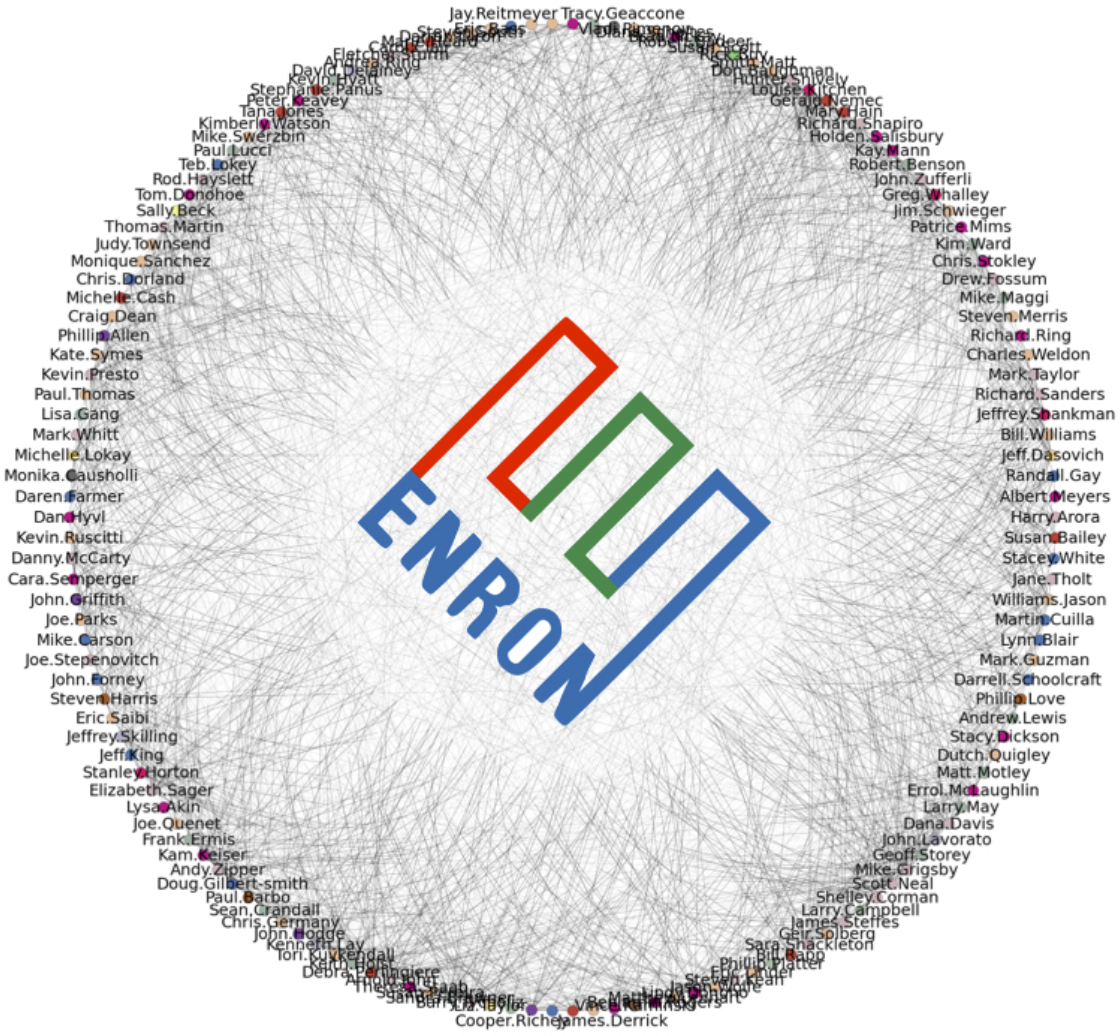


Network Analysis of Enron Email Data

Akhil Ketkar
akhilketkar@g.harvard.edu

Arjun Sanghvi
asanghvi@g.harvard.edu

December 10, 2014

1 Introduction

The world has never been more connected. The internet is nearly ubiquitous and acts as both a catalyst for human interaction and a lens through to which study it. Increased generation and accessibility of data, in addition to the proliferation of massive social networks, has led to substantial network analysis research in the past two decades.

In this paper, we examine structural properties of the Enron email network to better understand the organization and dynamics of the corporation in its final tumultuous years. In comparison to previous analyses, we provide further temporal granularity on the order of months.

The Enron email dataset is interesting because it contains real email data from employees at a major organization that was involved in a massive fraud. The dataset contains a large amount of information that can be used to answer a number of interesting questions in areas such as Social Network Analysis, Organizational Behavior, such as: who are the key actors in the information network, are there communities in within the network, how do these features of the network evolve over time, does information flow over a network look different in a "crisis" etc. In addition to network or graph theoretic techniques, the dataset can be analyzed from an NLP perspective.

2 Brief Background on Enron

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. Before its bankruptcy on December 2, 2001, Enron employed approximately 20,000 staff and was one of the world's major electricity, natural gas, communications, and pulp and paper companies, with claimed revenues of nearly $111 billion during 2000. Fortune named Enron "America's Most Innovative Company" for six consecutive years.

At the end of 2001, it was revealed that its reported financial condition was sustained by an institutionalized, systematic, and creatively planned accounting fraud, known since as the Enron scandal. Enron has since become a well-known example of willful corporate fraud and corruption. The scandal also brought into question the accounting practices and activities of many corporations in the United States and was a factor in the creation of the Sarbanes Oxley Act of 2002. The scandal also affected the greater business world by causing the dissolution of the Arthur Andersen accounting firm.

3 Data and Resulting Graphs

3.1 Dataset

There are several versions of the Enron email dataset. The version that we are utilizing for this paper comes from the work of Jitesh Shetty and Jafar Adibi [1] at ISI. Shetty

and Adibi cleaned the dataset by dropping emails that were blank, duplicates of unique emails, had junk data, or were returned by the system due to transaction failures. The final dataset consists of 252,759 emails in 3000 user defined folders from 151 people. Shetty and Abidi loaded the information into a MySql database that contains four tables containing information about the employees, messages, recipients and reference information.

