

# Graph Theoretic and Spectral Analysis of Enron Email Data

Anurat Chapanond, Mukkai S. Krishnamoorthy, Bülent Yener \*

## Abstract

Analysis of social networks to identify communities and model their evolution has been an active area of recent research. This paper analyzes the Enron email data set to discover structures within the organization. The analysis is based on constructing an email graph and studying its properties with both graph theoretical and spectral analysis techniques. The graph theoretical analysis includes the computation of several graph metrics such as degree distribution, average distance ratio, clustering coefficient and compactness over the email graph. The values of metrics in Enron email graph are compared to those in another email data set. It is shown that the degree distribution of the Enron email graph obeys the power law and there is a giant component that contains 62% of the nodes. The spectral analysis shows that the email adjacency matrix has a rank-2 approximation.

## 1 Introduction

There has been an increasing research focus on identifying communities within social networks and modeling their evolution over time. Real data for social network analysis can be obtained from email communications, chat-friendship (i.e., buddy list) lists, or from a non-electronic medium such as membership of clubs or board of directors of Fortune-500 companies.

In this paper, we consider the Enron email data set; this is the only substantial collection of real email data set that is public [8]. We provide both graph-theoretic and spectral analysis of the data set

to identify and quantify its structural information. Our approach is based on constructing an adjacency matrix representing the email communication graph. We compute interesting graph properties, such as diameter, clustering coefficient and betweenness of the Enron email graph. We compare the graph properties of Enron email graph with the RPI email graph [2]. The comparison shows that there are both similarities (e.g., both degree distributions obey the power law) and differences (e.g., different density of connectivity among communities of practice) between them. We also perform spectral analysis of the email data (as a matrix). We show that this matrix has a low rank-2 approximation.

There has been prior work on Enron data. In [10] authors automate classification of email messages into user-specific folders and extract from chronologically ordered email streams using SVM(Support Vector Machines). In [5] authors construct a database and provide a brief statistical report. In cite[9] language usage in a social network is studied. In [6] email response times are predicted from email logs.

This paper is organized as follows. In Section 2 we explain how to process the email data set to construct an undirected simple graph (i.e., without self loops). In Section 3 we introduce graph metrics and compare their values to that in RPI email graph. Section 4 presents the spectral analysis and show that rank-2 approximation is possible. In Section 5 we display the email graph using a novel visualization tool. Section 5.1 forms the conclusion.

## 2 Data Pre-Processing

Enron email data are stored in text file format [9]. There were 150 employees from Enron with email logs recorded during a 19 month period (from De-

---

\*Department of Computer Science Rensselaer Polytechnic Institute, Troy, NY 12180, email(chapaa; moorthy; yener)@cs.rpi.edu. This research is supported in part by NSF ITR Award #0324947.

ember 1999 to June 2001). Each log file contains email headers e.g. Message-ID, Date, From, To, Subject and email content. The attachments, although specified by X-Filename, are not included in the log.

## 2.1 Resolving Multiple Email Address

We extracted the From and To fields of email headers to build sender- and receiver-email list.

However, there could be several email addresses for an employee, thus we first identify all the email addresses of the same person. For example the following email addresses belong to the same person: vince.kaminski@enron.com, vince.j.kaminski@enron.com, vince.j.kaminski@enron.com, j.kaminski@enron.com, kaminski@enron.com, vincent.j.kaminski@enron.com, j'.kaminski@enron.com, j.kaminski@enron.com.

While some of these email addresses could be identified automatically, manual inspection is necessary to handle the employees with the same last name or unexpected characters in the emails.

## 2.2 Construction of the Email Graph

A matrix of number of emails that are sent between Enron employees is constructed. The matrix can be used to construct a directed simple graph, in which vertices represent employees and undirected edges are added between employees who corresponded through email<sup>1</sup>. However we constructed an undirected simple graph using the following threshold; the minimum number of emails between each employee and the minimum number of emails sent by each of them.

