
Computational & Mathematical Organization Theory, 11, 265–281, 2005

Graph Theoretic and Spectral Analysis of Enron Email Data

ANURAT CHAPANOND
MUKKAI S. KRISHNAMOORTHY
BÜLENT YENER
Department of Computer Science Rensselaer Polytechnic Institute, Troy, NY 12180
email: chapaa@cs.rpi.edu
email: moorthy@cs.rpi.edu
email: yener@cs.rpi.edu

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 spectral analysis shows that the email adjacency matrix has a rank-2 approximation. It is shown that preprocessing of data has significant impact on the results, thus a standard form is needed for establishing a benchmark data.

Keywords: email graph, graph metrics, spectral analysis, social network analysis

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 (http://www.chron.com/content/chronicle/special/01/enron/index.html). 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 compute the graph properties of Enron email graph and 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 Corrada-Emmanuel et al. (2004) the authors automate classification of email messages into user-specific folders and extract from

chronologically ordered email streams using SVM (Support Vector Machines). In Han and Kamber (2001) the authors construct a database and provide a brief statistical report. In http://www-2.cs.cmu.edu/∼enron/ language usage in a social network is studied. In Adibi and Shetty (http://www.isi.edu/ adibi/Enron/Enron_Dataset_Report.pdf) email response times are predicted from email logs. In Diesner and Carley (2005) authors investigate the Enron email data set from a social network analytic perspective; various network analytic techniques are applied. In Priebe et al. (2005) the authors introduce a theory of scan statistics on graphs and apply the ideas to the problem of anomaly detection in a time series of Enron email graphs. In McCallum et al. (2005) the authors apply the Author-Recipeint-Topic (ART) model for social network analysis on Enron email data set. In Browne and Berry (2005) the authors apply a non-negative matrix factorization approach for the extraction and detection of concepts or topics on Enron email data set. And in Keila and Skillicorn (2005) authors investigate the structures present in the Enron email data set using singular value decomposition and semidiscrete decomposition.

This paper is organized as follows. In Section 2 we explain how to process email data set to construct a directed simple graph (i.e., without self loops), present the spectral analysis and show that rank-2 approximation is possible, and apply different filters to reduce noise. In Section 3 introduce graph metrics and comute their values. Section 4 we present the comparison of Enron directed and undirected graph, and the impact of SVD-based filtering and graph-based filtering. In Section 5 we display the email graph using a novel visualization tool. Section 6 concludes the work.

2. Data Processing

Enron email data are stored in text file format (http://www-2.cs.cmu.edu/∼enron/). There were 150 employees from Enron with email logs recorded for 3.5 years (from October 1998 to June 2002). 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 Graphs*

Using the emails sent between Enron employees we construct a directed simple graph G_0, in which vertices represent employees and a directed edge is established from a node with *from* address to another node with *to* address.[1] G_0 has 152 nodes 1895 edges. We note that G_0 may contain some noise since it considers every email sent; thus, we derived two graphs from it using two noise reduction techinques.

The first filter uses a threshold, based on the minimum number of emails between employees and the minimum number of emails sent by each of them, to produce an undirected simple graph G_u with 152 nodes 152 edges.

The second filter is SVD-based which produces a directed graph G_d by removing edges between nodes which are considered noise nodes by SVD method on the adjacency matrix A of G_0. The graph G_d has 152 nodes 1874 edges.

2.2.1. Threshold-Based Noise Filtering. 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:

T1: The employees must have exchanged at least 30 emails with each other.
T2: Each member of the pair has sent at least 6 emails (20%) to the other (to reduce the number of one-way relationships).

We note that in Tyler et al. (2003) authors also used $T1 = 30$ and $T2 = 5$ emails as threshold values. The thresholds are chosen to a) remove some edges in the email graph, and b) to construct an undirected graph. By removing edges with small number of emails we enhance the real connection between people; the edges with small number of emails are considered as noise here. We are also interested in the interaction between people. The threshold we use to construct the undirected graph emphasizes an interaction by considering two-way communication. Clearly this techniques induces a sparse graph which has 57 connected components thus it aids the clustering approach to identify communities.

2.2.2. SVD-Based Noise Filtering. We perform a spectral analysis on Enron email data similar to Drineas et al. (2004). We show that the Enron email matrix has also a low rank (i.e., rank 2) approximation by computing Singular Value Decomposition (SVD) Golub and Van Loan (1984) of the $m \times m$ adjacency matrix A of Enron email graph G_0.