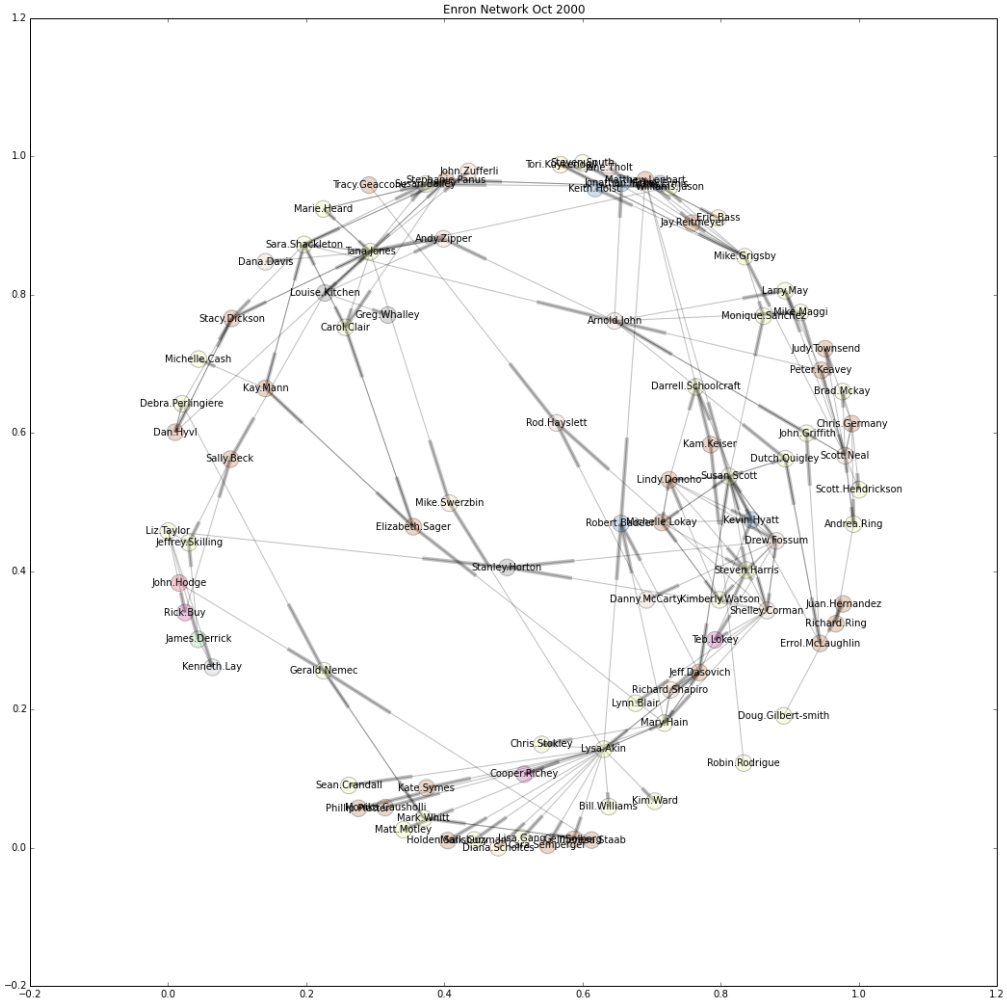
We chose this version of the dataset mainly because it is has been structured into a relational database. Unfortunately Shetty and Adibi's website (which was used as a source by a number of papers) was taken down recently. So we retrieved a copy of the the data from Joel Pfeiffer's website [2]. In addition to the email data, we also used data on the positions and specific roles of various employees from Youngser Park[4].

We made several modifications to the dataset to arrive at the version that was used in the analysis such as: normalizing the email addresses for all the employees, adding missing emails addresses, removing emails that were not sent from other employees etc. We then incorporated the employee position data into the email data to create the final version of the dataset.

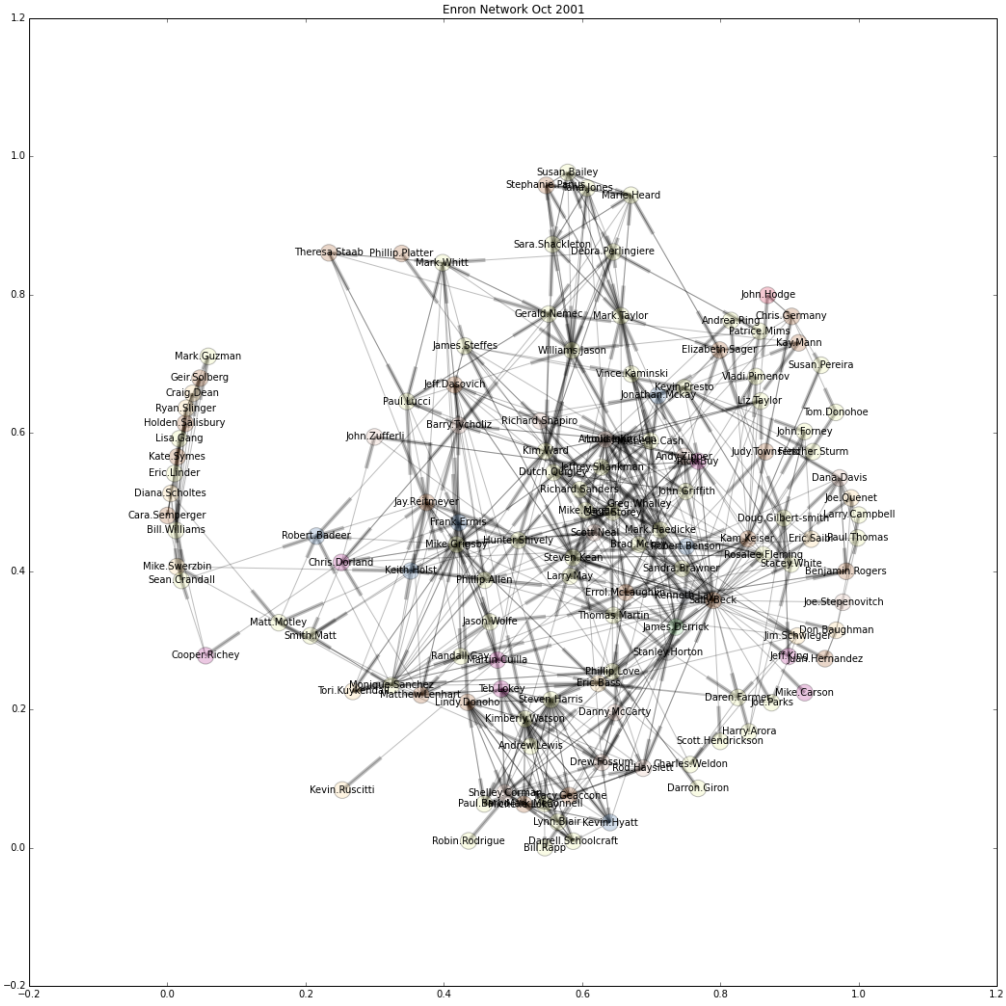
3.2 Graphs

All subsequent analysis in the paper relies on the network representation of the dataset mentioned above. The nodes of the network are the employees of Enron and the edges represent individual email exchanges. Certain parts of the paper use directed graphs where an edge goes from a sender to the receiver of the email while other parts use undirected graphs where an edge simply denotes an email sent between the two nodes.

