Automated Social Hierarchy Detection through Email Network Analysis

Ryan Rowe
Columbia University
Department of Applied Mathematics
New York, NY 10027
rrr2107@columbia.edu

Germán Creamer
Columbia University
Center for Computational Learning Systems
New York, NY 10027
ggc14@columbia.edu

Shlomo Hershkop
Columbia University
Computer Science Department
New York, NY 10027
shlomo@cs.columbia.edu

Salvatore J Stolfo
Columbia University
Computer Science Department
New York, NY 10027
sal@cs.columbia.edu

ABSTRACT

This paper provides a novel algorithm for automatically extracting social hierarchy data from electronic communication behavior. The algorithm is based on data mining user behaviors to automatically analyze and catalog patterns of communications between entities in a email collection to extract social standing. The advantage to such automatic methods is that they extract relevancy between hierarchy levels and are dynamic over time.

We illustrate the algorithms over real world data using the Enron corporation's email archive. The results show great promise when compared to the corporations work chart and judicial proceeding analyzing the major players.

General Terms

Social Network, Enron, Behavior Profile, Link Mining, Data Mining

1. INTRODUCTION

There is a vast quantity of untapped information in any collection of electronic communication records. The recent bankruptcy scandals in publicly held US companies such as Enron and WorldCom, and the subsequent Sarbanes-Oxley Act have increased the need to analyze these vast stores of electronic information in order to define risk and identify any conflict of interest among the entities of a corporate household. Corporate household is 'a group of business units united or regarded united within the corporation, such as suppliers and customers whose relationships with the corporation must be captured, managed, and applied for various purposes' [23]. The problem can be broken into three distinct phases; entity identification, entity aggregation, and transparency of inter-entity relationships [22].

Identifying individual entities is straightforward process, but the relationships between entities, or corporate hierarchy is not a straightforward task. Corporate entity charts sometimes exist on paper, but they do not reflect the day to day reality of a large and dynamic corporation. Corporate insiders are aware of these private relationships, but can be hard to come by, especially after an investigation. This information can be automatically extracted by analyzing the email communication data from within a corporation.

Link mining is a set of techniques that uses different types of networks and their indicators to forecast or to model a linked domain. Link mining has been applied to many different areas [27] such as money laundering [17], telephone fraud detection [9], crime detection [30], and surveillance of the NASDAQ and other markets [17, 13]. Perlich and Huang [25] show that customer modeling is a special case of link mining or relational learning [26] which is based on probabilistic relational models such as those presented by [12, 33, 34]. A recent survey of the literature can be found in [11]. In general models classify each entity independently according to its attributes. Probabilistic relational models classify entities taking into account the joint probability among them. The application of link mining to corporate communication is of course limited by restrictions to disseminate internal corporate data. Thus testing algorithms against real world data is hard to come by. An exception to this situation is the publicly available Enron email dataset.

The Enron Corporation's email collection described in section 2, is a publicly available set of private corporate data released during the judicial proceedings against the Enron corporation. Several researchers have explored it mostly from a Natural Language Processing (NLP) perspective [20, 19, 24]. Social network analysis (SNA) examining structural features [6] has also been applied to extract properties of the Enron network and attempts to detect the key players around the time of Enron's crisis; [7] studied the patterns of

Joint 9th WEBKDD and 1st SNA-KDD Workshop '07 August 12, 2007, San Jose, California , USA

communication of Enron employees differentiated by their hierarchical level; [16] interestingly enough found that word use changed according to the functional position, while [5] conducted a thread analysis to find out employees' responsiveness. [29] used an entropy model to identify the most relevant people, [8] presents a method for identity resolution in the Enron email dataset, and [1] applied a cluster ranking algorithm based on the strength of the clusters to this dataset.

The work presented in this paper differs in two major ways. First, the relationship between any two users are calculated based on behavior patterns of each specific user not just links. This allows the algorithm to judge the strength of communication links between users based on their overall communication pattern. Second, we assume a corporate householding perspective and propose a methodology to solve the problem of transparency of inter-entity relationships in an automatic fashion. Our approach determines link mining metrics which can reproduce approximate social hierarchy within an organization or a corporate household, and rank its members. We use our metric to analyze email flows within an organization to extract social hierarchy. We analyze the behavior of the communication patterns without having to take into account the actual contents of the email messages.

By performing behavior analysis and determining the communication patterns we are able to automatically:

- Rank the major officers of an organization.
- Group similarly ranked and connected users in order to accurately reproduce the organizational structure in question.
- Understand relationship strengths between specific sets of users.

This work is a natural extension of previous work on the Email Mining Toolkit project (EMT) [31, 32]. New functionality has been introduced into the EMT system for the purposes of automatically extracting social hierarchy information from any email collection.

The rest of the paper is organized as follows: Section 2 describes the Enron email corpus, section 3 presents the methods used to rank the Enron's officers; section 4 presents the results; section 4 discusses the results, and section 5 presents the conclusions.

