

Segmentation and Automated Social Hierarchy Detection through Email Network Analysis ^{*}

Germán Creamer¹³ Ryan Rowe² Shlomo Hershkop³ and Salvatore J. Stolfo³

¹ Center for Computational Learning Systems, Columbia University, New York, NY 10027

² Department of Applied Mathematics, Columbia University, New York, NY 10027

³ Department of Computer Science, Columbia University, New York, NY 10027
{ggc14, rrr2107}@columbia.edu, {shlomo, sal}@cs.columbia.edu

Abstract. We present our work on automatically extracting social hierarchies from electronic communication data. Data mining based on user behavior can be leveraged to analyze and catalog patterns of communications between entities to rank relationships. The advantage is that the analysis can be done in an automatic fashion and can adopt itself to organizational changes 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, Corporate Householding.

1 Introduction

There is a vast quantity of untapped information in any collection of electronic communication records. Current techniques of manual sifting and hard coded keyword searches do not scale to the task of analyzing these collections. 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].

^{*} This work is based on an earlier work: Automated Social Hierarchy Detection through Email Network Analysis in Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis ACM, 2007. <http://doi.acm.org/10.1145/1348549.1348562>

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 [28] such as money laundering [17], telephone fraud detection [9], crime detection [31], and surveillance of the NASDAQ and other markets [17, 13]. Perlich and Huang [26] show that customer modeling is a special case of link mining or relational learning [27] which is based on probabilistic relational models such as those presented by [12, 34, 35]. 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 [19, 21, 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 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. [30] 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 segments of users.

This work is a natural extension of previous work on the Email Mining Toolkit project (EMT) [32, 33]. 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 describes the research design; section 5 presents the results; section 5 discusses the results, and section 6 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, 29, 20, 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 [29]. Our final main dataset has 149 users after cleaning. We call this dataset as the ENRON dataset. 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. We also used a supplementary file provided by [29] to assign the position of each user. When we apply the occupational classification suggested by the former authors to our dataset, we find that 38.5% of the users are classified as “employee” or “N/A”. The classification “employee” does not bring any additional information more than indicating that the user is formally working at Enron. We reviewed the emails of those employees that were not well classified and imputed a position based on their signatures, the content of the email or lists of traders that circulated internally. We found out that

an important part of the “unknown” employees were traders or were acting as traders.

We also used another segment of the major FERC dataset that includes only the emails among the 54 workers that we identified as members of the North American West Power Traders division. We called this dataset as TRADER. The importance of this dataset is that [25] presents an organigram of the above division.

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 employees 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 $V = v_1, v_2, \dots, v_n$ 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 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 Research Design

We ranked the employees of both datasets ENRON and TRADERS using the social score (see Figure 5 and 6). We separated the ENRON dataset in four equal-sized segments where the top and low segments have the employees with the highest and lowest social scores respectively. We also classified the workers into four occupational categories:

1. Senior managers: CEO, chief risk officer, chief operating officer, presidents, vice presidents, and managing directors.
2. Middle managers: directors, managers, senior managers, lawyers, senior specialists, legal specialists, assistants to president, and risk management head. Assistants to president may qualify as regular “employees”, however they communicate and take similar decisions to those that a middle manager may take.
3. Traders. Some traders might be more important than a middle manager according to their performance, however we keep them in a separate category because of Enron’s leadership as an energy trading company.
4. Employees: employees, employee associates, analysts, assistant traders, and administrative assistants.

We expect that there is a relationship between the occupational category and the segment that each employee belongs to. For instance, senior managers should be mostly in the first segment, and middle managers in the first and second segments. An exception is the last category because 23 workers still keep the generic title “employees.” So they could be distributed among all the segments.

We built a 4 x 4 contingency table with the four segments and the four occupational categories (see Table 1). We wanted to test the hypothesis, using the Chi Square statistics, that there is a relationship between the occupational categories and the four segments of employees ranked by their social scores. So, we compared the ENRON contingency table with a contingency table that homogeneously distributes the same number of workers among the four segments (see Table 2). The null hypothesis is that the ENRON contingency table is not different from the expected contingency table.

In the case of the TRADERS dataset, the above analysis was not appropriate because it has fewer users and a flatter structure than the rest of the organization. We evaluated if the social score is capable of identifying the most important employees in the organizational structure or those that are in the top of the departmental organigram.