This is an example of the network drawn from emails sent in October 2000. In this period Enron's share price was close to it's all time highs.



Compare the network in Oct 2000 to this network created from emails sent in October 2001 at which point the fraud at Enron has been exposed and Enron is nearing bankruptcy.



3.3 Terminology

1. A is the adjacency matrix of the network which is n x n for a network with n nodes.
2. D is the diagonal matrix containing node degrees on it's diagonals
3. L is the laplacian of the matrix defined as $D - A$
4. B Is the modularity matrix defined as $B_{ij} = A_{ij} - \frac{k_i k_j}{2m}$ where k_a is degree of node a

4 Centrality Measures

One of the key questions one can ask about an information network such as the Enron email network is: Who are the most important actors in the network? There are various measures of centrality one can use to answer that question.

4.1 Eigenvector Centrality

Eigenvector based centrality measures are based on the idea that a node is important if it is connected to other important nodes. Let A be the adjacency matrix of a graph and x the vector of centrality measures. We would like the centrality of each node to be proportional to the centralities of all the nodes connected to it. In matrix form then we would like vector x to satisfy the

$$Ax = \kappa x \tag{1}$$

Thus x is an eigenvector of the matrix A. Even though all eigenvectors of A satisfy the above equation, the eigenvector corresponding to the largest eigenvalue is used in practice. In order to compute the largest eigenvalue we use the power method.

Table 1: Employees with Greatest Eigenvector Centrality

Name	**Eigenvector Centrality**	**Position**
Sally Beck	0.238627	Chief Operating Officer
Liz Taylor	0.229042	Administrative Assistant to President
Louise Kitchen	0.220102	President of Enron Online
Kenneth Lay	0.205464	CEO
John Lavorato	0.201416	CEO, Enron America
Mike Grigsby	0.156134	Vice President
Kevin Presto	0.148424	Vice President
Scott Neal	0.147877	Vice President, trader
Barry Tycholiz	0.145584	Vice President
Arnold John	0.145032	Vice President

Katz centrality and PageRank are small modifications on the idea of eigenvector centrality. This measure identifies several key members of upper management at Enron despite the fact that they don't send or receive a lot of email.

4.2 Other measures of Centrality

4.2.1 Closeness Centrality

This is defined as reciprocal of sum of shortest path distances from v to all other nodes. The closer a node is to others, the greater its centrality.

Table 2: Employees with Greatest Closeness Centrality

Name	**Closeness Centrality**	**Position**
Liz Taylor	0.653509	Administrative Assistant to President
Sally Beck	0.626050	Chief Operating Officer
John Lavorato	0.618257	CEO, Enron America
Louise Kitchen	0.598394	President of Enron Online
Kenneth Lay	0.579767	CEO
Jeff Dasovich	0.541818	Executive, Government Relations
Kevin Presto	0.528369	Vice President
Susan Scott	0.526502	Assistant Trader
Mike Grigsby	0.520979	Vice President
Arnold John	0.520979	Vice President

4.2.2 Betweenness Centrality

Betweenness centrality of a node is the sum of the fraction of all-pairs shortest paths that pass through that node.

Table 3: Employees with Greatest Betweenness Centrality

Name	**Betweenness Centrality**	**Position**
Louise Kitchen	0.100766	President of Enron Online
Susan Scott	0.069708	Assistant Trader
Mike Grigsby	0.060098	Vice President
Kenneth Lay	0.048120	CEO
Kim Ward	0.046483	Associate Director, Enron North America
Jeff Dasovich	0.045877	Executive, Government Relations
Liz Taylor	0.043397	Administrative Assistant to President
Sally Beck	0.043037	Chief Operating Officer
Kevin Presto	0.040577	Vice President
Bill Williams	0.037401	Trader

5 Communities

The basic goal of community detection is to separate the network into groups of vertices that have few connections between them. Cut size is a simple metric that can be used to find communities in a network. The cut size is simply the number of edges you have to remove from the network so that the two groups of nodes become disconnected.

$$R = \frac{1}{2} \sum_{i,j in different groups} A_{ij} \tag{2}$$

where the factor of $\frac{1}{2}$ compensates for the fact that each edge appears twice in the adjacency matrix.

Modularity is another (arguably better) measureThe idea is that we find the fraction of edges that run between vertices of the same type, and then we subtract from that figure the fraction of such edges we would expect to find if edges were positioned at random without regard for vertex type [**?**]. When the fraction of edges between vertices of the same type is significantly greater than what would be expected at random, we can say that the network exhibits high homophily. The number of edges that run between nodes of the same type is given by

$$\sum_{edges(i,j)} \delta(c_i, c_j) = \frac{1}{2} A_{ij} \delta(c_i, c_j) \tag{3}$$

where c_i and c_j are the classes of nodes and $\delta(m, n)$ is the Kronecker delta which is equal to 1 if $m = n$ and 0 otherwise.

The expected number of edges between nodes of the same type is given by

$$\frac{1}{2} \sum_{i,j} \frac{k_i k_j}{2m} \delta(c_i, c_j) \tag{4}$$

where k_i denotes the number of edges of node i (it's degree) and m is the total number of edges in the network. Subtracting (3) from (2) gives us modularity Q

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{5}$$

5.1 Partitioning using Cut Size

We can use these ideas to find communities within a network. The following discussion is based on Mark Newman's work [5].