In matrix notation, SVD for the matrix A 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. $\Sigma = \mathbf{diag}(\sigma_1(A), \ldots, \sigma_r(A))$ is an $r \times r$ diagonal matrix containing the singular values of A.

The plot of the singular values are shown in figure 1. The largest two singular values of the Enron email matrix A are 1277 and 1550 and the rest of the singular values are much smaller than these two values. Thus, we observe 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.

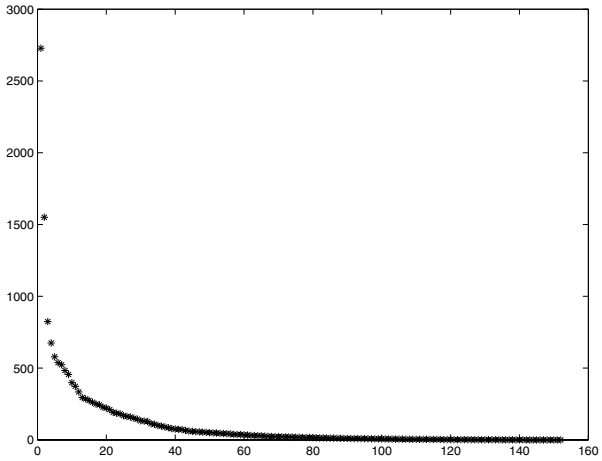

Figure 1. The singular values of Enron email matrix A shows that largest two singular values will be sufficient for noise reductions and extracting the structure.

We apply SVD-based filtering method by removing edges between nodes which are considered noise edges by SVD method. The resulting graph is used to create clusters with different metrics. We then compared these with clusters from graph method using undirected graph in which noise is removed by applying the thresholds we have described.

3. Properties of Enron Email Graph

In this section we investigate the properties of Enron email graph with respect to some graph metrics.

3.1. Degree Distribution of Enron Email Graph

We examine the power law property of the Enron graphs G_0 and G_u. The degree distribution for undirected graph G_u, in-degree distribution for directed graph and out-degree distribution for directed graph G_0 are plotted using exponential binning procedure described in Newman (2003) as shown in figure 2 in log-log scale.

Unlike Tyler et al. (2003) the undirected email graph plot does not obey power law distribution, moreover all of the Enron plots do not show a straight line thus not obeying power law distribution.

3.2. 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.

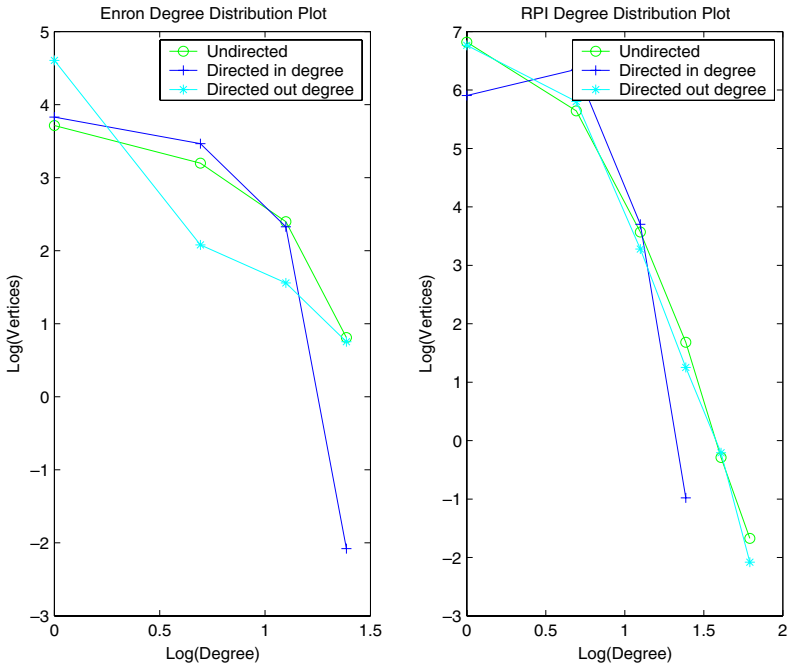

Figure 2. The log-log degree distribution plots for the Enron email graphs G_u and G_0.

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 email graphs is shown in figure 2.

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.

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.

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. 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. $C_i = \frac{2.E_i}{k.(k-1)}$ 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 (Drineas et al., 2004). 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.

Betweenness—The betweenness of an edge is defined as the number of shortest paths that traverse it (Tyler et al., 2003). 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 (Tyler et al., 2003).

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 Karypis et al. (1998).

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

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 Karypis et al. (1998).