5 Results and Discussion

We have performed the data processing and analysis using EMT [33]. 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

Table 1. Actual distribution of Enron’s employees in occupational categories and segments defined by the social score

	Sr.Mgrs	Mgrs	Traders	Employees	Total
1	20	11	0	6	37
2	11	8	10	8	37
3	5	14	11	7	37
4	3	12	15	8	38
Total	39	45	36	29	149

Table 2. Expected distribution of Enron’s employees in occupational categories and segments defined by the social score

	Sr.Mgrs	Mgrs	Traders	Employees	Total
1	10	11	9	7	37
2	9	11	9	8	37
3	10	11	9	7	37
4	10	12	9	7	38
Total	39	45	36	29	149

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 separate analysis of the complete Enron dataset and the North American West Power Traders division of Enron.

5.1 Analysis of complete ENRON dataset

In the case of the ENRON dataset, the Chi Square test rejects the null hypothesis with a probability of 99.6%. Hence, the four segments defined by the social score has also aggregated Enron’s employees in a different way than a simple homogeneous distribution. In order to evaluate if the aggregation given by the social score also corresponds to the organizational hierarchy, we ranked the occupational groups in a scale of one to four based on a weighted average of the distribution of each occupational group in the four segments where one represents the highest hierarchy (see Table 3).

Table 3. Weighted ranking of each occupational category. The ranking is based on the distribution of each group of employees in the four hierarchical segments

Occupational category	Weighted ranking
Senior managers	1.77
Middle managers	2.6
Traders	3.14
Employees	2.59

Table 3 shows a direct relationship of the ranking and the hierarchy of each occupational category, at exception of the generic category “employees” which has a ranking similar to the one of the middle managers. We suppose that this category may hide workers from other categories that were not well classified. Senior managers are present in the first (51.3%) and second (28.2%) segments of the ENRON contingency table (see Table 4), so their ranking is 1.77.

Table 4. Proportional distribution of Enron’s employees in occupational categories and segments defined by the social score

	Sr. Mgrs	Mgrs	Traders	Employees
1	51.28%	24.44%	0.00%	20.69%
2	28.21%	17.78%	27.78%	27.59%
3	12.82%	31.11%	30.56%	24.14%
4	7.69%	26.67%	41.67%	27.59%

Middle managers have a ranking of 2.6. There is clearly a major jump with senior managers and their hierarchical level is higher than the one of the traders. The preeminence of Enron as an energy trading company leads to a slight distinction between the hierarchy of managers and traders. Even though managers organized the company, traders were the main drivers of the company. Therefore, the ranking of the traders is just slightly below the ranking of the managers.

Traders are mostly concentrated in the third and fourth segments (30.56% and 41.7% respectively) which is consistent with a ranking of 3.14. Most of the traders do not have a large number of emails. This can be explained because of the parallel communication systems of the traders (instantaneous message, phone, Bloomberg or trading terminal). They also communicate mostly among themselves, hence their social scores might be reduced in relation to the scores of the rest of the organization.

Employees are almost equally distributed in the last three segments and with smaller presence in the first segment. The even distribution of “employees” is easily explained by its generic category. According to the emails, many of them have a lot of influence in the company, however the emails studied did not indicate their occupational position. So we kept them in this generic category. When we eliminate this group of workers from our calculations, the probability of rejecting the null hypothesis using the Chi Square test is 99.95%.

5.2 Analysis of North American West Power Traders division

As one can see in Table 5 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.