Let us define quantities s_i for each vertex i, which represent the division of the network:

$$s_i = +1 \text{ if node belongs in group 1} \tag{6}$$

$$s_i = -1 \text{ if node belongs in group 2} \tag{7}$$

Then

$$\frac{1}{2}(1 - s_i s_j) = 1 \text{ if i and j are in the same group} \tag{8}$$

$$\frac{1}{2}(1 - s_i s_j) = 0 \text{ if i and j are in different groups} \tag{9}$$

This allows us to rewrite equation 2 for cut size as

$$R = \frac{1}{4} \sum_{ij} A_{ij}(1 - s_i s_j) \tag{10}$$

Simplifying the above equation and using the fact that $\sum_j A_{ij} = k_i$ we get

$$R = \frac{1}{4}\sum_{ij}(k_i\delta_{ij} - A_{ij})s_i s_j = \frac{1}{4}\sum_{ij} L_{ij}s_i s_j \tag{11}$$

where L is the Laplacian of the matrix. We can rewrite this equation in matrix form as

$$R = \frac{1}{4}s^T L s \tag{12}$$

where s is the vector of all s_i. Equation 12 gives us a optimization problem: find s such that the quantity R is minimized. s is constrained to only have values that are +1 or -1.

This turns out to be a very difficult problem to solve so we relax some of the constraints on s. s is allowed to take any value subject to $|s| = \sqrt{n}$ where n is the number of nodes and $1^T s = n_1 - n_2$ where n_1 and n_2 are the sizes of the partitions we need.

We can differentiate with respect to s and introduce two Lagrange multipliers to solve this problem

$$\frac{\partial}{\partial s_i}\left[\sum_{jk} L_{jk}s_j s_k + \lambda\left(n - \sum_j s_j^2\right) + 2\mu\left((n_1 - n_2) - \sum_i s_j\right)\right] = 0 \tag{13}$$

This results in

$$Ls = \lambda s + \mu 1 \tag{14}$$

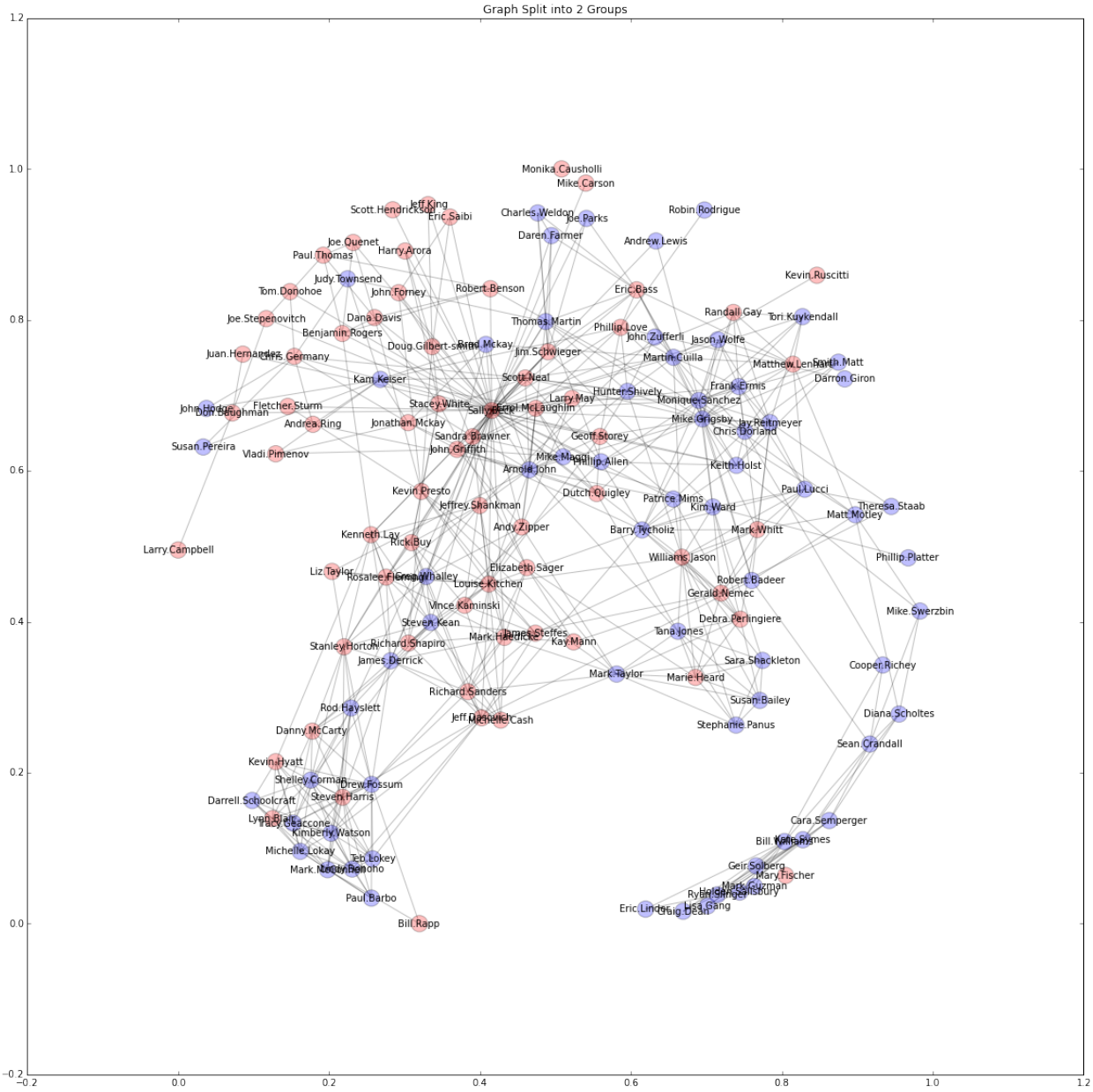
It is easy to see that the vector 1 is an eigenvector of a Laplacian matrix since the sum of the off-diagonal entries of a Laplacian is the degree which is 0 minus the diagonal element of the matrix. Knowing this allows us to compute $\mu = -\frac{n_1 - n_2}{n}\lambda$.