2. ENRON ANTECEDENTS AND DATA

The Enron email data set is a rich source of information showcasing the internal working of a real corporation over a period between 1998-2002. There seems to be multiple versions of the "official" Enron email data set in the literature [6, 28, 21, 4]. In the midst of Enron's legal troubles in 2002, the Federal Energy Regulatory Commission (FERC) made a dataset of 619,449 emails from 158 Enron employees available to the public removing all attachment data. Cohen first put up the raw email files for researchers in 2004, the format was mbox style with each message in its own text file [4]. Following this, a number of research groups around the country obtained and manipulated the dataset in a variety of ways in attempts to correct inconsistencies and integrity issues within the dataset. Like [6], the version of the dataset we use to conduct our own research was treated and provided by Shetty and Adibi from ISI [28]. The ISI treatment of the Enron corpus consisted of deleting extraneous, unneeded emails and fixing some anomalies in the collection data having to do with empty or illegal user email names and bounced emails messages. In addition duplicates and blank emails were removed.

It should be noted that [3] has found that there is indication that a significant number of emails were lost either in converting the Enron data set or through specific deletion of key emails. So although we are working with most of the emails, we will make the assumption that the algorithm is robust although some emails are not part of the analysis. In addition the FERC dataset only covers about 92% of Enron employess at the time.

3. SNA ALGORITHM

The social network analysis algorithm works as follows:

For each email user in the dataset analyze and calculate several statistics for each feature of each user. The individual features are normalized and used in a probabilistic framework with which users can be measured against one another for the purposes of ranking and grouping. It should be noted that the list of email users in the dataset represents a wide array of employee positions within the organization or across organizational departments.

Two sets of statistics are involved in making the decision about a given user's "importance." First, we collect information pertaining to the flow of information, both volumetric and temporal. Here we count the number of emails a user has sent and received in addition to calculating what we call the **average response time** for emails. This is, in essence, the time elapsed between a user sending an email and later receiving an email from that same user. An exchange of this nature is only considered a "response" if a received message succeeds a sent message within three business days. This restriction has been implemented to avoid inappropriately long response times caused by a user sending an email, never receiving a response, but then receiving an unrelated email from that same user after a long delay, say a week or two. These elapsed time calculations are then averaged across all "responses" received to make up the average response time.

Second, we gather information about the nature of the connections formed in the communication network. Here we rank the users by analyzing **cliques** (maximal complete subgraphs) and other graph theoretical qualities of an email network graph built from the dataset. Using all emails in the dataset, one can construct an undirected graph, where vertices represent accounts and edges represent communication between two accounts. We build such a graph in order to find all cliques, calculate degree and centrality measures and analyze the social structure of the network. When all the cliques in the graph have been found, we can determine which users are in more cliques, which users are in larger cliques, and which users are in more important cliques. We base it on the assumption that users associated with a larger set and frequency of cliques will then be ranked higher. Finally all of the calculated statistics are normalized and combined, each with an individual contribution to an overall social score with which the users are ultimately ranked.

3.1 Information Flows

First and foremost, we consider the volume of information exchanged, i.e. the number of emails sent and received, to

be at least a limited indicator of importance. It is fair to hypothesize that users who communicate more, should, on average, maintain more important placement in the social hierarchy of the organization. This statistic is computed by simply tallying the total number of emails sent and received by each user.

Furthermore, in order to rate the importance of user i using the amount of time user j takes to respond to emails from user i, we must first hypothesize that a faster response implies that user i is more important to user j. Additionally, when we iterate and average over all j, we will assume that the overall importance of user i will be reflected in this overall average of his or her importance to each of the other people in the organization. In other words, if people generally respond (relatively) quickly to a specific user, we can consider that user to be (relatively) important. To compute the average response time for each account x, we collect a list of all emails sent and received to and from accounts y_1 through y_n, organize and group the emails by account y_1 through y_n, and compute the amount of time elapsed between every email sent from account x to account y_j and the next email received by account x from account y_j. As previously mentioned, communication of this kind contributes to this value only if the next incoming email was received within three business days of the original outgoing email.

3.2 Communication Networks

The first step is to construct an undirected graph and find all cliques. To build this graph, an email threshold N is first decided on. Next, using all emails in the dataset, we create a vertex for each account. An undirected edge is then drawn between each pair of accounts which have exchanged at least N emails. We then employ a clique finding algorithm, Algorithm 457, first proposed by Bron and Kerbosch [2]. This recursively finds all maximal complete subgraphs (cliques).

a. *Number of cliques*: The number of cliques that the account is contained within.

b. *Raw clique score*: A score computed using the size of a given account's clique set. Bigger cliques are worth more than smaller ones, importance increases exponentially with size.

c. *Weighted clique score*: A score computed using the "importance" of the people in each clique. This preliminary "importance" is computed strictly from the number of emails and the average response time. Each account in a clique is given a weight proportional to its computed preliminary. The weighted clique score is then computed by adding each weighed user contribution within the clique. Here the 'importance' of the accounts in the clique raises the score of the clique.