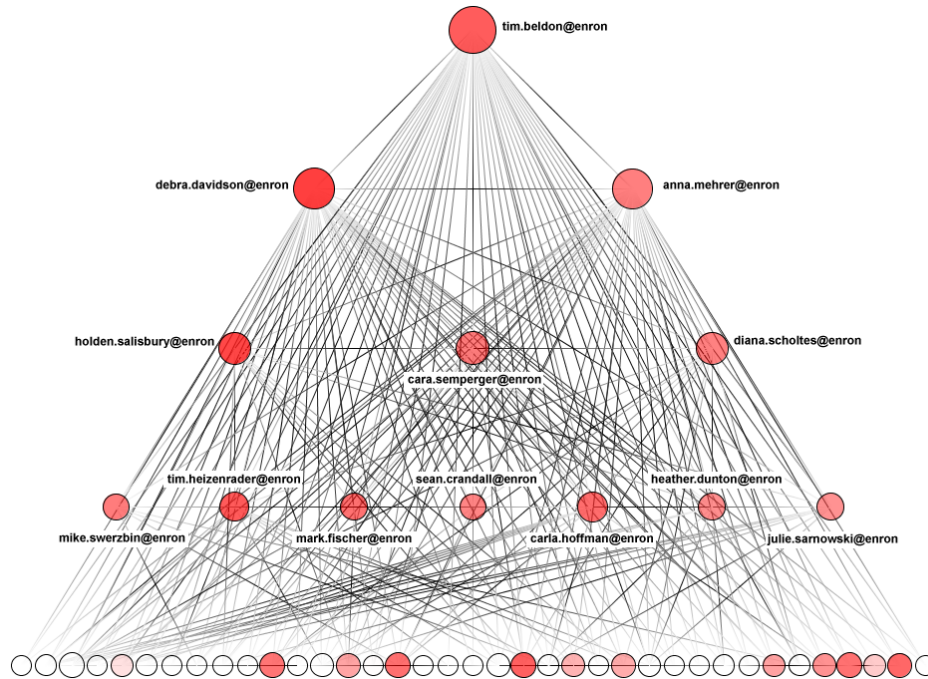


Fig. 1. Enron North American West Power Traders Extracted Social Network

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 segment, 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 organizational structure.⁴

⁴ Researchers interested in this line of research can find organigrams of public companies in their annual reports.

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 household or 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.

6 Conclusions and future work



Fig. 2. Analysis of our own emails

Understandingly, real world organizational data is hard to come by because of privacy concerns. The data in the the Enron dataset provides an excellent starting point for testing tools in a general setting. 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.

In our analysis of the Enron dataset, we have been able to recognize and rank the major officers, group them by their hierarchy, and capture the relationship among the segment of users. We think that this approach contributes to the

definition of corporate household in the case of Enron, and can be easily extended to other corporations.

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

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. Introducing the Enron corpus. In *First Conference on Email and Anti-spam (CEAS)*, Mountain View, CA, 2004.
20. B. Klimt and Y. Yang. Introducing the Enron corpus. In *CEAS*, 2004.
21. B. Klimt and Y. Yang. The Enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, Pisa, Italy, 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. R. McCullough. Memorandum related to reading Enron's scheme accounting materials, 2004. <http://www.mresearch.com/pdfs/89.pdf>.
26. 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.
27. C. Perlich and F. Provost. Acora: Distribution-based aggregation for relational learning from identifier attributes. *Journal of Machine Learning*, 2005.
28. T. E. Senator. Link mining applications: Progress and challenges. *SIGKDD Explorations*, 7(2):76–83, 2005.
29. J. Shetty and J. Adibi. The Enron email dataset database schema and brief statistical report, 2004.
30. 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.
31. M. Sparrow. The application of network analysis to criminal intelligence: an assessment of the prospects. *Social networks*, 13:251–274, 1991.
32. 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.
33. 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.
34. 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.
35. B. Taskar, M. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Proceedings of Neural Information Processing Systems, 2004.*, 2004.