If we define a new vector $x = s + \frac{\mu}{\lambda}1$ we can see that $Lx = \lambda x$. We now need to pick a λ such that R (the cut size) is minimized. Solving for R we get

$$R = \frac{n_1 n_2}{n}\lambda \tag{15}$$

The cut size is thus proportional to the eigenvalue λ. Given that we want to minimize R we should choose λ to be the smallest eigenvalue of L. But we know that all the eigenvalues of L are non-negative and the smallest of them is the 0 vector. So the best option is to choose the eigenvector v_2 corresponding to the second lowest eigenvalue λ_2 of the Laplacian matrix.

Applying this technique to the Enron network to try to split it into roughly equal halves we get the following using data from October 2001



The modularity of this division is 0.16. As previously mentioned a positive number means there is homophily in the network.

5.2 Modularity Maximization Using Spectral Techniques

We can also use the idea of modularity to find communities within a network. As discussed earlier modularity is defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{16}$$

or

$$Q = \frac{1}{2m} \sum_{i,j} B_{ij} \delta(c_i, c_j) \tag{17}$$

We set up a vector s in the same fashion as we did in the previous section so that $\delta(c_i, c_j) = \frac{1}{2}(s_i s_j + 1)$. Substituting that back into the definition of modularity we get

$$Q = \frac{1}{4m} s^T B s \tag{18}$$

This equation is once again of a similar form as the cut size equation in the previous section.
To solve this optimization problem we once again relax the constraint on s to only take values of +/- 1 and allow s to be any number subject to $|s| = \sqrt{n}$. Differentiating with respect to the elements of s and introducing a Lagrangian we get

$$\frac{\partial}{\partial s_i} \left[\sum_{jk} B_{jk} s_j s_k + \beta \left(n - \sum_j s_j^2 \right) \right] = 0 \tag{19}$$

which ultimately gives

$$Bs = \beta s \tag{20}$$

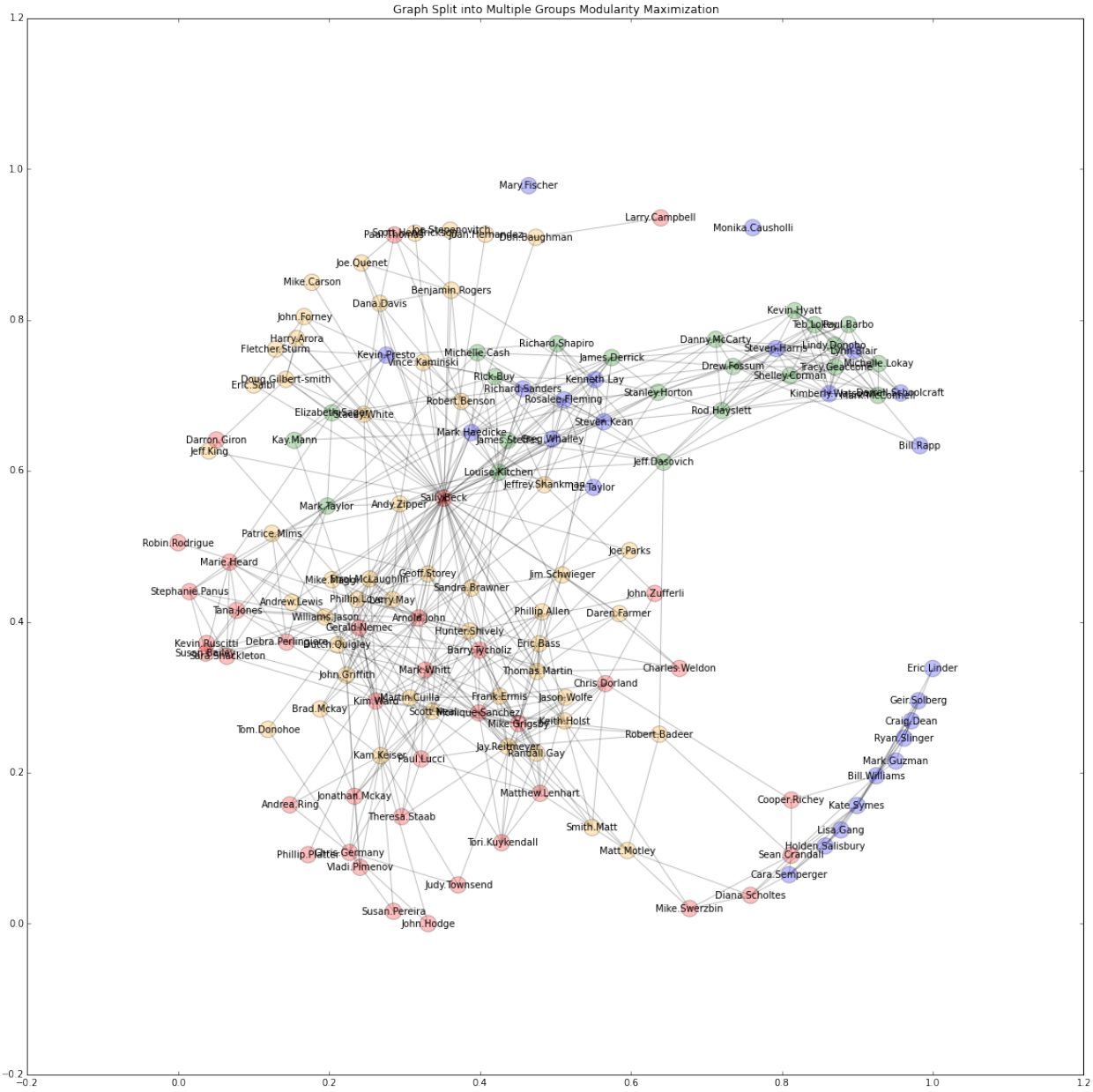
In other words s is an eigenvector of B. Substituting back into the equation for modularity you get