More specifically, the raw clique score R is computed with the following formula:

$$R = 2^{n-1}$$

where n is the number of users in the clique. The weighted clique score W is computed with the following formula:

$$W = t \cdot 2^{n-1}$$

where t is the time score for the given user.

Finally, the following indicators are calculated for the graph $G(V, E)$ where $V = v_1, v_2, ..., v_n$ is the set of vertices, E is the set of edges, and e_{ij} is the edge between vertices v_i and v_j:

- Degree centrality or degree of a vertex v_i: $deg(v_i) \doteq \sum_j a_{ij}$ where a_{ij} is an element of the adjacent matrix A of G
- Clustering coefficient: $C \doteq \frac{1}{n}\sum_{i=1}^{n} CC_i$, where $CC_i \doteq \frac{2|\{e_{ij}\}|}{deg(v_i)(deg(v_i)-1)} : v_j \in N_i,\ e_{ij} \in E$. Each vertex v_i has a neighborhood N defined by its immediately connected neighbors: $N_i = \{v_j\} : e_{ij} \in E$.
- Mean of shortest path length from a specific vertex to all vertices in the graph G: $L \doteq \frac{1}{n}\sum_j d_{ij}$, where $d_{ij} \in D$, D is the geodesic distance matrix (matrix of all shortest path between every pair of vertices) of G, and n is the number of vertices in G.
- Betweenness centrality $B_c(v_i) \doteq \sum_i \sum_j \frac{g_{kij}}{g_{kj}}$. This is the proportion of all geodesic distances of all other vertices that include vertex v_i where g_{kij} is the number of geodesic paths between vertices k and j that include vertex i, and g_{kj} is the number of geodesic paths between k and j [10].
- "Hubs-and-authorities" importance: "hub" refers to the vertex v_i that points to many authorities, and "authority" is a vertex v_j that points to many hubs. We used the recursive algorithm proposed by [18] that calculates the "hubs-and-authorities" importance of each vertex of a graph $G(V, E)$.

3.3 The Social Score

We introduce the social score S, a normalized, scaled number between 0 and 100 which is computed for each user as a weighted combination of the number of emails, response score, average response time, clique scores, and the degree and centrality measures introduced above. The breakdown of social scores is then used to:

i. Rank users from most important to least important

ii. Group users which have similar social scores and clique connectivity

iii. Determine n different levels (or echelons) of social hierarchy within which to place all the users. This is a clustering step, and n can be bounded.

The rankings, groups and echelons are used to reconstruct an organization chart as accurately as possible. To compute S , we must first scale and normalize each of the previous statistics which we have gathered. The contribution, C , of each metric is individually mapped to a [0, 100] scale and weighted with the following formula:

$$w_x \cdot C_x = w_x \cdot 100 \cdot \left[\frac{x_i - \inf x}{\sup x - \inf x}\right]$$

where x is the metric in question, w_x is the respective weight for that metric, the $\sup x$ and $\inf x$ are computed across all i users and x_i is the value for the i^{th} user. This normalization is applied to each of the following metrics:

1. number of emails
2. average response time
3. response score
4. number of cliques
5. raw clique score
6. weighted clique score
7. degree centrality
8. clustering coefficient
9. mean of shortest path length from a specific vertex to all vertices in the graph
10. betweenness centrality
11. "Hubs-and-Authorities" importance

Finally, these weighted contributions are then normalized over the chosen weights w_x to compute the social score as follows:

$$S = \frac{\sum_{\text{all } x} w_x \cdot C_x}{\sum_{\text{all } x} w_x}$$

This gives us a score between 0 and 100 with which to rank every user into an overall ranked list. Our assumption is that although the number of emails, average response time, number and quality of cliques, and the degree and centrality measures are all perfectly reasonable variables in an equation for "importance," the appropriate contribution, i.e. weight, of each will vary by situation and organization, and therefore can be adjusted to achieve more accurate results in a variety of cases.

3.4 Visualization

As part of this research, we developed a graphical interface for EMT, using the JUNG library, to visualize the results of social hierarchy detection by means of email flow.

After the results have been computed, the statistics calculated and the users ranked, the option to view the network is available. When this option is invoked, a hierarchical, organized version of the undirected clique graph is displayed. Nodes represent users, while edges are drawn if those two users have exchanged at least m emails. Information is provided to the user in two distinct ways, the qualities of a user are reflected in the look of each node, where the relative importance of a user is reflected in the placement of each node within the simulated organization chart.

Although every node is colored red, its relative size represents its social score. The largest node representing the highest ranked individual, the smallest representing the lowest. The transparency of a given node is a reflection of the user's time score. A user boasting a time score near to 1 will render itself almost completely opaque where a user with a very low time score will render almost entirely transparent.

The users are divided into one of n echelons using a grouping algorithm, we use $n = 5$ in this paper. Currently, the only grouping algorithm which has been implemented is a straight scale level division. Users with social scores from 80-100 are placed on the top level, users with social scores from 60-80 are placed on the next level down, etc. If the weights are chosen with this scale division in mind, only a small percentage of the users will maintain high enough social scores to inhabit the upper levels, so a tree-like organizational structure will be manifested. Different, more sophisticated, ranking and grouping algorithms have been considered and will be implemented, and will be discussed in the following section on future work.