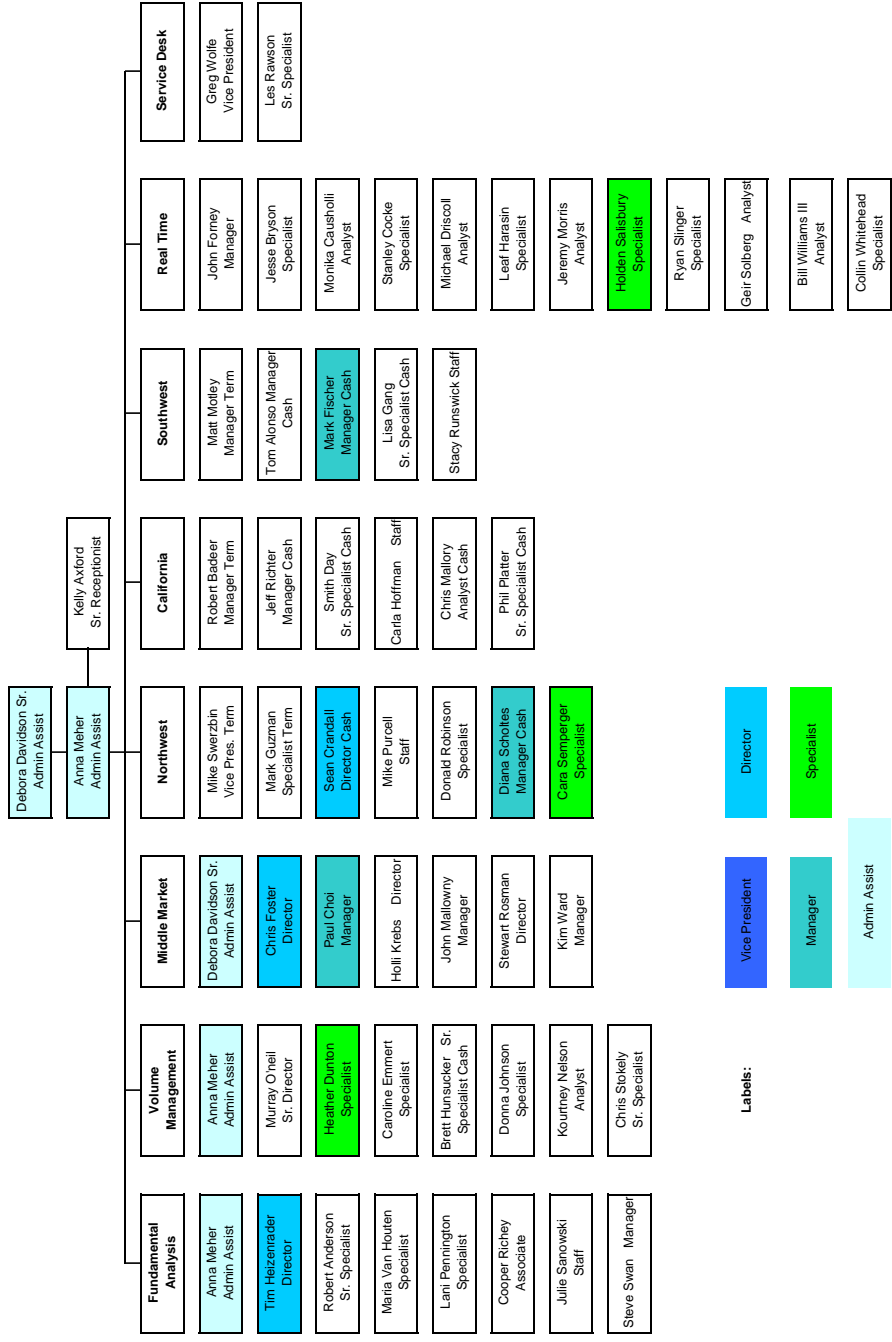


Fig. 3. Network Chart with highlighted results

Table 5. Social scores of employees. Enron North American subsidiary

<i>Name</i>	<i>Position</i>	<i># Email</i>	<i>Avg Time</i>	<i>Response</i>	<i># Cliques</i>	<i>RCS</i>	<i>WCS</i>	<i>Degree</i>	<i>Btw</i>	<i>Hubs</i>	<i>Avg.Dist.</i>	<i>CC</i>	<i>Score</i>
Tim Beldon	Vice President	1266	2493	0.641	236	251140	1261588	83.00	370.35	0.04	1.00	0.40	75.68
Debra 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

Note: CC: Clustering coefficient, Btw: Betweenness, Avg.Dist.: average distance, CC: clustering coefficient, Score: social score, Response: response score.