### Choice of Threshold

The undirected email graph is constructed as follows: in order for two employees to be connected by an edge in the graph two criterion must be met:

1. The employees must have exchanged at least 30 emails with each other.

<sup>1</sup>We construct an email graph without processing the email content to minimize the privacy concern.

T1	T2				
	0.05	0.10	0.15	0.20	0.25
25	-0.80	-0.83	-0.92	-0.95	-1.05
30	-0.87	-0.89	-1.01	-1.05	-1.15
35	-0.95	-1.06	-1.13	-1.18	-1.30
40	-1.01	-1.09	-1.24	-1.31	-1.41
45	-1.07	-1.20	-1.33	-1.46	-1.52

Table 1: Exponent value of power law degree distribution on different thresholds T1 and T2.

2. Each member of the pair has sent at least 6 emails to the other (to reduce the number of one-way relationships).

Changing the value used for each criterion will change the structure of the email graph. We found that the degree distribution of the email graph obeys the power law as shown in Figure ?? . We investigate the degree distribution of the email graph constructed by different thresholds.

We found that by varying the threshold we can construct an email graph with varying exponent value of the power law degree distribution. Table 1 shows different exponent value of the power law degree distribution for different thresholds. In the table, T1 is the minimum number of emails between employees and T2 is the minimum percentage of T1 of emails sent by each of them. We chose T1 of 30 emails and T2 of 20% or 6 emails. The resulting graph has the exponent value of -1.05.

We note that in [1] authors also used T1= 30 and T2= 5 emails as threshold values.

## 3 Structural Analysis with Graph Metrics

In this section we investigate the properties of Enron email graph with respect to some graph metrics and present a comparison to RPI email graph [2].

### 3.0.1 Graph metrics

The graph metrics we consider in this paper are degree distribution, diameter, average distance, average distance ratio, compactness, clustering coefficient, betweenness, relative interconnectivity, and relative closeness. We compare the values for two different email graphs, namely, Enron email graph with 150 nodes and RPI email graph with 1681 nodes. RPI email data set is constructed from a full SMTP (Simple Mail Transport Protocol) feed at Rensselaer Polytechnic Institutes central mail servers during a 24-hour period on 01/05/2004. Personally identifiable information in the logs was obscured using the HMAC message authentication protocol with a 128bit SSH1 hash as described in [2]. There are two differences between the construction of current RPI email data set and the one used in [2]: (i) the current set excludes the emails from and to outside of RPI domain, and (ii) it is subject to the thresholding as explained above.

**Degree distribution** - Degree distribution is the histogram of the degree of vertices in the graph. Degree distribution of an email graph reflects the power law property of the graph. It is used to determine an appropriate threshold for constructing the email graph. The degree distribution log graph for Enron and RPI email graph are shown in Figure ??

**Diameter** - Diameter is the longest of the shortest paths between any pair of vertices in a connected graph. It reflects how far apart two vertices are (from each other) in the graph. We computed the diameter of the giant component for both the Enron and the RPI email graphs. The Enron graph has a 9 diameter and the RPI graph has 27. The RPI graph has higher value of diameter than the Enron graph because the RPI graph has about ten times more number of vertices. We note that the diameters are surprisingly high in both graphs with respect to the number of vertices.

**Average distance (*AvgDist*)** - Average distance is the average length of shortest path between each vertex in the graph. The vertices that do not have a shortest path between them will be given the number of vertices in the graph as the length of their shortest path.

**Average distance ratio** - Average distance ratio is defined as  $\frac{NodeNo - AvgDist}{NodeNo}$  where *NodeNo* is the total number of vertices in the graph. Average distance ratio can have value between 0 and 1. The graph with only isolated vertices will have the average distance ratio of 0 and the complete graph will have the average distance ratio of 1. Average distance ratio reveals the spanning of edges in the graph; the more spanning the graph is the higher the value of average distance ratio. The value for the Enron graph is 0.36 and for the RPI graph is 0.11. This may indicate that the Enron email graph reflects the organizational structure.

