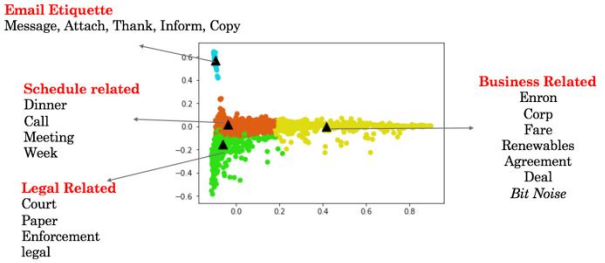


We narrowed down the data to the emails which were circulated among smaller groups and were sent at suspicious non-office work hours. We created a network which analyzed this flow of information. Our hypothesis was validated when we found the top 2 perpetrators (The Chairman and the CEO) in the network with others who were accused of fraud.

Community Collusion

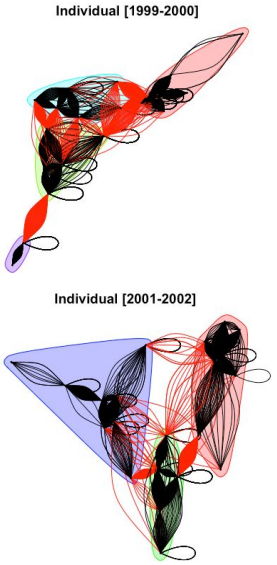
Natural Language Processing

We incrementally trained the complete email corpus with google news dataset Word2Vec model to get efficient word vectors which capture semantic meaning of the text. Using those we identified keywords relevant to suspicious behavior and secrecy / rumor language. We then found the top 15 employees whose mails contained a large number of occurrences of these words. We also discovered that 8 of them were from our previous network analysis. Finally, we implemented LDA based topic-modelling to find recurring topics being discussed in the emails. We found general topic



such as *business* (enron, corp, deal, renewable) and *legal* (court, paper, enforcement).

Email Network Community Collusion:



We have used another innovative technique to visualize the flow of information between the individuals who exchange email frequently and can be identified in a community through the plots. It is observed that the information flow between the communities decreased after the company fell in 2001. The communities expanded and the information flow increased within the community itself.

Summary

We used a two-pronged approach – Network Based Analysis and Natural Language Processing to identify Enron employees who were potentially indulgent in fraudulent or unethical behavior. The network analysis identified a small network of individuals who showed suspicious behavior and NLP unearthed that their emails indeed contained numerous semantic indications of unethical behavior.