Table 6. Social scores of employees. Enron

<u>Name</u>	<u>Position</u>	<u># Email</u>	<u>Avg.Time</u>	<u>Response</u>	<u># Cliques</u>	<u>RCS</u>	<u>WCS</u>	<u>Degree</u>	<u>Btw</u>	<u>Hubs</u>	<u>Avg.Dist.</u>	<u>CC</u>	<u>Score</u>
Liz Taylor	Assistant to president	1724	1552	0.547	122	5790	24960	69	5087.36	0.03	1.57	0.12	64.76
Louise Kitchen	President, Enron Online	2480	1542	0.594	121	6452	27700	50	1459.04	0.03	1.76	0.21	60.07
Sally Beck	Chief operating officer	1107	2075	0.481	96	4978	22024	61	2899.41	0.03	1.70	0.14	54.63
Kenneth Lay	CEO, chairman	510	2986	0.347	98	5918	26441	52	1975.09	0.03	1.74	0.19	53.41
Michelle Lokay	Administrative assistant	1676	1583	0.568	16	6400	35576	16	85.37	0.00	2.52	0.75	47.31
Kimberly Watson	Employee	1803	1934	0.570	14	5890	33169	18	101.61	0.00	2.64	0.64	46.83
Lindy Donoho	Employee	997	1766	0.537	16	6400	35576	17	71.14	0.00	2.60	0.74	46.26
Steven Harris	Sr. manager *	1269	1515	0.565	16	6400	35576	18	104.80	0.00	2.50	0.69	46.20
Mike Grigsby	VP *	3463	1970	0.576	83	2096	6936	34	1006.34	0.02	1.89	0.28	44.11
Kevin Hyatt	Director, Pipeline business	623	1398	0.569	13	5248	29098	16	85.37	0.00	2.52	0.75	41.21
Lynn Blair	Manager *	975	1528	0.562	9	4480	25133	16	65.66	0.00	2.61	0.74	40.12
Drew Fossum	VP	949	1416	0.578	18	4892	27119	19	222.89	0.01	2.39	0.53	39.86
Teb Lokey	Regulatory affairs manager	324	1037	0.615	6	3200	18215	13	4.78	0.00	3.11	0.87	37.67
Jeff Dasovich	Director	4173	1737	0.765	22	856	3748	25	804.22	0.01	2.01	0.26	37.52
Darrell Schoolcraft	Manager *	371	2738	0.385	4	2816	16150	10	1.38	0.00	3.18	0.96	37.17
Shelley Corman	VP,regulatory affairs	922	1472	0.570	14	3624	19831	18	263.32	0.01	2.43	0.56	36.28
Scott Neal	VP -- trader	607	2353	0.437	53	2786	11015	31	616.46	0.02	1.97	0.29	35.89
Greg Whalley	President	613	611	0.670	32	3416	16964	21	122.24	0.02	2.02	0.47	35.63
Tana Jones	Senior legal specialist	3565	1374	0.683	25	644	2353	20	336.92	0.01	2.17	0.35	33.53
Rod Hayslett	VP, CFO and treasurer	853	1208	0.628	9	2816	15429	17	633.12	0.01	2.11	0.57	33.41
Andy Zipper	VP,Enron online	394	2389	0.430	41	2736	10828	16	49.78	0.01	2.15	0.56	33.26
Bill Rapp	Attorney	400	2104	0.478	5	2564	14670	12	101.66	0.00	2.70	0.79	33.26
Steven Kean	VP & chief of staff	740	1650	0.535	23	2744	14169	19	159.24	0.01	2.03	0.50	33.02
Sara Shackleton	VP	3343	1394	0.703	14	328	1192	12	34.45	0.00	2.56	0.59	32.89
Arnold John	VP	874	1233	0.597	41	2228	8488	20	165.76	0.01	2.07	0.43	32.76
Mark McConnell	Manager, TW *	692	1619	0.550	4	1536	8505	12	9.02	0.00	2.87	0.88	32.26
Rosalee Fleming	Employee	817	1840	0.506	11	2304	12269	17	41.10	0.01	2.05	0.62	31.88
Tracy Geaccone	Employee	614	771	0.676	5	1536	8481	12	11.48	0.00	2.82	0.83	31.83
Hunter Shively	VP -- trader	441	2013	0.483	33	2372	9822	20	212.59	0.01	2.09	0.41	31.70
Richard Shapiro	VP,regulatory affairs	2630	964	0.667	19	888	3853	18	182.02	0.01	2.14	0.45	31.46
Phillip Allen	Managing director	567	360	0.704	40	1960	7777	18	86.07	0.01	2.05	0.51	31.43
Paul Barbo	Sr. manager *	492	1973	0.489	4	2054	12093	13	192.58	0.00	2.68	0.68	31.04
Stanley Horton	President,Enron gas pipeline	410	1629	0.533	10	2576	14133	18	357.72	0.01	2.08	0.51	30.98
Jeffrey Skilling	Ceo	168	886	0.633	12	2536	13586	14	30.27	0.01	2.12	0.64	30.76
Rick Buy	Chief risk management	350	559	0.678	12	2072	11337	13	33.82	0.01	2.09	0.74	30.56