**Compactness** - Compactness is the ratio between the number of existing edges and the number of all possible edges  $\frac{2E}{N^2 - N}$  where *E* is the total number of edges and *N* is the total number of vertices in the graph. Compactness can have value between 0 and 1. The graph with only isolated vertices will have the compactness of 0 and the complete graph will have the compactness of 1. Compactness is the statistic that is not affected by the structure of the graph since only the number of edges is used to compute. The value for Enron graph is 0.0067 and for the RPI graph is 0.0006. We note that the denominator has  $N^2$ , therefore the value of compactness is heavily affected by the size of the graph.

**Clustering coefficient** - Clustering coefficient  $C_i$  is defined as the percentage of the connections between the neighbors of vertex *i*, i.e.  $C_i = \frac{2 \cdot E_i}{k \cdot (k-1)}$  where *k* is the number of neighbors of vertex *i* and  $E_i$  is the number of existing connections between its neighbors. Clustering coefficient is the average value of  $C_i$  for all vertex *i* [2]. Clustering coefficient reflects the connectivity information in the neighborhood environment of a vertex. It provides the transitivity information since it controls whether two different vertices are connected or not, assuming that they are connected to the same vertex. The value for the Enron graph is 0.033 and for the RPI graph is 0.119.

**Betweenness** - The betweenness of an edge is defined as the number of shortest paths that traverse it [1]. The edge with high betweenness is said to be the inter-community edge where the edge with low betweenness is said to be the intra-community edge. By

repeatedly removing an edge with high betweenness the resulting graph will contain a group of clusters where each cluster represents a community of practice [1].

**Relative interconnectivity**  $RI(C_i, C_j)$  between two clusters  $C_i$  and  $C_j$  is defined as the absolute interconnectivity between  $C_i$  and  $C_j$ , normalized with respect to the internal interconnectivity of the two clusters  $C_i$  and  $C_j$  [3].

$$RI(C_i, C_j) = \frac{|EC_{\{C_i, C_j\}}|}{|EC_{C_i}| + |EC_{C_j}|}$$

Where  $EC_{\{C_i, C_j\}}$  is the edge-cut of the cluster containing both  $C_i$  and  $C_j$  so that the cluster is broken into  $C_i$  and  $C_j$ , and  $EC_{C_i}$  ( $EC_{C_j}$ ) is the size of its min-cut bisector for cluster  $C_i$  ( $C_j$ ).

**Relative closeness** -  $RC(C_i, C_j)$  between a pair of clusters  $C_i$  and  $C_j$  is the absolute closeness between  $C_i$  and  $C_j$ , normalized with respect to the internal closeness of the two clusters  $C_i$  and  $C_j$  [3].

$$RC(C_i, C_j) = \frac{\bar{S}_{EC_{\{C_i, C_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{C_i}} + \frac{|C_j|}{|C_i|+|C_j|}\bar{S}_{EC_{C_j}}}$$

Where  $\bar{S}_{EC_{\{C_i, C_j\}}}$  is the average weight of the edges that connect vertices in  $C_i$  to vertices in  $C_j$  and  $\bar{S}_{EC_{C_i}}$  ( $\bar{S}_{EC_{C_j}}$ ) is the average weight of the edges that belong to the min-cut bisector of cluster  $C_i$  ( $C_j$ ).

Relative interconnectivity and relative closeness are metrics used to determine the similarity in graph structure between two clusters. In this paper we use the metrics to determine the similarity of community of practice in the graph. By the definition of relative closeness, our graph, an undirected simple graph with equal edge weights, will always have the value of 1 for relative closeness of any clusters. The connectivity between each cluster is also of interest. It can be used to analyze the pattern or type of community of practice in the graph.