$$Q = \frac{n}{4m} \beta \tag{21}$$

since $s^T s = n$

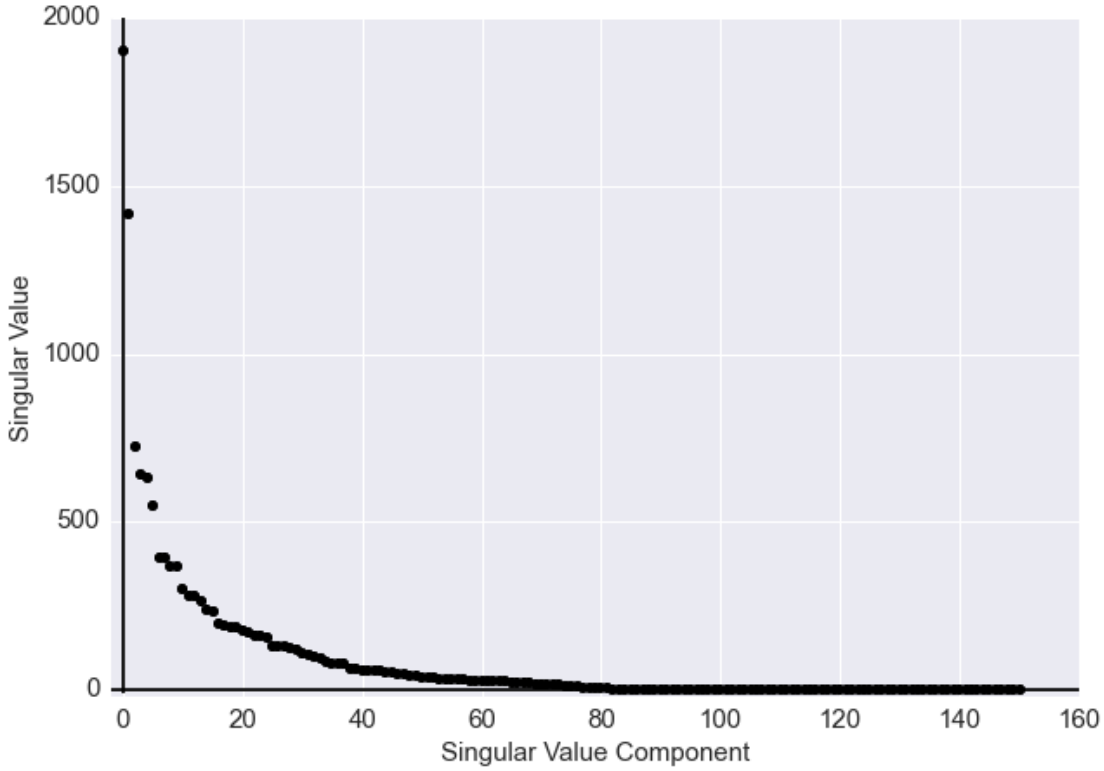
In this case we want to maximize modularity so we need to pick s accordingly. If u_1 is the eigenvector corresponding to the largest eigenvalue then we need to maximize $s^T u_1$. The best we can do is make each term in $s^T u_1$ positive. So if u_{1_i} is negative we set s_i to be -1 and +1 otherwise. This give us a relatively simple algorithm whereby we only need to look at the sign of each element of the eigenvector corresponding to the largest eigenvalue of the modularity matrix to find partitions in the data.

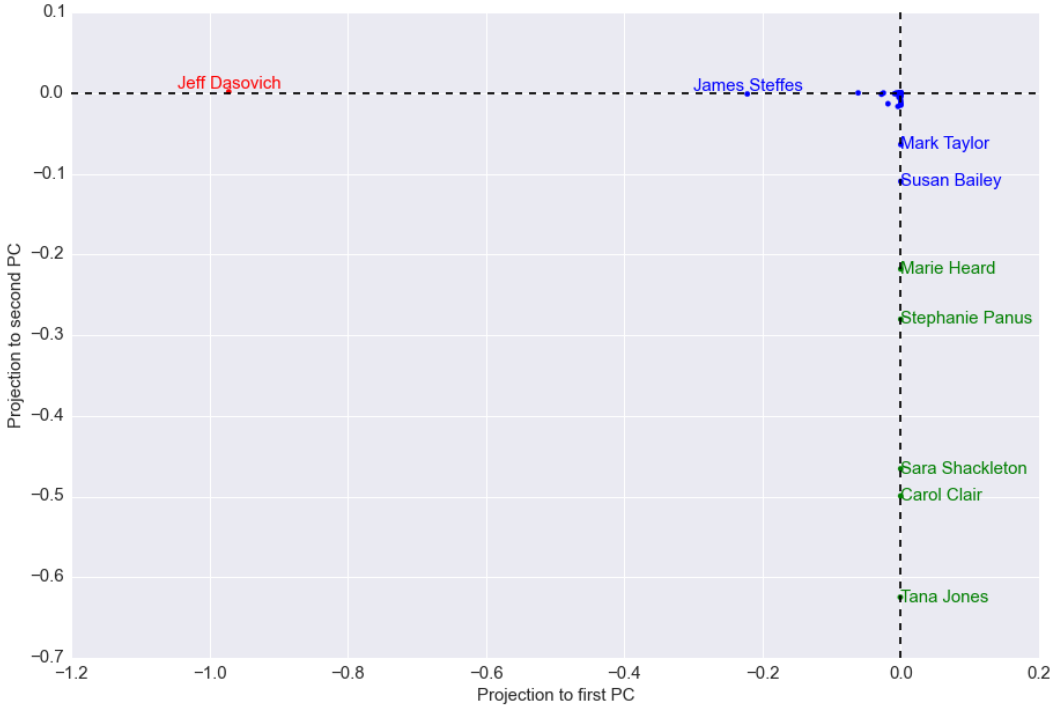
This technique can also be applied repeatedly to divided the network into more than two partitions. Applying this technique to the Enron network using data from October 2001 we get



5.3 Singular Value Decomposiiton

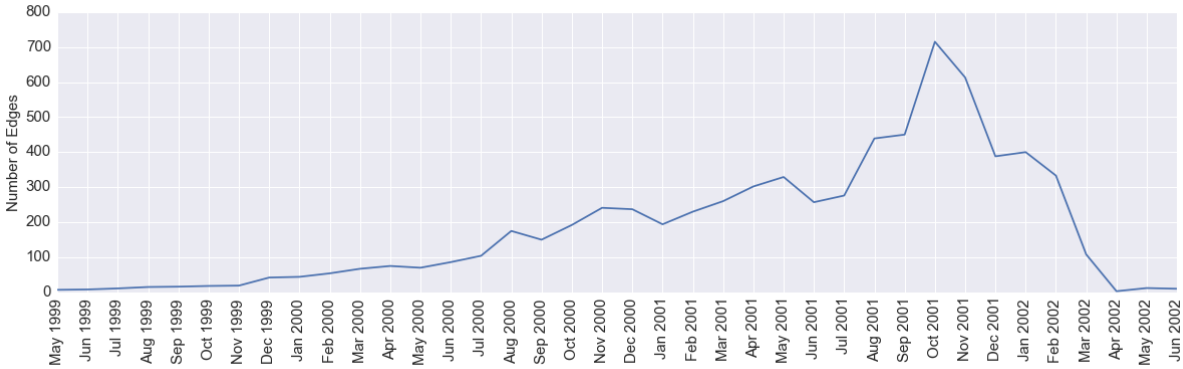
We can look at singular values of the adjacency matrix to obtain low rank approximations of the network and find communities within the network.