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

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 clusters is also of interest. It can be used to analyze the pattern or type of community of practice in the graph.

4. Results and Interpretations

4.1. Comparison of Enron Directed and Undirected Graphs

We compute several graph metrics on several email graph settings. We construct both directed graphs and undirected graphs. Several directed graphs are constructed by changing value of threshold "minimum number of emails" (T1). Several undirected graphs are constructed by changing value of threshold "minimum number of emails" (T1) and "minimum number of emails from one side" (T2). T2 is all set to be 20 minimum emails more than 0, the original graph is used; however, there have to be edges from both directions for constructing an edge in undirected graph and when no. of minimum emails is 0, new graph is constructed by adding only edges with the weight more than the threshold value. The number of edges resulted from varying threshold is shown in the following figure 3. As the threshold increases the number of edges in the graph decreases since less number of edges will exceed the threshold value. The undirected graph also has smaller number of edges than the directed graph with the same threshold because another threshold "minimum number of sent emails for both nodes" is used to ensure that each edge has bi-directional communication. We also compute average degree for each graph. The average degree for directed graph is the number of out-degree per node and the average degree for undirected graph is the number of degree per node. Since the numbers of

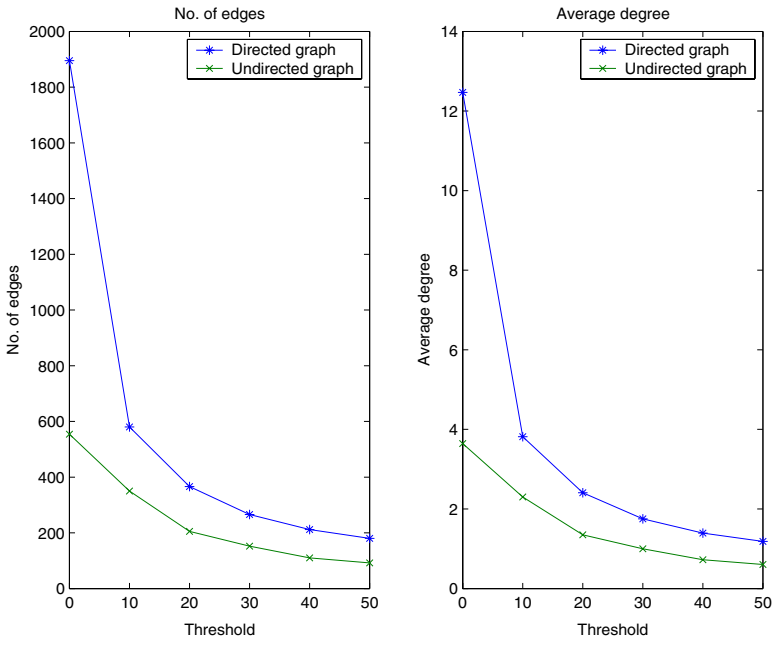

Figure 3. The Enron degree distribution plot for the directed and undirected graphs with different thresholds.

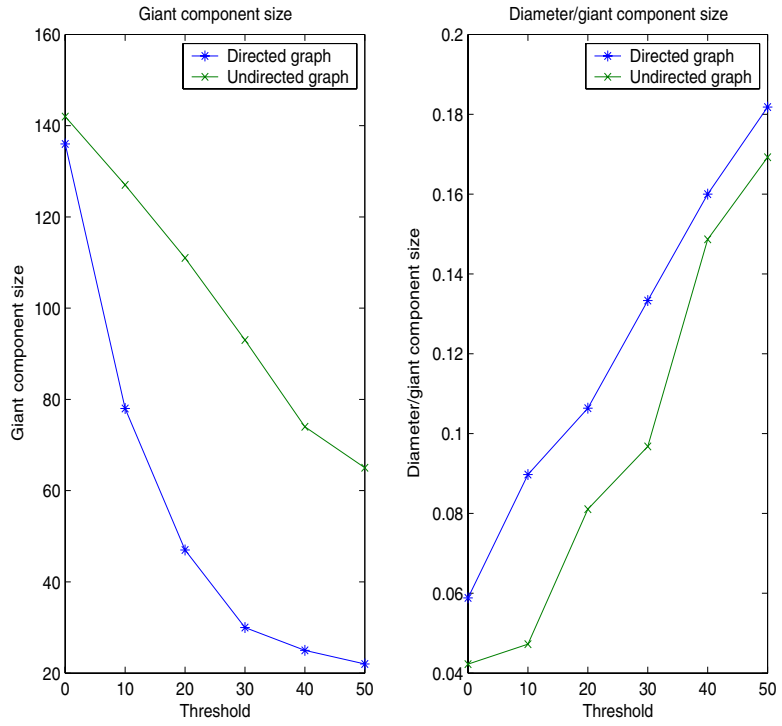

Figure 4. The enron diameter, giant component size and diameter/giant component size plot for directed and undirected graphs with different thresholds.

nodes are equal in all graphs the average degree graph looks similar to number of edges graph.