Mark Haedicke	Managing director,legal	414	1220	0.590	30	2000	10132	23	535.05	0.01	1.99	0.36	30.26
Stephanie Panus	Employee	1823	1999	0.551	30	608	2055	17	201.55	0.01	2.37	0.38	29.09
Matthew Lenhart	Trader	1112	1563	0.553	42	1060	3302	20	314.77	0.01	2.05	0.38	29.07
Marie Heard	Lawyer	1910	1735	0.574	19	552	2025	13	58.27	0.01	2.22	0.64	29.06
Susan Bailey	Legal specialist *	1623	2152	0.486	10	320	1157	10	7.11	0.00	2.67	0.76	29.03
Jeffrey Shankman	President,Enron global mkts	178	951	0.625	5	1680	9627	14	23.96	0.01	2.14	0.71	29.02
Kevin Presto	VP	659	1637	0.548	52	1016	2932	22	208.31	0.01	2.11	0.30	28.93
Kam Keiser	Employee	950	985	0.640	39	816	2428	24	352.67	0.01	2.09	0.30	28.56
Fletcher Sturm	VP	248	1364	0.568	26	1344	5640	17	53.18	0.01	2.11	0.52	28.27
Carol Clair	In house lawyer	1846	1194	0.683	11	192	627	9	36.69	0.00	2.53	0.64	28.20
Thomas Martin	VP	273	993	0.619	19	1736	7567	16	98.15	0.01	2.13	0.49	28.19
Susan Scott	Assistant trader *	1070	1891	0.507	28	780	2975	29	1277.64	0.01	2.05	0.14	27.85
Phillip Platter	Sr. specialist	51	222	0.721	2	32	89	4	0.00	0.00	3.09	1.00	26.82
Jay Reitmeyer	Trader *	579	2619	0.399	31	848	2607	15	175.84	0.01	2.14	0.48	26.68
Monique Sanchez	Trader *	663	1891	0.500	33	684	2020	24	307.16	0.01	2.08	0.27	26.52
James Steffes	VP,government affairs	1460	1161	0.666	17	384	1334	13	139.87	0.01	2.21	0.44	26.46
Barry Tycholiz	VP	1170	722	0.681	26	440	1317	16	146.66	0.01	2.07	0.43	26.32
Mike Maggi	Director	302	1031	0.612	9	648	2551	10	9.24	0.01	2.24	0.82	26.27
Phillip Love	Sr. manager -- traders *	588	1178	0.611	29	492	1335	23	504.59	0.01	2.19	0.26	26.12
Danny McCarty	VP	302	1138	0.600	4	1216	6720	14	201.46	0.01	2.18	0.58	25.87
Jane Tholt	VP	72	2127	0.463	2	12	24	3	0.00	0.00	2.86	1.00	25.83
Matt Smith	Trader *	390	2198	0.455	7	196	620	8	4.97	0.01	2.51	0.82	25.63
Elizabeth Sager	VP *	761	687	0.667	23	568	2073	21	389.14	0.01	2.09	0.34	25.61
Jason Williams	Trader *	792	1864	0.509	28	468	1418	17	220.80	0.01	2.10	0.39	25.24
Gerald Nemeec	Attorney *	992	1746	0.547	26	364	1061	20	644.56	0.01	2.09	0.26	25.06
Mark Taylor	Employee	679	1180	0.610	7	368	1503	11	46.54	0.01	2.28	0.67	24.94
Debra Perlingiere	Legal specialist *	820	1666	0.549	22	422	1347	20	335.68	0.01	2.30	0.29	24.91
Cara Semperger	Senior analyst cash	334	1878	0.503	10	196	520	10	28.84	0.00	2.87	0.60	24.74
Jason Wolfe	Trader *	392	889	0.631	17	256	742	10	14.23	0.01	2.17	0.69	24.55
Tori Kuykendall	Trader	468	474	0.689	10	226	705	10	19.66	0.00	2.63	0.58	24.52
Errol McLaughlin	Employee	885	1687	0.527	10	228	710	13	52.74	0.01	2.24	0.51	24.31
John Griffith	Managing director UK	413	1522	0.547	7	400	1454	11	22.96	0.01	2.28	0.64	24.27
Lysa Akin	Employee	382	1696	0.526	29	412	1036	20	591.63	0.00	2.41	0.26	24.20
Stacy Dickson	Employee	510	1017	0.618	5	40	99	6	6.23	0.00	2.91	0.60	24.08
Theresa Staab	Employee	187	2365	0.435	2	24	59	5	3.22	0.00	2.74	0.80	23.99
Richard Sanders	VP,Enron wholesale svcs.	256	1596	0.538	11	204	654	9	17.24	0.00	2.60	0.56	23.49
Kate Symes	Trader *	292	1360	0.571	12	220	567	14	85.52	0.00	2.82	0.40	23.48
Kim Ward	Trader *	803	1082	0.629	22	222	596	19	464.85	0.01	2.02	0.25	23.43
Chris Germany	Trader *	972	2515	0.419	15	210	517	16	272.50	0.00	2.53	0.21	23.41
Dana Davis	VP term	245	2775	0.375	23	452	1151	11	41.85	0.01	2.16	0.47	23.40

