Math 168: Short Essay 1
Prof. Mason Porter

Dhruv Chakraborty

October 10, 2020

The Enron Corpus

The Enron Scandal of 2001 is one of the best known corporate scandals in recent history. It lead to the bankruptcy and dissolution of not only one of the largest and richest oil companies in the world at the time (Enron), but also of their auditors - Arthur Andersen. This accounting scandal was investigated heavily by the Security Exchanges Commission (SEC) and the Federal Energy Regulatory Commission (FERC). As part of their investigation, the FERC subpoenaed all of Enron's email records, later also making them publicly available. This email data set is one of the most popular data sets ever produced, and has been used in fields ranging from Natural Language Processing to Network Analysis [1].

The data set itself has a natural communication network structure on it, with Enron employees serving as Nodes, and an edge (i, j) serving as an email correspondence between employees i and j. Moreover, the edges are directed and weighted by the number of emails sent and received between the employees. There are $36,692$ nodes representing employees and $183,831$ edges representing emails in the network. Although the initial data set had around a half a million emails, it has been cleaned and processed extensively after the initial collection and publication effort as part of the Federal investigation. The diameter i.e. longest shortest path of the network is only 11, suggesting just how connected the employees in even a massive global company are. It is my understanding that such results from network analysis, particularly when applied to a Facebook network, are responsible for

the canonical notion of "6 degrees of separation" between any two random people (on average).

There are several papers analyzing the Enron corpus, including ones that use anomaly detection and key player identification methods to discover the employees that actively engaged in fraud activities. However, a paper I found particularly interesting actually resulted from an undergraduate teaching seminar on network analysis [2], with students experimenting with the data set. In particular, the authors encourage the students to look at the data in many ways, with a specific interest in analyzing the difference between types of centralities and community detection techniques.

One of the primary methods used by the students was creating interactive visualizations, two of which are shown below.

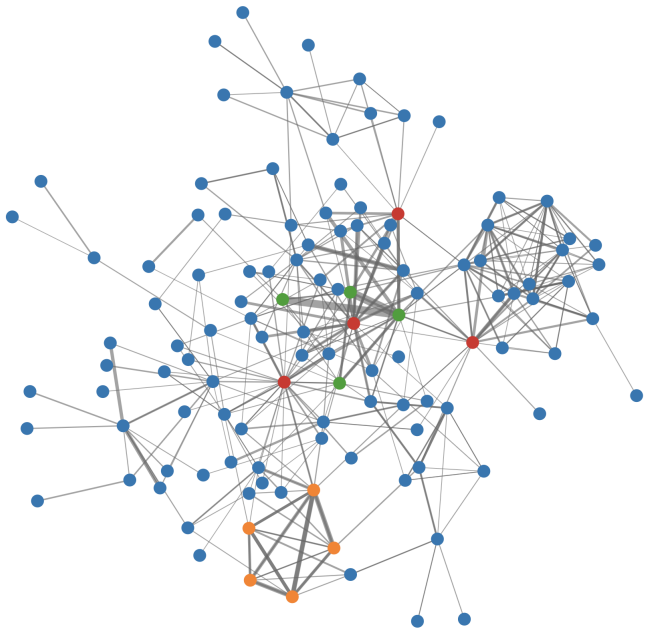

Figure 1: Enron email network colored by centrality measures

Figure 1 [3] above shows us the communication structure of some of Enron's most "central" employees, through many different measures. In particular, green indicates closeness centrality, red shows betweeness centrality

and orange is event centrality. The graph is interactive and has labels on hover, the full version can be found in the referenced link.

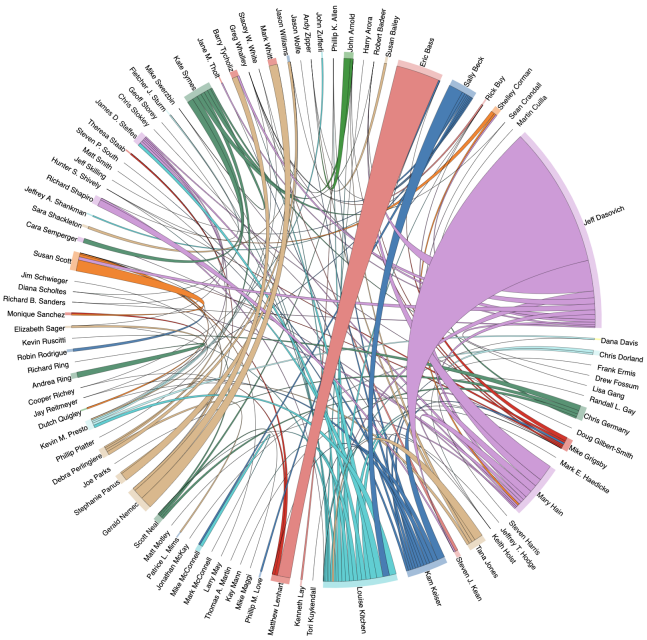

Figure 2: Alternate visualization of Enron email network colored by employee department

Figure 2 [4] shows us communication between key players in Enron, split by department. The overlap with Figure 1 is of particular interest, since Jeff Dasovich (who has the largest presence, in purple), is also present in Figure 1 as a green node (high closeness centrality).

Such results are quite interesting as they show the amount of insight and visualization that can be derived by simply considering a email network. Another result I found to be fascinating was one found by Leskovec et. all in "Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters."[5] Upon analyzing several large network data sets (including the Enron corpus), they discover the inherent differences between small and large clusters in networks - tight communities that are barely connected to the rest of the network and larger communities that tend to "blend into" the full network.

References

[1] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection* `http://snap.stanford.edu/data`, June 2014.

[2] J. S. Hardin, G. Sarkis and P. C. URC (2015) *Network Analysis with the Enron Email Corpus* , Journal of Statistics Education, 23:2, , DOI: 10.1080/10691898.2015.11889734

[3] Daniel Metz, Emily Proulx, David Khatami and Tim Kaye. *Enron Network Analysis - Rank* `http://enron-network.herokuapp.com/TOM`, 2015.

[4] Daniel Metz, Emily Proulx, David Khatami and Tim Kaye. *Enron Email Corpus - by Department* `http://obscure-meadow-3612.herokuapp.com/`, 2015.

[5] J. Leskovec, K. Lang, A. Dasgupta, M. Mahoney. *Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters.* Internet Mathematics 6(1) 29–123, 2009.