Figure 4 shows different plots on diameter, giant component size and diameter per giant component size. The threshold has different effect on diameter for different graphs. For directed graph we calculate diameter from strongly connected component and the diameter decreases when threshold increases and for undirected graph the diameter increases when threshold increases. For undirected graph the shortest path can be found easily because of its bi-directional property. With small threshold value there are many edges that act as shortcuts hence reduce the length of the path between nodes. When increasing the threshold value there are fewer shortcuts and the length of the path increases. For directed graph the node is unreachable when edges are removed therefore when threshold increases the diameter decreases. The giant component size decreases when the threshold increases as shown in the following. The undirected graph has stricter threshold than the directed graph and has the smaller size of giant component. The giant component size has a direct effect to the diameter because the diameter is calculated from the giant component. The diameter can be from one, in case of a complete graph, to the size of the giant component minus one. The following plot 4 shows diameter divided by giant component size on different

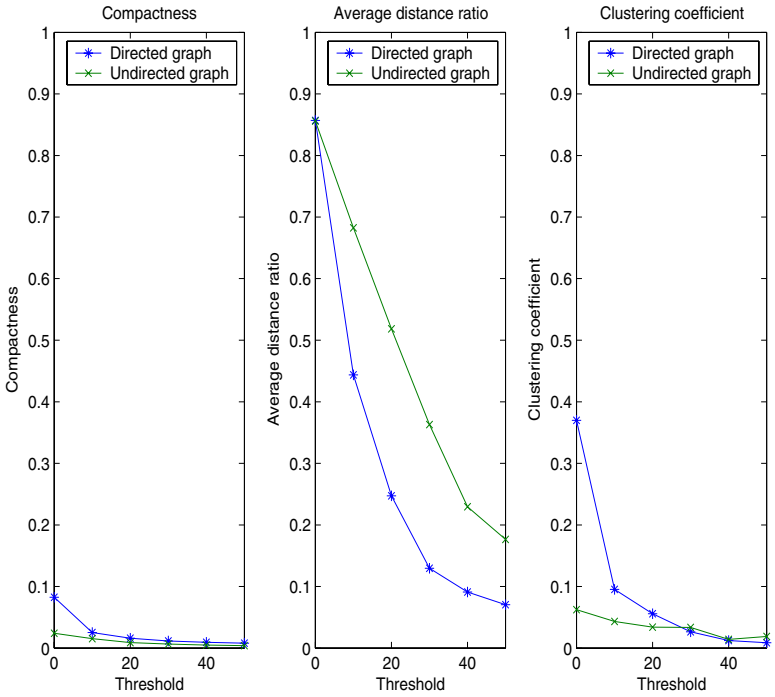

Figure 5. The enron compactness, average distance ratio, clustering coefficient plot for directed and undirected graphs with different thresholds.

threshold. The giant component size decreases when the threshold increases therefore it enhances the slope of both directed graph and undirected graph. Although the diameter changes differently for directed graph and undirected graph we found that both have plots for diameter per giant component size with about the same slope.

Three different metrics are used to compare the structure of the graphs. Figure 5 shows these metrics. Compactness is the number of all edges divided by all possible edges. For a directed graph the number of possible edges is twice of the same undirected graph. Therefore the undirected graph will have two times larger compactness as the same directed graph. However the following graph shows that when the threshold increases the compactness is equal for both directed and undirected graph. This is because the number of edges in directed graph is about twice of the number of edges in undirected graph. The average distance ratio is the probability of reachability from any other node. The plot shows that undirected graph is easier to reach than directed graph. Since the undirected graph has bi-directional property the average distance ratio is much higher than the directed graph. The average distance ratio also decreases when number of edges decreases. The clustering coefficient is the probability of the neighbors of a node forming a clique. The clustering coefficients for both directed graph and undirected graph are about the same when threshold increases. Therefore the clustering coefficient does not depend on the directionality of the graph.