### 3.1 Comparison of the Enron and RPI data sets

Table 2 shows the comparison of graph properties and metrics between the Enron and RPI graphs.

Graph properties and metrics	Enron	RPI
Number of vertices	150	1681
Number of edges	52	1932
Number of connected components	57	290
Size of giant component	93	535
Diameter	9	27
Average Distance Ratio	0.36	0.11
Compactness	0.0067	0.0006
Clustering Coefficient	0.033	0.119

Table 2: The comparison of graph properties and metrics for Enron and RPI data.

Different values from the metrics suggested that these two graphs have different organizational structures. We found that the Enron graph has a smaller giant component than RPI graph because of its smaller size. The giant component in the Enron graph contains 62% of the vertices. The Enron graph structure, with higher value of average distance ratio and compactness; seems to be more clustered than the RPI graph. However, the clustering coefficient shows that RPI graph is more clustered. We show in section 3.3 that this conflict can be explained by the analysis of their clusters.

The degree distribution for the Enron and the RPI graph are shown in Figure ???. These show that both graphs obey power law distribution.

### 3.2 Graph Clustering

We constructed the communities of practice from the Enron graph by the algorithm described in [1]. The algorithm is a clustering method that repeatedly removes an edge of the graph by betweenness metric until the graph reaches stopping criteria. The edge with highest betweenness will be removed until the component size is less than 6 or all edges in the component has betweenness less than the number of vertices in the component minus one. We then calculated relative interconnectivity between each cluster.

The Enron graph has 27 communities of practice excluding all communities with only one vertex. There are 50 links (relative interconnectivity between

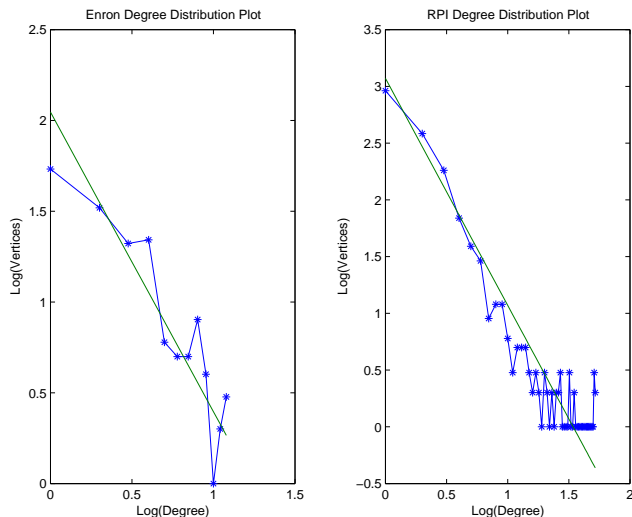


Figure 1: The log-log degree distribution plot for the Enron and the RPI email graph.

two clusters more than zero) between its communities. However the RPI graph has 472 communities of practice and there are only 212 links. We also found many cliques in RPI community of practice. In summary we found that the connectivity inside the communities of practice in Enron graph is sparser than that in the RPI graph but the connectivity between communities of practice in the Enron graph is denser than that in the RPI graph. This explains the conflict when comparing Enron and RPI graph metrics. Enron graph has a sparser connectivity inside the communities which results in a lower value of clustering coefficient but with a denser connectivity between communities Enron graph has a higher value of average distance ratio and compactness. We also found the pattern of the connectivity in RPI graph. We found that all the communities of practice have only a few links to the other communities. This is because the vertices mostly represent students or teachers and they are bound by the number of classes they involve. However Enron graph has different pattern; some communities could have high number of links where some communities have small number of links. Therefore we conclude that we can analyze pattern or type of community from the metric relative inter-

connectivity.

## 4 Spectral Analysis of the Enron Data Set

In this section, we perform a spectral analysis on Enron email data similar to what was done with the RPI email data [2]. We show that the Enron email matrix has also a low rank (i.e., rank 2) approximation. This is accomplished by performing Singular Value Decomposition [14] of the Enron email matrix (that was done using the preprocessing steps mentioned in the earlier sections). We also perform a simple clustering of the data based on the low rank approximation.