<u>Name</u>	<u>Position</u>	<u># Email</u>	<u>Avg.Time</u>	<u>Response</u>	<u># Cliques</u>	<u>RCS</u>	<u>WCS</u>	<u>Degree</u>	<u>Btw</u>	<u>Hubs</u>	<u>Avg.Dist.</u>	<u>CC</u>	<u>Score</u>
Larry May	Director	303	2362	0.433	8	428	1628	14	121.11	0.01	2.11	0.52	23.40
Mark Whitt	VP *	698	1399	0.589	9	176	539	13	109.44	0.01	2.24	0.46	23.34
Frank Ermis	Director	338	548	0.678	14	252	772	12	27.19	0.01	2.18	0.56	23.26
Sandra Brawner	Director	160	614	0.671	8	352	1332	9	13.67	0.01	2.24	0.64	23.23
Keith Holst	Director	496	2829	0.368	16	424	1292	11	75.99	0.01	2.16	0.45	23.18
Randall Gay	Manager *	409	1052	0.610	12	266	820	11	170.52	0.01	2.16	0.56	23.14
Darron Giron	Trader *	205	1307	0.580	2	8	13	3	0.38	0.00	2.85	0.67	23.13
Mary Hain	In house lawyer	817	1014	0.632	20	224	568	18	452.53	0.01	2.21	0.23	23.01
Don Baughman	Trader	277	2551	0.407	25	344	700	14	170.72	0.01	2.24	0.35	22.95
Jonathan Mckay	Director	207	640	0.666	20	504	1610	12	62.23	0.01	2.23	0.42	22.89
Michelle Cash	Legal specialist *	138	1891	0.497	8	224	760	9	17.51	0.00	2.48	0.58	22.85
Dan Hyvl	Employee	412	1495	0.562	7	92	279	9	57.62	0.00	2.59	0.53	22.83
Vince Kaminski	Risk management head	145	749	0.651	3	274	1380	7	34.56	0.01	2.35	0.67	22.71
Paul Lucci	Employee	254	1169	0.600	5	120	394	9	22.00	0.00	2.41	0.64	22.68
Doug Gilbert-smith	Manager	249	1500	0.551	14	192	470	12	44.13	0.01	2.22	0.47	22.62
Charles Weldon	Trader *	68	12	0.750	5	32	69	4	0.56	0.00	2.36	0.83	22.59
Harry Arora	VP	104	311	0.710	8	136	382	8	10.13	0.01	2.38	0.61	22.57
Bill Williams	Trader *	362	2271	0.446	10	176	473	14	266.01	0.00	2.72	0.33	22.54
Chris Dorland	Employee	202	669	0.662	28	372	990	16	115.63	0.01	2.11	0.28	22.12
Stacey White	Manager -- trader *	189	971	0.622	13	152	375	9	34.28	0.01	2.24	0.53	22.02
Vladi Pimenov	Trader *	85	1740	0.516	6	160	448	6	5.98	0.01	2.32	0.67	21.93
Eric Bass	Trader	632	962	0.641	15	172	470	14	204.18	0.01	2.09	0.32	21.93
Scott Hendrickson	Trader *	190	1535	0.544	3	24	37	4	6.28	0.00	2.52	0.67	21.84
Dutch Quigley	Trader *	440	1201	0.595	10	148	432	12	62.35	0.01	2.20	0.42	21.84
James Derrick	In house lawyer	236	1648	0.529	5	316	1406	10	75.22	0.01	2.14	0.53	21.83
Diana Scholtes	Trader	273	2185	0.458	10	176	495	11	250.41	0.00	2.30	0.51	21.78
Holden Salisbury	Cash analyst	132	1001	0.616	5	72	189	8	187.36	0.00	2.41	0.64	21.77
Kay Mann	Employee	374	1262	0.587	13	162	447	12	297.42	0.00	2.43	0.36	21.73
Martin Cuilla	Manager	215	1447	0.556	18	160	349	13	183.53	0.01	2.26	0.35	21.72
Geir Solberg	Trader *	111	1588	0.536	5	64	152	7	98.02	0.00	2.42	0.67	21.68
John Zufferli	VP *	186	835	0.639	12	112	247	8	18.01	0.01	2.26	0.54	21.57
John Hodge	Managing director	115	2081	0.470	3	36	76	5	18.37	0.00	2.61	0.60	21.54
Robin Rodrigue	Analyst *	50	1990	0.481	2	20	47	5	9.24	0.00	2.43	0.70	21.52
Judy Townsend	Trader *	323	1100	0.603	2	40	125	4	12.03	0.00	2.39	0.67	21.46
Sean Crandall	Director -- trader	214	1789	0.509	12	188	475	9	114.42	0.00	2.32	0.53	21.40
Mike Carson	Employee	112	597	0.671	9	112	282	6	7.63	0.01	2.22	0.60	21.20
Peter Keavey	Employee	54	11	0.750	2	18	46	4	1.12	0.00	2.78	0.50	21.15