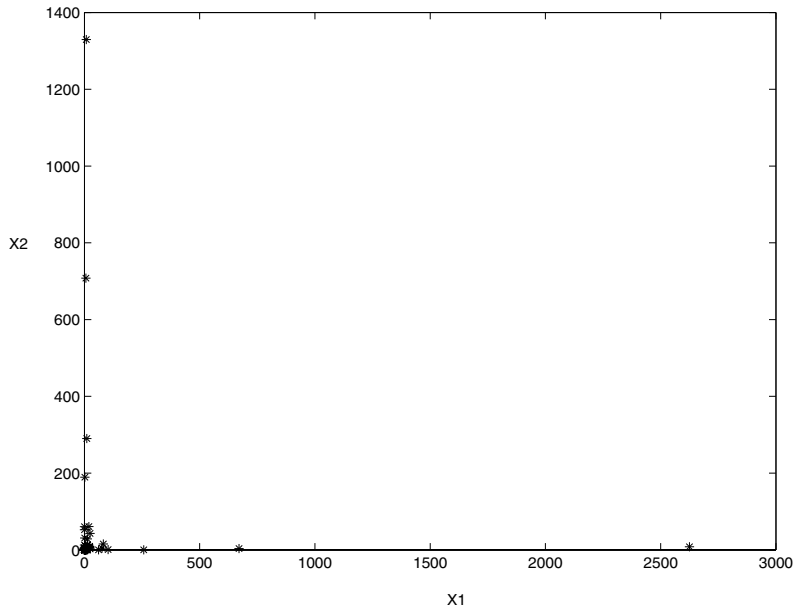

Figure 6. Projection of entries in Rank-2.

4.2. Cluster Analysis

4.2.1. SVD Based Clustering. SVD has been used extensively in analyzing large data sets (Han and Kamber, 2001). 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. Plotting the data in the second dimension, we computed another three clusters. Finally, we show the actual distribution of the entries of the matrix projected into the two dimension in the next figure 6.

Based on the SVD, we computed clusters from 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 the indices. We computed another three clusters from the second dimension. The first cluster consists of indices 55, 115, 125, and 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, and 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.

4.2.2. Clustering with Graph Metrics. We constructed the communities of practice from the Enron graph by the algorithm described in Tyler et al. (2003). 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 two clusters more than zero) between its communities. 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 the Enron graph has a higher value of average distance ratio and compactness. We also found that in the Enron graph 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 using the metric relative interconnectivity.

Different metrics are used to create clusters. Betweenness is used by betweenness algorithm (Tyler et al., 2003). Average Distance Ratio and Clustering Coefficient are used by K-mean clustering algorithm. In figure 7 we show the cluster size plottings using

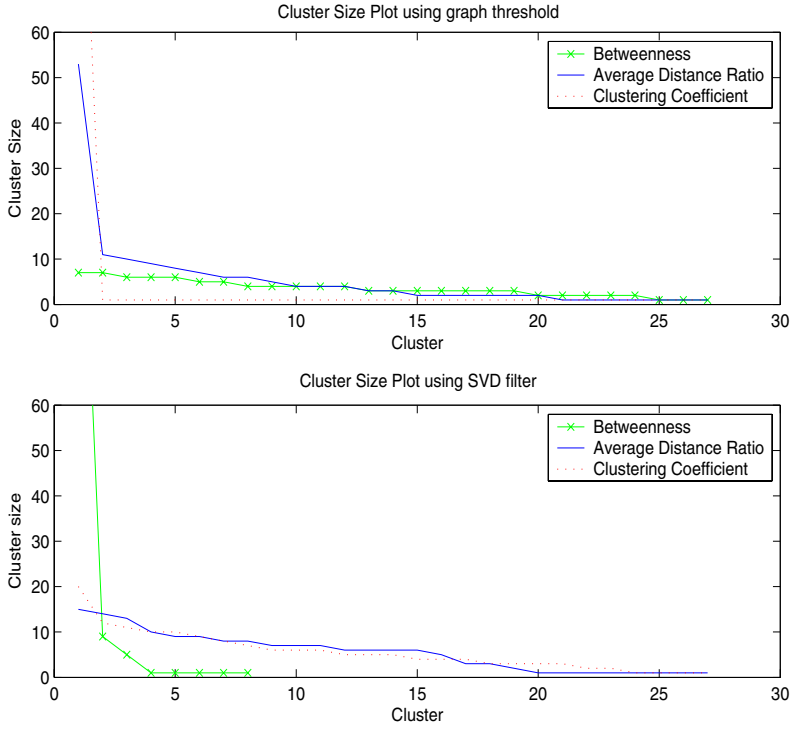

Figure 7. The cluster size ploting using graph threshold and SVD filter.

thresholding and SVD filtering methods. The largest values (omitted in the plotting) for clustering coefficient is 126 and the value in the SVD filter plot for betweenness is 132.