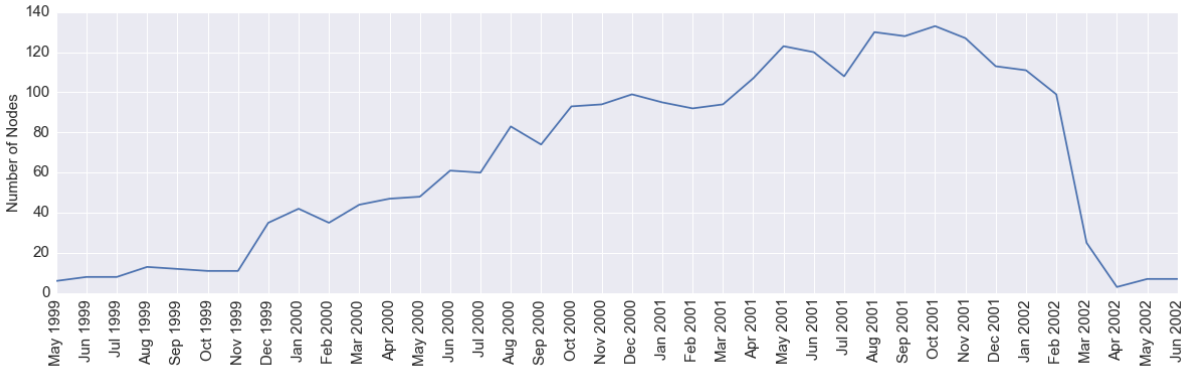

2000
1500
1000
500
0
Singular Value
0
20
40
60
80
100
120
140
160
Singular Value Component

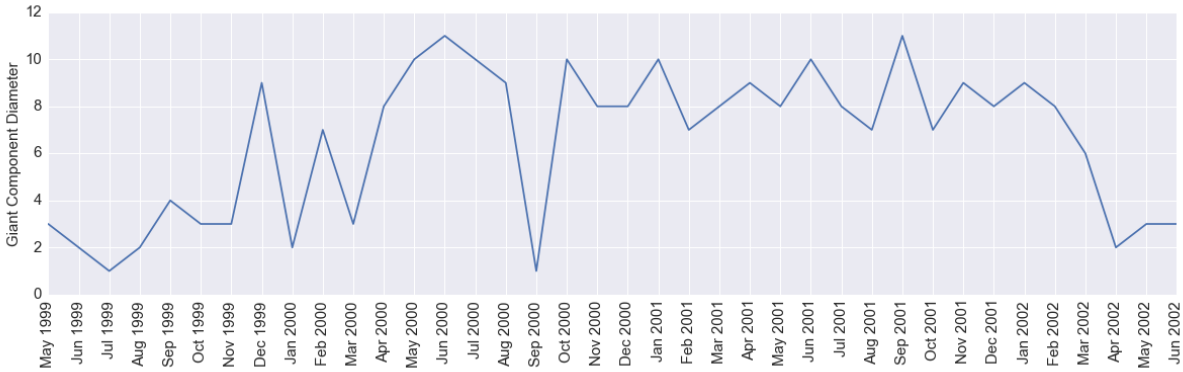


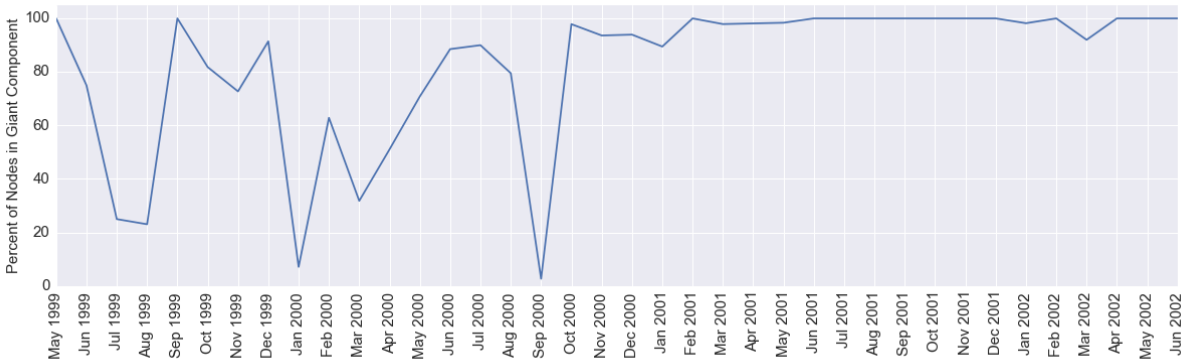
6 Evolution Across Time

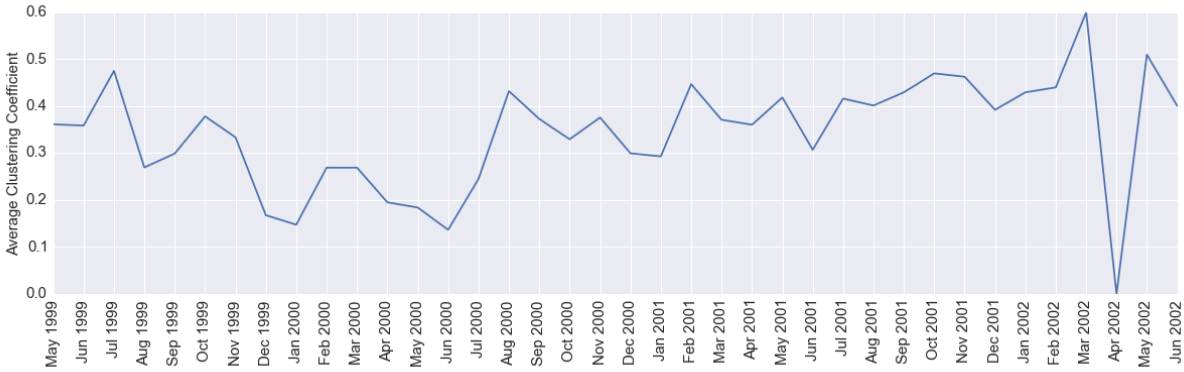
The Enron network was obviously an evolving network over time. This longitudinal information allows us to study how the network changed over time across various dimensions. We show a few of them below.