When a node is selected with the mouse, all users connected to the selected user through cliques are highlighted and the user, time score and social score populate a small table at the bottom of the interface for inspection. Nodes can be individually picked or picked as groups and rearranged at the user's discretion. If the organization is not accurate or has misrepresented the structure of the actual social hierarchy in question, the user can return to the analysis window and adjust the weights in order to emphasize importance in the correct individuals and then can recreate the visualization.

If the user would prefer to analyze the network graphically with a non-hierarchical structure, a more traditional graph/network visualization is available by means of the Fruchterman-Reingold node placement algorithm. This node placement algorithm will emphasize the clique structure and the connectedness of nodes in the graph rather than the hierarchical ranking scheme in the first visual layout.

4. RESULTS AND DISCUSSION

We have performed the data processing and analysis using EMT [32]. EMT is a Java based email analysis engine built on a database back-end. The Java Universal Network/Graph Framework (JUNG) library [15] is used extensively in EMT for the degree and centrality measures, and for visualization purposes (see section 3.4).

In order to showcase the accuracy of our algorithm we present the analysis of the North American West Power Traders division of Enron Corporation.

As one can see in Table 1 and Figure 1, when running the code on the 54 users contained with the North American West Power Traders division we can reproduce the very top of the hierarchy with great accuracy. The transparency of the vertices in the graph visualization (Figure 1) denotes the response score of the user, a combination of the number of responses and the average response time. By our assumptions made in section three, we have determined that lower average response times infer higher importance, and appropriately, Tim Belden and Debra Davidson have fast average response times, causing more opaque colored node representations.

Once we turn to the lower ranked individuals, differences in our computed hierarchy and the official hierarchy are quite noticeable in Figure 3. As we move down the corporate ladder, the conversational flows of dissimilar employees can in fact be quite similar. Despite the discrepancies of our selections with the lower ranked officers, we find that consistently we are able to pick out the most important 2 or 3 individuals in any given subset, affording us the power to build a hierarchy from small groups up. Not only does the head of Enrons Western trading operation, Tim Belden, appear on the top of our list, both his administrative assistants appear with him. Additionally, in the first fourteen positions we are also able to identify the majority of directors, and an important number of managers and specialists. Figure 3 highlights these positions and their key role in the

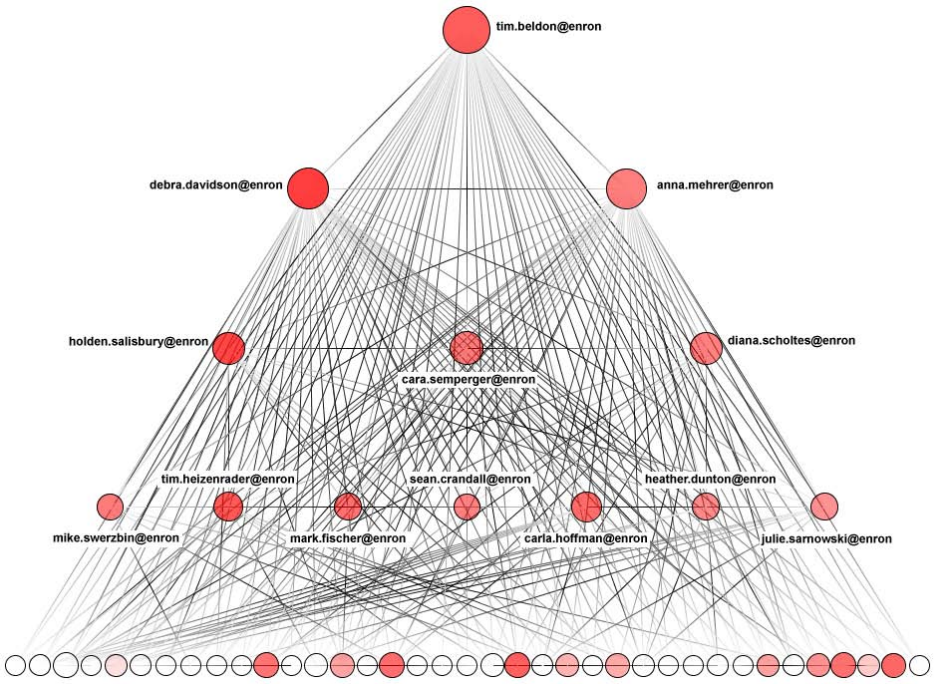

Figure 1: Enron North American West Power Traders Extracted Social Network

organizational structure.[1]

The placement of accounts other than the top two or three is in fact giving us insight into the true social hierarchy of this particular Enron business unit over the course of time from which the emails were gathered. This differs noticeably from the official corporate hierarchy, which can be expected as the data reflects the reality of the corporate communication structure.

With this sort of technique, it may be possible to view a snapshot of a corporate community (or any number of sub-communities) and effectively determine the real relationships and connections between individuals, a set of insights an official corporate organization chart simply could not offer.