The figure shows that values of metrics vary depending on the filtering methods to process the email data. For example Betweenness metric with SVD filtering results in a big cluster whereas graph filtering creates many small clusters. In contrast, Clustering Coefficient with graph filter results in a big cluster where SVD filtering creates many small clusters. The Average Distance ratio metric seems to create many small clusters for both filters.

Table 1 shows the result of applying different metrics and filters on the Enron email graph. First we apply SVD method on the Enron email graph without filtering. We then apply several metrics i.e. Betweenness, Clustering Coefficient, and Average Distance Ratio on different filters i.e., threshold-based and SVD-based. The results are shown for the clusters with higest value of the metric used (C1 (cluster 1) to C5 (cluster 5) for each filtering method. Clearly, filtering of data has remarkable impact on the results since there is no agreement on clustering of users to identify communities.

5. Visualization: Email Graph to Organization Hierarchy

The following image (figure 8) shows the visualization of the Enron graph. The layout was done with GraphDraw, a graph tool in Java (Preston and Krishnamoorthy, 2004)

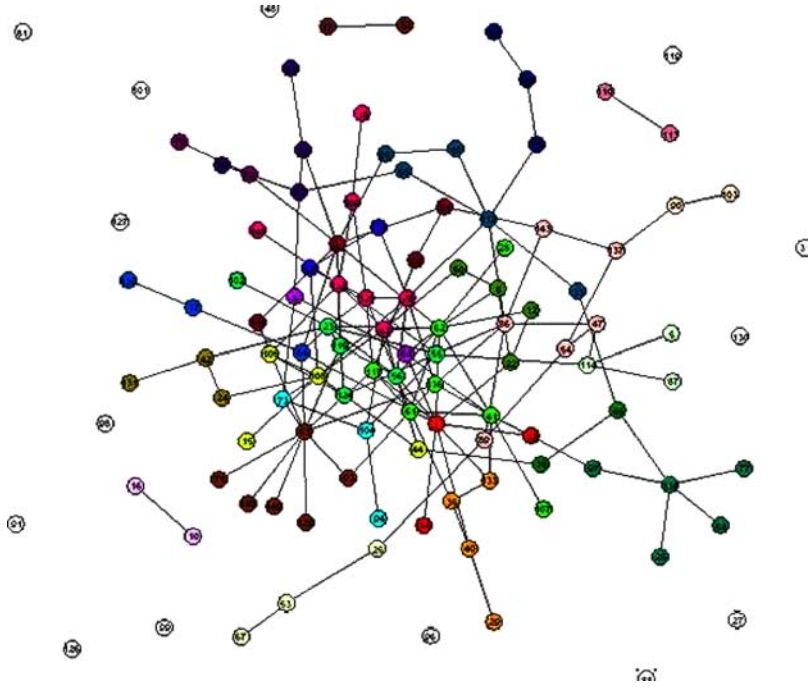

Figure 8. The visualization for enron email graph color-coded by the cluster of community of practice.

Table 1. The clusters resulted from different metrics and filters.

Metric	Filter	Clusters
SVD on G_0	N/A	C1[20, 44, 57, 126] C2[1, 8, 23, 43, 56, 61, 63, 73, 105, 109, 117, 115]
Cluster. Coeff (C.C.)	Tresholding $\rightarrow G_u$	C1[1–18, 20, 21, 24–38, 40-54, 57-60, 62, 64-69, 71–82, 84–106, 108–113, 115, 117–131, 133–140]
C. C.	SVD $\rightarrow G_d$	C1[123] C2[95] C3[10, 97] C4[5] C5[12, 22, 77, 68, 107]
Betweenness	Tresholding $\rightarrow G_u$	C1[1, 19, 34] C2[133, 35, 40, 29] C3[84, 37, 135] C4[61, 41, 62, 137, 117, 107, 28] C5[64, 6]
Betweenness	SVD $\rightarrow G_d$	C1[0–15, 17, 19–30, 32, 33, 35, 35-41, 43-66, 68-80, 84–101, 103–106, 108, 109, 111–113, 115–119, 121, 123–131, 133–145, 147, 149, 150]
Av. Dist. Ratio	Tresholding $\rightarrow G_u$	C1[61] C2[1, 62, 60, 114, 40, 42, 85, 124] C3[19, 135, 22, 55] C4[0, 131, 111, 116, 54, 83] C5[63, 107, 72, 103]
Av. Dis. Ratio	SVD $\rightarrow G_d$	C1[63] C2[8] C3[64] C4[1] C5[20, 57, 61, 94, 119, 23, 56]

Table 2. The payment and spanning tree level for each Enron executives.