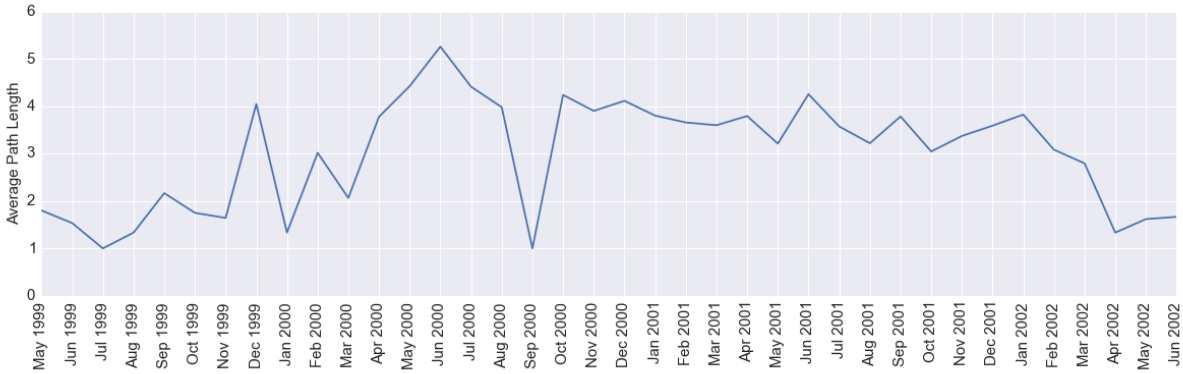


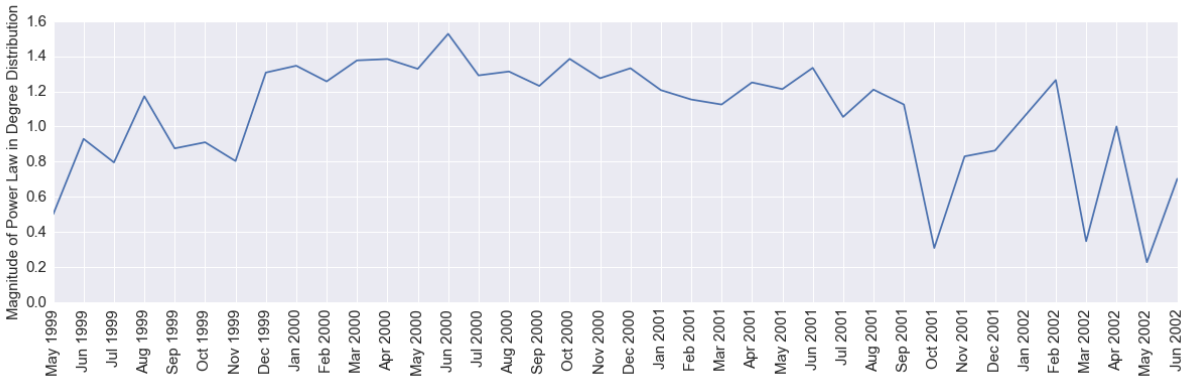

Number of Nodes
0
20
40
60
80
100
120
140
May 1999
Jun 1999
Jul 1999
Aug 1999
Sep 1999
Oct 1999
Nov 1999
Dec 1999
Jan 2000
Feb 2000
Mar 2000
Apr 2000
May 2000
Jun 2000
Jul 2000
Aug 2000
Sep 2000
Oct 2000
Nov 2000
Dec 2000
Jan 2001
Feb 2001
Mar 2001
Apr 2001
May 2001
Jun 2001
Jul 2001
Aug 2001
Sep 2001
Oct 2001
Nov 2001
Dec 2001
Jan 2002
Feb 2002
Mar 2002
Apr 2002
May 2002
Jun 2002


Giant Component Diameter
0
2
4
6
8
10
12
May 1999
Jun 1999
Jul 1999
Aug 1999
Sep 1999
Oct 1999
Nov 1999
Dec 1999
Jan 2000
Feb 2000
Mar 2000
Apr 2000
May 2000
Jun 2000
Jul 2000
Aug 2000
Sep 2000
Oct 2000
Nov 2000
Dec 2000
Jan 2001
Feb 2001
Mar 2001
Apr 2001
May 2001
Jun 2001
Jul 2001
Aug 2001
Sep 2001
Oct 2001
Nov 2001
Dec 2001
Jan 2002
Feb 2002
Mar 2002
Apr 2002
May 2002
Jun 2002


Percent of Nodes in Giant Component
0
20
40
60
80
100
May 1999
Jun 1999
Jul 1999
Aug 1999
Sep 1999
Oct 1999
Nov 1999
Dec 1999
Jan 2000
Feb 2000
Mar 2000
Apr 2000
May 2000
Jun 2000
Jul 2000
Aug 2000
Sep 2000
Oct 2000
Nov 2000
Dec 2000
Jan 2001
Feb 2001
Mar 2001
Apr 2001
May 2001
Jun 2001
Jul 2001
Aug 2001
Sep 2001
Oct 2001
Nov 2001
Dec 2001
Jan 2002
Feb 2002
Mar 2002
Apr 2002
May 2002
Jun 2002







7 Conclusion

References

[1] Shetty, J., Adibi, J. (n.d.). The Enron Dataset Database Schema and Brief Statistical Report

[2] https://www.cs.purdue.edu/homes/jpfeiff/enron.html data retrieved on Nov 11 2014

[3] http://www.nytimes.com/2006/01/18/business/worldbusiness/18iht-web.0117enron.time.html

[4] http://cis.jhu.edu/ parky/Enron/employees

[5] Networks: An Introduction by M.E.J Newman