Juan Hernandez	Senior specialist logistics	118	2266	0.445	5	40	59	6	8.68	0.00	2.39	0.60	21.12
Benjamin Rogers	Employee associate	82	304	0.711	4	40	82	5	5.95	0.00	2.45	0.60	21.05
Jim Schwieger	Trader	181	2022	0.478	14	348	1119	10	36.73	0.01	2.28	0.36	20.98
Joe Stepenovitch	VP,energy mkting & trading	112	5530	0.000	3	24	28	5	5.68	0.00	2.47	0.60	20.97
Eric Saibi	Trader	166	3086	0.332	5	52	113	6	11.82	0.00	2.26	0.60	20.81
Daren Farmer	Logistics manager	69	780	0.646	3	24	57	5	10.90	0.00	2.46	0.60	20.77
Robert Badeer	Director	404	500	0.684	6	38	84	7	59.38	0.00	2.49	0.43	20.63
Mike Swerzbin	Trader	156	1238	0.583	11	144	338	11	214.45	0.01	2.14	0.38	20.46
Ryan Slinger	Trader	122	1151	0.596	5	52	111	7	286.25	0.00	2.44	0.48	20.32
Eric Linder	Trader *	39	N/A	0.000	1	8	12	2	0.00	0.00	3.25	1.00	20.05
Kevin Ruscitti	Trader	74	968	0.621	3	16	29	4	13.60	0.00	2.89	0.33	19.93
John Forney	Manager,real time trading	211	3207	0.319	11	74	105	8	23.08	0.00	2.36	0.36	19.87
Lisa Gang	Director	73	N/A	0.000	1	16	35	4	0.00	0.00	3.12	1.00	19.77
Geoff Storey	Director -- trader	171	254	0.719	8	80	214	8	38.31	0.01	2.24	0.36	19.58
Richard Ring	Employee	41	973	0.620	2	10	22	3	5.51	0.00	2.85	0.33	19.45
Paul Thomas	Trader *	68	4685	0.115	6	60	92	8	76.00	0.00	2.32	0.39	19.30
Susan Pereira	Trader *	23	N/A	0.000	1	8	13	2	0.00	0.00	3.07	1.00	19.15
Joe Parks	Trader *	120	2490	0.414	3	28	62	5	28.88	0.00	2.42	0.40	18.87
Patrice Mims	Employee *	130	420	0.696	2	40	123	5	20.79	0.00	2.27	0.40	18.42
Steven South	Trader *	27	N/A	0.000	1	4	4	2	0.00	0.00	2.86	1.00	18.13
Chris Stokley	Employee *	68	3851	0.228	5	26	43	6	48.58	0.00	2.26	0.33	18.05
Mark Guzman	Trader	84	N/A	0.000	2	48	117	6	2.20	0.00	3.04	0.80	17.83
John Lavorato	CEO,Enron America	274	N/A	0.000	24	640	1846	16	71.23	0.01	2.24	0.28	16.08
Jeff King	Manager	94	N/A	0.000	11	96	146	5	1.04	0.00	2.34	0.80	15.63
David Delainey	CEO, Enron N.A. & E.energy	86	N/A	0.000	3	128	406	7	5.38	0.01	2.22	0.81	15.45
Tom Donohoe	Employee *	31	1287	0.577	2	6	7	2	4.08	0.00	2.61	0.00	15.11
Robert Benson	Director	100	N/A	0.000	7	112	218	6	3.08	0.01	2.29	0.73	14.95
Andrew Lewis	Director -- trader	74	N/A	0.000	3	32	60	5	1.83	0.01	2.32	0.80	14.82
Larry Campbell	Senior specialist	32	N/A	0.000	1	8	8	3	1.73	0.00	2.82	0.67	14.60
Albert Meyers	Trader *	37	814	0.641	2	4	4	2	13.73	0.00	2.50	0.00	14.54
Brad Mckay	Employee	120	N/A	0.000	10	176	379	9	35.19	0.01	2.34	0.56	14.40
Matt Motley	Director	90	N/A	0.000	6	84	179	6	35.49	0.00	2.58	0.53	13.50
Craig Dean	Trader	71	N/A	0.000	2	8	8	3	13.73	0.00	2.49	0.67	13.06
Andrea Ring	Trader *	118	N/A	0.000	3	22	37	6	88.75	0.00	2.32	0.47	11.56
Cooper Richey	Manager	91	N/A	0.000	5	36	55	7	46.05	0.00	2.26	0.43	11.24
Monika Causholli	Analyst risk management	30	N/A	0.000	2	6	4	2	7.91	0.00	2.57	0.00	6.51
Joe Quenet	Trader	37	N/A	0.000	3	10	6	2	0.77	0.00	2.50	0.00	6.31

Note: * indicates that the position is inferred by the content of emails. TW: Transwestern Pipeline Company, N.A.: North America, VP: vice president, E: Enron
CC: Clustering coefficient, Btw: Betweenness, Avg.Dist.: average distance, CC: clustering coefficient, Score: social score, Response: response score.