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

The visualization is automatically created by using a force-directed algorithm from email graph.

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 http://www.chron.com/content/chronicle/special/01/enron/index.html. 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.

6. Summary and Conclusions

In this paper it is shown that the Enron email data has low rank approximation and preprocessing of data, to filter out noise, has significant impact on the properties of the graph

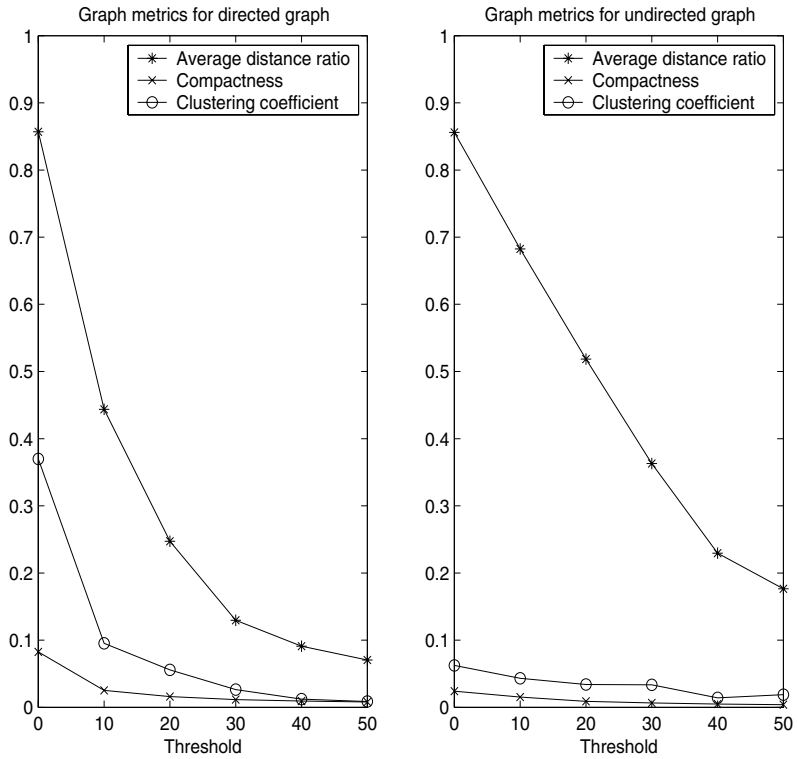

Figure 9. The metrics comparison for directed and undirected graphs.

representing email communications. In particular, identification of clusters representing tightly coupled users is very sensitive to the filtering of data.

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 dense. Furthermore, in the Enron graph some communities have a high number of links while other communities have a small number of links. Thus the metric *relative interconnectivity* can be used to analyze the pattern or type of community.

The graph metrics for directed and undirected graphs (figure 9) both have the same trend with average distance ratio the highest value, clustering coefficient second and compactness the lowest value. From these three metrics we conclude that the graph is well distributed; the edges connecting the nodes are dispersed over all nodes as we can see that the average distance ratio is high. The graph is sparse since the compactness has small value, the graph the clusters are not dense since the clustering coefficient also has small value.

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 significant 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 the 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

This research is supported in part by NSF ITR Award #0324947.

Note

1. We construct an email graph without processing the email content to minimize the privacy concern.

References

Adibi, J. and J. Shetty, The Enron Email Dataset Database Schema and Brief Statistical Report, http://www.isi.edu/ adibi/Enron/Enron_Dataset_Report.pdf.

Browne, M. and M.W. Berry (2005), "Email Surveillance Using Nonnegative Matrix Factorization," in *Proceeding of SIAM International Conference on Data Mining, SIAM Workshop on Link Analysis*, Counterterrorism and Security.

Chapanond, A. and M.S. Krishnamoorthy (2004), "User Classification for P2P Network," Unpublished Manuscript, Rensselaer Polytechnic Institute, Troy, NY.

Corrada-Emmanuel, A., A. McCallum, and X. Wang (2004), Language Use in a Social Network: The Enron Email Dataset, CNLP Seminars.

Diesner, J. and K. Carley (2005), "Exploration of Communication Networks from the Enron Email Corpus," in *Proceeding of SIAM International Conference on Data Mining, SIAM Workshop on Link Analysis*, Counterterrorism and Security.