5. CONCLUSIONS AND FUTURE WORK

Although real world data is hard to come by, the Enron dataset provides an excellent starting point for these tools. When we analyzed the algorithm on our own email data the social hierarchy of our lab was very apparent. Figure 2 clearly shows professor, PhD, lab students, and outsiders.

The next immediate concern is to apply these tools to the Enron dataset in a comprehensive and formal manner over time based data sets. The dataset contains enough email volume and generality to provide us with very useful results if we are interested in knowing how social structure changes over time. By varying the feature weights it is possible to use the mentioned parameters to:

a. Pick out the most important individual(s) in an organization,

b. Group individuals with similar social/email qualities, and

c. Graphically draw an organization chart which approximately simulates the real social hierarchy in question

In order to more completely answer our question, as previously mentioned, a number of additions and alterations to the current algorithms exist and can be tested. First, the concept of average response time can be reworked or augmented by considering the order of responses, rather than the time between responses, like in [14]. For example, if user a receives an email from user b before receiving an email from user c, but then promptly responds to user c before responding to user b, it should be clear that user c carries more importance (at least in the eyes of user a). Either replacing the average response time statistic with this, or introducing it as its own metric may prove quite useful.

Another approach is to consider common email usage times for each user and to adjust the received time of email to the beginning of the next common email usage time. For example, if user a typically only accesses her email from 9-11am and from 2-5pm, then an email received by user a at 7pm can be assumed to have been received at 9am the next morning. We hypothesize that this might correct errors currently introduced in the average response time calculations due to different people maintaining different work schedules.

In addition to the continued work on the average response time algorithms, new grouping and division algorithms are being considered. Rather than implementing the straight scale division algorithm, a more statistically sophisticated formula can be used to group users by percentile or standard deviations of common distributions. Furthermore, rather than ignoring the clique connections between users at this step, the graph edges could very well prove important in how

[1]Researchers interested in this line of research can find organigrams of public companies in their annual reports.

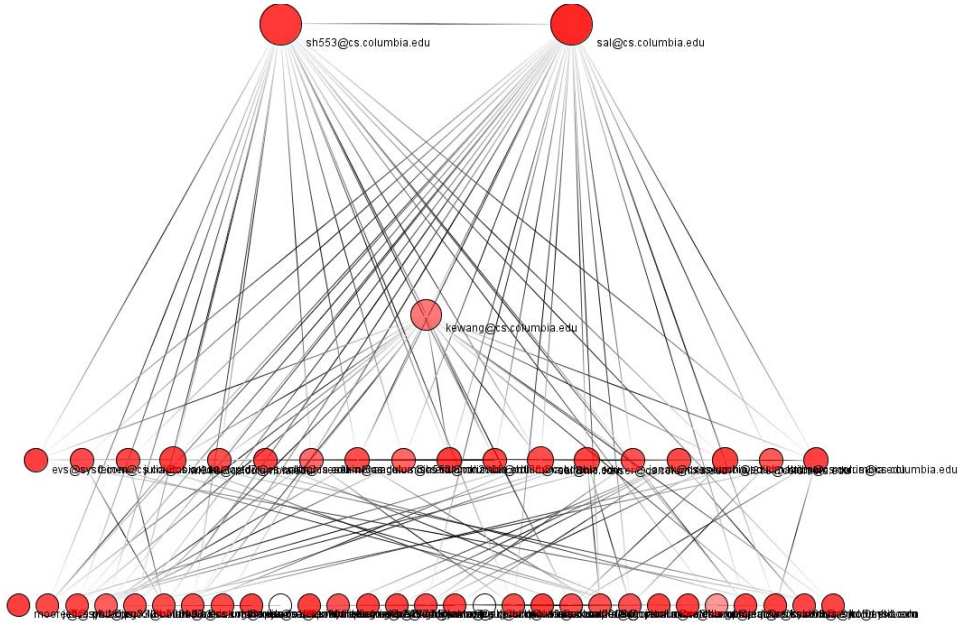

Figure 2: Analysis of our own emails

to arrange users into five different levels of social ranking, by grouping users with respect to their connections to others.

6. ADDITIONAL AUTHORS

7. REFERENCES

[1] Z. Bar-Yossef, I. Guy, R. Lempel, Y. S. Maarek, and V. Soroka. Cluster ranking with an application to mining mailbox networks. In *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, pages 63–74, Washington, DC, USA, 2006. IEEE Computer Society.

[2] C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, 1973.

[3] G. Carenini, R. T. Ng, and X. Zhou. Scalable discovery of hidden emails from large folders. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 544–549, New York, NY, USA, 2005. ACM Press.

[4] W. Cohen. Enron data set, March 2004.

[5] D. G. Deepak P and V. Varshney. Analysis of enron email threads and quantification of employee responsiveness. In *Proceedings of the Text Mining and Link Analysis Workshop on International Joint Conference on Artificial Intelligence*, Hyderabad, India, 2007.