In matrix notation, SVD for Enron email matrix of  $m \times m$  is defined as  $A = U\Sigma V^T$  where  $U$  and  $V$  are orthogonal (thus  $U^T U = I$  and  $V^T V = I$ ) matrices of dimensions  $m \times r$  and  $m \times r$  respectively, containing the left and right singular vectors of  $A$ .  $\Sigma = \text{diag}(\sigma_1(A), \dots, \sigma_r(A))$  is an  $r \times r$  diagonal matrix containing the singular values of  $A$ . SVD has been extensively used in analyzing large data [5]. The plot of the singular values are shown in Figure 4.

The largest two singular values of the Enron email matrix are 2277 and 1550 and the rest of the singular values are much smaller than these two values. So, we claim that Enron email matrix has a low rank (2) approximation. In other words, all the entries in the Enron email matrix can be approximately obtained using two principal components.

Once we obtained that the matrix has a low rank approximation, we projected the matrix in each of the dimensions. Plotting the data in the first dimension, we computed three clusters in the first dimension. The first cluster consisting of indices 20,44,57 and 126, which are Jeffrey Dasovich, Mary Hain, Steven Kean, and James Steffes, the second consisting of indices 1,8,23,43,56,61,63,73,105,109,117 and 133, which are Philip Allen, Sally Beck, David Delainey, Mark Haedicke, Wincente Kaminski, Louise Kitchen, John Lovorato, Kay Mann, Elizabeth Sager, Richard Sanders, Richard Shapiro, and Mark Taylor, and the third cluster containing the rest of

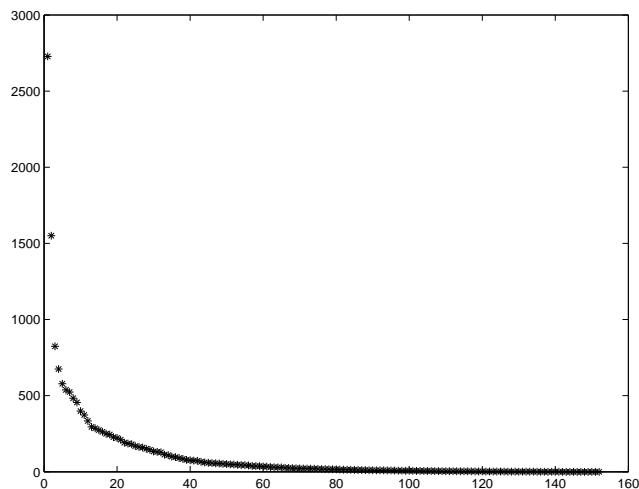


Figure 2: The singular values of Enron email matrix shows that largest two singular values will be sufficient for noise reductions and extracting the structure.

the indices. Plotting the data in the second dimension, we computed three clusters. The first cluster consists of indices 55,115,125,135, which are Tana Jones, Sara Shackleton, Carol St Clair, Paul Thomas, the second cluster consisting of indices 8,43,47,54,73,87,90,105,109, which are Sally Beck, Mark Haedicke, Marie Heard, Kay Mann, Stephanie Panus, Debra Perlingiere, Elizabeth Sager, Richard Sanders and the third cluster containing the rest of the indices. The clusters that are obtained using this are more or less consistent with the clusters that are obtained using the graph model. Finally, we show the actual distribution of the entries of the matrix projected into the two dimension in the next Figure 3

## 5 Visualization: Email Graph to Organization Hierarchy

The following image 4 shows the visualization of the Enron graph. The layout was done with GraphDraw, a graph tool in Java [13] The visualization is automatically created by using a force-directed algorithm from email graph.