Drineas, P., M.S. Krishnamoorthy, M.D. Sofka and B. Yener (2004), "Studying E-mail Graphs for Intelligence Monitoring and Analysis in the Absence of Semantic Information," in *IEEE International Conference on Intelligence and Security Informatics*.

Enron Email Dataset, http://www-2.cs.cmu.edu/ enron/.

Golub, G. and F. Van Loan (1984), *Matrix Computations*, Johns Hopkins University Press.

Han, J. and M. Kamber (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

Houston Chronicles, http://www.chron.com/content/chronicle/special/01/enron/index.html.

Kalman, Y. and S. Rafacli (2005), Email Chronemics: Unobtrusive Profiling of Response Times, HICSS-38, Hawaii.

Karypis, G., E.H. Han and V. Kumar (1998), "CHAMELEON: A Hierarchical Clustering Algorithm of Spatial Data," in *Proceedings of the 8th Symposium Spatial Data Handling*, Vancouver, Canada, pp. 45–55.

Keila, P.S. and D.B. Skillicorn (2005), "Structure in the Enron Email Dataset," in *Proceeding of SIAM International Conference on Data Mining, SIAM Workshop on Link Analysis*, Counterterrorism and Security.

Klimt, B. and Y. Yang (2004), "The Enron Corpus: A New Dataset for Email Classification Research," To be published in *Proceedings of the European Conference on Machine Learning (ECML)*.

Loch, C.H., J.R. Tyler, and R. Lukose (submitted), "Conversational Structure in Email and Face to Face Communication," Draft, submitted to *Organization Science*.

McCallum, A., A. Corrada-Emmanuel, and X. Wang (2005), "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks, with Application to Enron and Academic Email", in *Proceeding of SIAM International Conference on Data Mining, SIAM Workshop on Link Analysis, Counterterrorism and Security*.

Newman, M.E.J. (2003), "The Structure and Function of Complex Networks," In SIAM Review, June 2003.

Preston, N. and M. Krishnamoorthy (2004), "GraphDraw: A Graph Drawing System to study Social Networks," Unpublished Manuscript, Rensselaer Polytechnic Institute, Troy, NY.

Priebe, C.E., J.M. Conroy, D.J. Marchette, and Y. Park (2005), "Scan Statistics on Enron Graphs," in *Proceeding of SIAM International Conference on Data Mining, SIAM Workshop on Link Analysis*, Counterterrorism and Security.

Tyler, J.R., M.D. Wilkinson and B.A. Huberman (2003), "Email as Spectroscopy: Automated Discovery of Community Structure within Organizations," in *Proceeding of the International Conference on Communities and Technologies*, Netherlands, kluwer Academic Publishers.

Anurat Chapanond is currently a Ph.D. student in Computer Science, RPI. Anurat graduated B. Eng. degree in Computer Engineering from Chiangmai University (Thailand) in 1997, M. S. in Computer Science from Columbia University in 2002. His research interest is in web data mining analyses and algorithms.

M.S. Krishnamoorthy received the B.E. degree (with honors) from Madras University in 1969, the M. Tech degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, in 1971, and the Ph. D. degree in Computer Science, also from the Indian Institute of Technology, in 1976.

From 1976 to 1979, he was an Assistant Professor of Computer Science at the Indian Institute of Technology, Kanpur. From 1979 to 1985, he was an Assistant Professor of Computer Science at Rensselaer Polytechnic Institute, Troy, NY, and since, 1985, he has been an Associate Professor of Computer Science at Rensselaer. Dr. Krishnamoorthy's research interests are in the design and analysis of combinatorial and algebraic algorithms, visualization algorithms and programming environments.

Bulent Yener is an Associate Professor in the Department of Computer Science and Co-Director of Pervasive Computing and Networking Center at Rensselaer Polytechnic Institute in Troy, New York. He is also a member of Griffiss Institute of Information Assurance.

Dr. Yener received MS. and Ph.D. degrees in Computer Science, both from Columbia University, in 1987 and 1994, respectively. Before joining to RPI, he was a Member of Technical Staff at the Bell Laboratories in Murray Hill, New Jersey.

His current research interests include bioinformatics, medical informtatics, routing problems in wireless networks, security and information assurance, intelligence and security informatics. He has served on the Technical Program Committee of leading IEEE conferences and workshops. Currently He is an associate editor of ACM/Kluwer Winet journal and the IEEE Network Magazine. Dr. Yener is a Senior Member of the IEEE Computer Society.