[6] J. Diesner and K. Carley. Exploration of communication networks from the enron email corpus. In *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, Newport Beach CA, 2005.

[7] J. Diesner, T. L. Frantz, and K. M. Carley. Communication networks from the enron email corpus. *Journal of Computational and Mathematical Organization Theory*, 11:201–228, 2005.

[8] T. Elsayed and D. W. Oard. Modeling identity in archival collections of email: a preliminary study. In *Third Conference on Email and Anti-spam (CEAS)*, Mountain View, CA, July 2006.

[9] T. Fawcett and F. Provost. Activity monitoring: noticing interesting changes in behavior. In *Proceedings of the Fifth ACM SIGKDD International conference on knowledge discovery and data mining (KDD-99)*, pages 53–62, 1999.

[10] L. Freeman. Centrality in networks: I. conceptual clarification. *Social networks*, 1:215–239, 1979.

[11] L. Getoor and C. P. Diehl. Link mining: A survey. *SIGKDD Explorations*, 7(2):3–12, 2005.

[12] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. *Journal of Machine Learning Research*, 3:679–707, 2002.

[13] H. G. Goldberg, J. D. Kirkland, D. Lee, P. Shyr, and D. Thakker. The NASD securities observation, news analysis and regulation system (sonar). In *IAAI 2003*, 2003.

[14] S. Hershkop. *Behavior-based Email Analysis with Application to Spam Detection*. PhD thesis, Columbia University, 2006.

[15] D. F. Joshua O'Madadhain and S. White. Java universal network/graph framework, 2006. JUNG 1.7.4.

[16] P. Keila and D. Sillicorn. Structure in the enron email dataset. *Journal of Computational and Mathematical Organization Theory*, 11:183–199, 2005.

[17] J. D. Kirkland, T. E. Senator, J. J. Hayden, T. Dybala, H. G.Goldberg, and P. Shyr. The nasd regulation advanced detection system (ads). *AI Magazine*, 20(1):55–67, 1999.

[18] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 1999.

[19] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, Pisa, Italy, 2004.

[20] B. Klimt and Y. Yang. Introducing the enron corpus. In *First Conference on Email and Anti-spam (CEAS)*, Mountain View, CA, 2004.

[21] B. Klimt and Y. Yang. Introducing the enron corpus. In *CEAS*, 2004.

[22] S. Madnick, R. Wang, and X. Xian. The design and implementation of a corporate householding knowledge processor to improve data quality. *Journal of Management Information Systems*, 20(3):41–69, Winter 2003.

[23] S. Madnick, R. Wang, and W. Zhang. A framework for corporate householding. In C. Fisher and B. Davidson, editors, *Proceedings of the Seventh International Conference on Information Quality*, pages 36–40, Cambridge, MA, November 2002.

[24] A. McCallum, A. Corrada-Emmanuel, and X. Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. In *NIPS'04 Workshop on 'Structured Data and Representations in Probabilistic Models for Categorization'*, Whistler, B.C., 2004.

[25] C. Perlich and Z. Huang. Relational learning for customer relationship management. In *Proceedings of International Workshop on Customer Relationship Management: Data Mining Meets Marketing*, 2005.

[26] C. Perlich and F. Provost. Acora: Distribution-based aggregation for relational learning from identifier attributes. *Journal of Machine Learning*, 2005.

[27] T. E. Senator. Link mining applications: Progress and challenges. *SIGKDD Explorations*, 7(2):76–83, 2005.

[28] J. Shetty and J. Adibi. The enron email dataset database schema and brief statistical report, 2004.

[29] J. Shetty and J. Adibi. Discovering important nodes through graph entropy: the case of enron email database. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Chicago, Ill, August 2005.

[30] M. Sparrow. The application of network analysis to criminal intelligence: an assessment of the prospects. *Social networks*, 13:251–274, 1991.

[31] S. Stolfo, G. Creamer, and S. Hershkop. A temporal based forensic discovery of electronic communication. In *Proceedings of the National Conference on Digital Government Research*, San Diego, California, 2006.

[32] S. J. Stolfo, S. Hershkop, C.-W. Hu, W.-J. Li, O. Nimeskern, and K. Wang. Behavior-based modeling and its application to email analysis. *ACM Transactions on Internet Technology*, 6(2):187–221, May 2006.

[33] B. Taskar, E. Segal, and D. Koller. Probabilistic classification and clustering in relational data. In B. Nebel, editor, *Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence*, pages 870–878, Seattle, US, 2001.

[34] B. Taskar, M. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Proceedings of Neural Information Processing Systems, 2004.*, 2004.