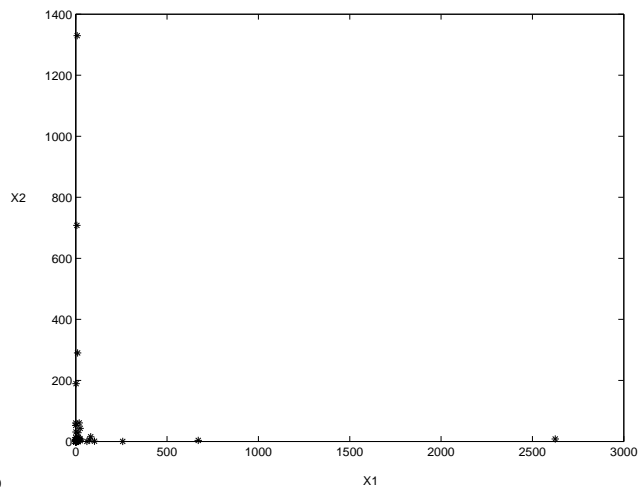


Figure 3: Projection of entries in Rank-2.

Each vertex will try to push the other vertices away while each edge acts like a spring that pulls the vertices together. The graph has been color-coded by cluster of community of practice. The vertices with the same color are in the same community of practice. The giant connected component of the Enron graph is shown but some isolated vertices are omitted.

Visual inspection of the graph reveals the organization leadership tends to end up in the center. We did not know the hierarchy of the Enron organization however we looked at the highly paid executives [8]. We found that the resulting email graph showed somewhat the hierarchy of the organization.

Using a BFS algorithm a spanning tree with the root of the tree being the vertex corresponding to Enron CEO (level 0). We found that the level of vertices corresponds to the salary of the employee; i.e. the higher payment an employee receives, the lower level (smaller number) the vertex is.

### 5.1 Summary and Conclusions

In this paper we construct a graph from the Enron email data set and analyze its both graph theoretical and spectral properties. We also compare the En-

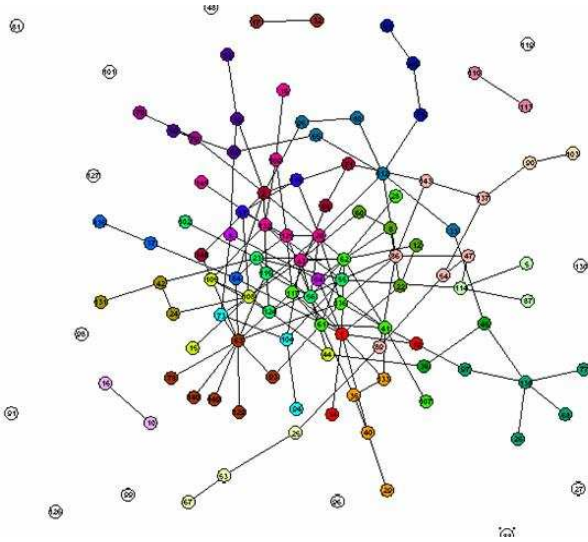


Figure 4: The visualization for Enron email graph color-coded by the cluster of community of practice.

ron email graph to the RPI email graph. Some of the observations can be summarized as follows: The degree distribution of the Enron email graph obeys power law and there is a giant component that contains 62% of the vertices. The graph metrics considered for analyzing the properties of email graphs are useful to capture the social structure. For example based on the *betweenness* metric we observe that the connectivity between communities of practice in the Enron email graph is denser than that in the RPI email graph. Furthermore, in the Enron graph some communities have a high number of links while other communities have a small number of links. This is in contrast with the RPI email graph in which communities of practice have only a few links to the other communities. This may be because the vertices mostly represent students or faculty and the communities are related to the classes. Thus the metric *relativeinterconnectivity* can be used to analyze the pattern or type of community.

The visualization of the email graph shows somewhat the hierarchy of the organization with respect to the salary structure.