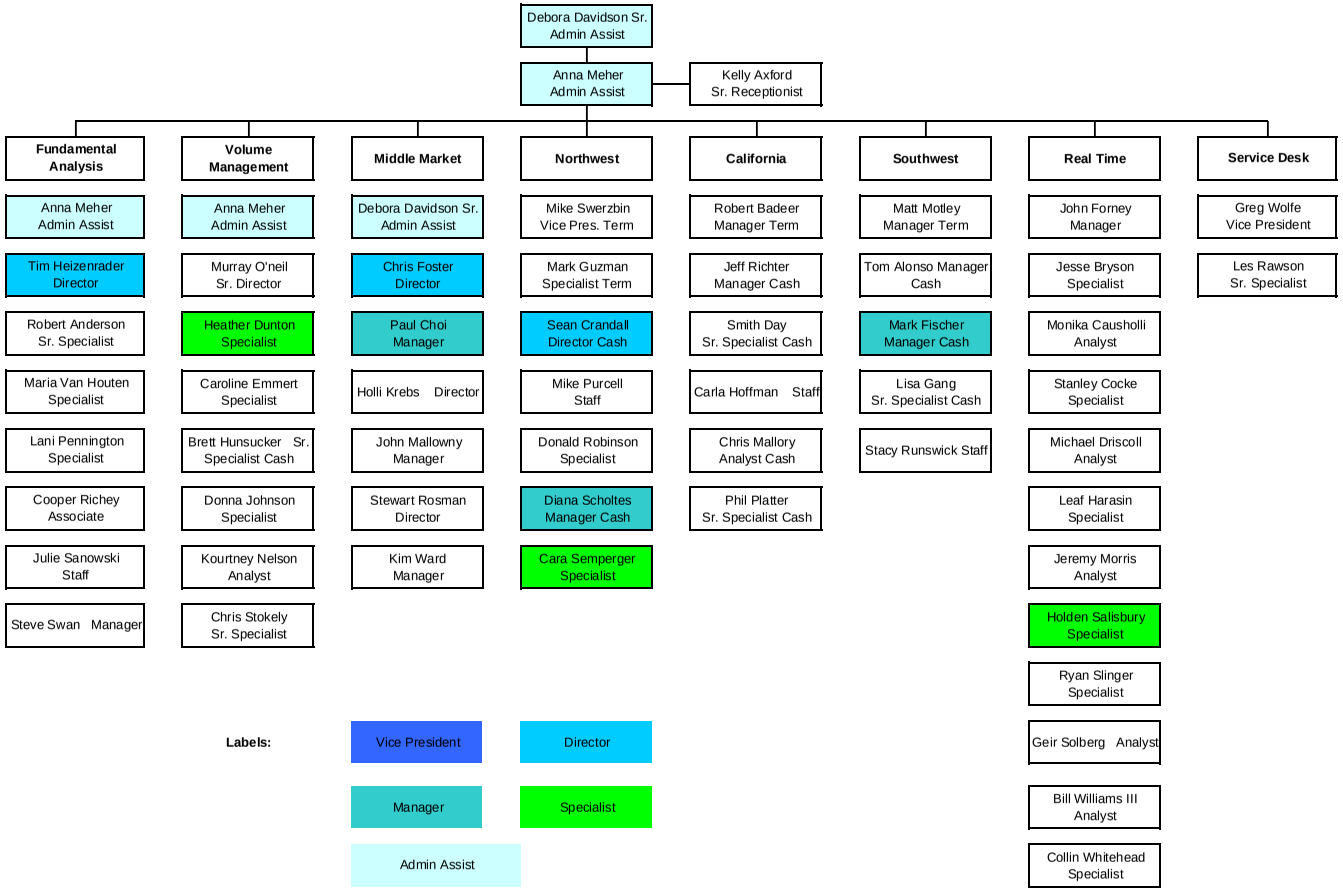

Figure 3: Network Chart with highlighted results

Name	Position	# Email	Avg Time	ResponseScore	# Cliques	RCS	WCS	Degree	Betweenness	Hubs	AvgDistance	ClusteringCoeff.	SocialScore
Tim Beldon	Vice President	1266	2493	0.641	236	251140	1261588	83.00	370.35	0.04	1.00	0.40	75.68
Debora Davidson Sr.	Admin Assist	537	17	0.757	235	251136	1261586	66.00	278.35	0.04	1.02	0.41	63.51
Anna Meher	Admin Assist	544	1833	0.506	231	250368	1259149	62.00	260.94	0.04	1.04	0.42	62.84
Carla Hoffman	Staff	739	1319	0.576	221	249232	1255447	55.00	143.98	0.04	1.13	0.49	61.67
Cara Semperger	Specialist	693	2707	0.506	137	167232	859288	63.00	82.96	0.03	1.25	0.52	53.68
Diana Scholtes	Manager Cash	468	2443	0.496	124	203520	1061153	45.00	21.44	0.03	1.43	0.70	53.31
Sean Crandall	Director Cash	412	2151	0.478	91	126912	657157	42.00	40.04	0.03	1.42	0.62	43.64
Holden Salisbury	Specialist	400	951	0.723	83	104192	532137	49.00	37.29	0.03	1.40	0.61	43.03
Mark Fischer	Manager Cash	346	1580	0.553	75	125952	676349	34.00	15.56	0.02	1.49	0.72	42.90
Heather Dunton	Specialist	329	2530	0.442	60	88736	462950	43.00	51.56	0.03	1.40	0.59	39.51
Bill Williams III	Analyst	257	3254	0.326	49	81408	437255	36.00	25.12	0.03	1.47	0.68	37.98
Paul Choi	Manager	157	N/A	0	91	130112	624944	44.00	48.03	0.03	1.38	0.60	36.02
Tim Heizenrader	Director	268	843	0.645	50	56960	298395	33.00	19.45	0.02	1.55	0.71	35.56
Chris Foster	Director	210	1612	0.56	46	58624	283552	35.00	23.18	0.02	1.49	0.66	34.74
Donald Robinson	Specialist	214	1486	0.545	23	34688	203384	27.00	6.67	0.02	1.62	0.81	33.03
Jeff Richter	Manager Cash	208	4393	0.12	34	43456	200427	25.00	12.80	0.02	1.57	0.74	32.53
Mike Swerzbin	Vice Pres. Term	269	1752	0.517	23	36672	195602	31.00	14.80	0.02	1.57	0.70	32.51
Stewart Rosman	Director	118	1386	0.567	20	40448	206036	26.00	6.85	0.02	1.62	0.81	32.25
Julie Sarnowski	Staff	284	2289	0.428	43	43008	220023	28.00	25.94	0.02	1.53	0.63	32.14
Stacy Runswick	Staff	188	2837	0.356	25	24064	134823	32.00	11.12	0.02	1.58	0.74	31.83
Mike Purcell	Staff	139	1338	0.626	11	15360	91653	24.00	5.02	0.02	1.66	0.79	30.36
Chris Mallory	Analyst Cash	180	N/A	0	56	78720	383567	27.00	9.92	0.02	1.55	0.76	30.19
Tom Alonso	Manager Cash	302	N/A	0	42	67584	362249	26.00	9.89	0.02	1.55	0.75	29.67
Greg Wolfe	Vice President	116	N/A	0	59	81920	388975	35.00	25.82	0.02	1.47	0.65	29.23
Matt Motley	Manager Term	223	N/A	0	26	56320	292362	23.00	3.04	0.02	1.62	0.88	28.93
Kim Ward	Manager	147	3901	0.206	4	768	2437	13.00	0.39	0.01	1.81	0.95	28.92
Jesse Bryson	Specialist	71	2346	0.428	17	6720	29988	23.00	7.42	0.02	1.66	0.77	28.10
Phil Platter	Sr. Specialist Cash	205	N/A	0	54	66528	315399	33.00	34.34	0.02	1.49	0.63	27.90
John Forney	Manager	63	5194	0.007	33	13504	47359	29.00	24.06	0.02	1.53	0.61	27.69
Geir Solberg	Analyst	127	3157	0.299	19	5760	23945	23.00	7.59	0.02	1.66	0.73	27.67
Stanley Cocke	Specialist	79	2689	0.367	21	14976	62360	26.00	18.15	0.02	1.57	0.64	27.40
Ryan Slinger	Specialist	111	1151	0.597	9	1344	5467	18.00	3.79	0.01	1.74	0.78	27.10
John Mallowny	Manager	140	N/A	0	16	41728	224918	31.00	6.50	0.02	1.60	0.81	26.74
Kourtney Nelson	Analyst	167	N/A	0	41	36032	176304	29.00	21.81	0.02	1.53	0.63	23.97
Lisa Gang	Sr. Specialist Cash	120	N/A	0	12	13056	65253	22.00	7.37	0.02	1.64	0.75	21.34
Monika Causholli	Analyst	44	N/A	0	12	3072	10871	16.00	2.21	0.01	1.74	0.86	20.58
Kelly Axford	Sr. Receptionist	76	N/A	0	4	2560	13698	15.00	1.68	0.01	1.75	0.87	20.51
Holli Krebs	Director	39	N/A	0	2	256	966	9.00	0.08	0.01	1.85	0.96	20.33
Les Rawson	Sr. Specialist	79	N/A	0	16	6656	26614	23.00	7.65	0.02	1.66	0.74	20.19
Jeremy Morris	Analyst	66	N/A	0	6	1024	3597	12.00	0.87	0.01	1.79	0.89	20.09
Robert Anderson	Sr. Specialist	44	N/A	0	2	256	958	8.00	0.15	0.01	1.85	0.96	20.06
Smith Day	Sr. Specialist Cash	14	N/A	0	1	32	75	6.00	0.00	0.01	1.91	1.00	20.00
Mark Guzman	Specialist Term	159	N/A	0	14	5248	20018	18.00	6.84	0.01	1.68	0.75	19.97
Caroline Emmert	Specialist	45	N/A	0	3	1024	4138	12.00	0.84	0.01	1.79	0.91	19.90
Steve Swan	Manager	28	N/A	0	2	192	622	9.00	0.20	0.01	1.85	0.93	19.55
Maria Van Houten	Specialist	20	N/A	0	2	128	411	7.00	0.11	0.01	1.87	0.95	19.44
Cooper Richey	Associate	36	N/A	0	7	1536	5001	14.00	2.68	0.01	1.75	0.82	18.89

Table 1: The raw data for the Enron North American subsidiary.