We also investigate whether there is any signifi-

Employee	Payment	Level
Kenneth Lay	\$103,559,793.00	0
Philip Allen	\$4,484,442.00	1
David Delainey	\$4,749,979.00	2
Mark Haedicke	\$3,859,065.00	2
Louise Kitchen	\$3,471,141.00	2
Rick Buy	\$2,355,702.00	2
Wincenty Kaminski	\$1,085,821.00	2
Richard Shapiro	\$1,057,548.00	2
Mitchell Taylor	\$1,092,663.00	2
Sally Beck	\$969,068.00	2
John Lavorato	\$10,425,757.00	3
Jeffrey Shankman	\$3,038,702.00	4
Michael Mcconnell	\$2,101,364.00	4
Steven Kean	\$1,747,522.00	4
James Derrick	\$550,981.00	4
Roderick Hayslett	\$0.00	6

Table 3: The payment and spanning tree level for each Enron executives.

cant link between Enron employees and people from White House. We add a vertex that represents people from White house, e.g. president@whitehouse.gov, vice.president@whitehouse.gov. Our preliminary investigation shows that there are emails being sent and received between Enron employees and White House during the logging period but after the filtering process there is no link between this group of Enron employees and White House people. We also examined the link between Enron employees and the six people who had been prosecuted - Sheila Kahanek, Dan Boyle, Daniel Bayly, Robert Furst, William Fuhs, and James Brown. By adding another vertex representing these people we found that there is no link between them and this group of Enron employees.

**Acknowledgments:** The authors would like to thank Michael D. Sofka for providing the RPI email data set.

## References

- [1] Tyler, J. R., Wilkinson, M. D., and Huberman. B. A., "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations", in Pro-

- ceeding of the International Conference on Communities and Technologies, Netherlands, kluwer Academic Publishers (2003).
- [2] Drineas, P., Krishnamoorthy, M. S., Sofka, M. D., and Yener, B., "Studying E-mail Graphs for Intelligence Monitoring and Analysis in the Absence of Semantic Information", 2004.
  - [3] Karypis, G., Han, E.-H., and Kumar, V., "CHAMELEON: A hierarchical clustering algorithm of spatial data", In Proc. 8th Symp. Spatial Data Handling, pages 45-55, Vancouver, Canada, 1998.
  - [4] Chapanond, A., and Krishnamoorthy M. S., "User Classification for P2P network", manuscript (2004).
  - [5] Han, J., and Kamber, M., Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, 2001.
  - [6] Adibi, J., and Shetty, J., The Enron Email Dataset Database Schema and Brief Statistical Report, [http://www.isi.edu/~adibi/Enron/Enron\\_Dataset\\_Report.pdf](http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf).
  - [7] Kalman, Y., Rifacli, S., Email Chronemics: Unobtrusive Profiling of Response Times, HICSS-38, Hawaii, 2005.
  - [8] Houston Chronicles, <http://www.chron.com/content/chronicle/special/01/enron/index.html>.
  - [9] Enron Email Dataset, <http://www-2.cs.cmu.edu/~enron/>.
  - [10] Corrada-Emmanuel, A., McCallum, A., and Wang, X., Language Use in a Social Network: The Enron Email Dataset, CNLP Seminars, 2004.
  - [11] Klimt, B., and Yang, Y., The Enron Corpus: A New Dataset for Email Classification Research, To be published in proceedings of the European Conference on Machine Learning (ECML), 2004.
  - [12] Loch, C. H., Tyler, J. R., and Lukose, R., "Conversational Structure in Email and Face to Face communication", Draft, submitted to Organization Science.
  - [13] N. Preston and M. Krishnamoorthy, "GraphDraw- A Graph Drawing System to study Social Networks," Unpublished Manuscript, Rensselaer Polytechnic Institute, Troy, NY, 2004.
  - [14] G. Golub and F. Van Loan, "Matrix Computations", Johns Hopkins University Press, 1984.