Exploration of Communication Networks from the Enron Email Corpus[1]

Jana Diesner (diesner@cs.cmu.edu)
Kathleen M. Carley (kathleen.carley@cmu.edu)
Carnegie Mellon University

Abstract

The Enron email corpus is appealing to researchers because it is a) a large scale email collection from b) a real organization c) over a period of 3.5 years. In this paper we contribute to the initial investigation of the Enron email dataset from a social network analytic perspective. We report on how we enhanced and refined the Enron corpus with respect to relational data and how we extracted communication networks from it. We apply various network analytic techniques in order to explore structural properties of the networks in Enron and to identify key players across time. Our initial results indicate that during the Enron crisis the network had been denser, more centralized and more connected than during normal times. Our data also suggests that during the crisis the communication among Enron's employees had been more diverse with respect to people's formal positions, and that top executives had formed a tight clique with mutual support and highly brokered interactions with the rest of organization. The insights gained with the analyses we perform and propose are of potential further benefit for modeling the development of crisis scenarios in organizations and the investigation of indicators of failure.

Key Words: Enron, social network analysis, dynamic social networks, communication networks, DyNetML, ORA

1 Introduction

The Enron email corpus is appealing to researchers because it is a) a large scale email collection from b) a real organization c) over a period of 3.5 years. For research related to Social Networks, Organizational Theory, and Organizational Behavior this dataset is of particular interest and potential value because it enables the long term examination of interactions and processes within and among the entities of an organization. The Enron corpus contains a large amount of information on interaction, communication, knowledge, cognition, resources, tasks and relationships on an individual and group level in Enron. In order to explore and understand how these factors might have impacted the network, its design, culture, and life cycle, we need to extract and analyze this information in an effective and efficient way.

There is a growing body of research on various aspects of the Enron email corpus. To date, most publications have focused on Natural Language Processing (NLP) of the data: Klimt and Yang [17][18] and Bekkerman [2] explored the classification of emails, such as the organization of messages in user-defined folders and thread detection. Corrada-Emmanuel used the MD5 digest to generate mappings of the dataset, such as mapping of authors and recipients [8]. Shetty and Adibi [33] provide information on quantitative features of the corpus, such as the distribution of the number of emails per user and over time (months, years). They generated a social network that represents 151 Enron employees. In this network each exchange of at least 5 emails between any pair of agents across the entire time range (1998 to 2002) was considered as a link.

Essentially, the research community is exploring the Enron dataset from a mainly NLP perspective. In this paper we contribute to this initial investigation from a network analytic perspective: We describe how we enhanced and refined the Enron email database with respect to relational data. Moreover, we report on how we extracted network data from our instance of the corpus and demonstrate the application of various social network analytic techniques to the exploration of structural and behavioral features of the organization under investigation. The network analytic perspective enables us to investigate vulnerabilities of the system and its adaptivity to changing situations. The insights gained with the analyses we perform and propose are of potential further benefit for modeling the development of crisis scenarios in organizations and the investigation of indicators of failure. Note that the work presented in this paper is research in progress; the results of our sample study cannot be generalized for the Enron corporation or other organizations, but show what knowledge we can gain from analyzing an email corpus from a network analytic perspective and what kind of questions we can answer.

[1] This paper is part of the Dynamics Networks project in CASOS (Center for Computational Analysis of Social and Organizational Systems, http://www.casos.cs.cmu.edu) at Carnegie Mellon University. This work was supported in part by the Office of Naval Research (ONR), United States Navy Grant No. 9620.1.1140071 on Dynamic Network Analysis under the direction of Rebecca Goolsby. Additional support on measures was provided by the DOD and the NSF. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation or the U.S. government. We thank Corinne Coen (SUNY, Buffalo) for her advice on this project, Eduard Hovy (USC, ISI) for pointing us to ISI's work on Enron, and the CASOS lab for their help on this work; especially Andrew Dougherty and Dan Woods.

Section 2 provides a synopsis of the Enron case and develops our research questions. Section 3 describes the dataset. In section 4 we report on how we refined the database and extracted relational data from it. Next we describe our methodology for analyzing the extracted data. Section 6 presents initial analyses results. Section 7 reports on the limitations and of our study. Section 8 points out directions for future work.

2 The Enron Case

Enron - What happened?

Enron was formed in 1985 under the direction of Kenneth Lay through the merger of Houston Natural Gas, a utility company, and Internorth of Omaha, a gas pipeline company. The company was based in Houston, Texas. Within 15 years Enron became the nation's seventh-biggest company in revenue by buying electricity from generators and selling it to consumers. The company quickly adapted to the deregulation of the energy market by positioning themselves as an energy broker: Enron identified areas where energy needs where higher than energy capacities, built power plants in such regions, sold the plants before their value diminished, and moved on to new areas with mismatches of power needs and capacities [28]. Later the company applied and expanded their middlemen skills and derivate trades to newer markets such as TV ad time and bandwidth. In 2002, Enron employed 21,000 people in more than 40 countries [10].

From 1985 on, Arthur Andersen, LLP (Andersen) had been Enron's auditor. Andersen earned tens of millions of dollars from accounting and internal and external consulting services for Enron, which was one of Andersen's largest clients worldwide. Enron employed many former Andersen workers.

In 1999, Enron officials began to separate losses from equity and derivate trades into "special purpose entities" (SPE); partnerships that were excluded from the company's net income reports. An example of such an SPE was Raptor, a liaison of Enron executives, who bought equity shares in two companies, New Power Co. and Avici, with loaned stock money from Enron. Enron profited from the increase of the value of the SPE's shares but had Raptor booking the losses, thus excluding them from their financial reports. The systematic omission of negative balance sheets and income statements from SPEs in Enron's reports resulted in an off-balance-sheet-financing system [28].

In December of 2000, president and chief operating officer Jeffrey Skilling took over the position of chief executive from Kenneth Lay. Lay remained chairman while the Enron stock hit a 52-week high of $84.87. In August 2001 Skilling surprisingly resigned, stating personal reasons for quitting. Lay was named as Enron's chief officer and CEO again in 2001 [20]. In the same month Sherron Watkins, Enron's Vice-President of Corporate Development who became famous as Enron's whistle-blower, wrote an anonymous letter to Lay in which she accused Enron of possible fraud and improprieties such as the SPEs [31]. Andersen knew of the information provided by Sherron Watkins.

In October 2001 the losses transferred from Enron to the SPE's totaled over $618 million and Enron publicly reported this amount as net loss for the third quarter. By the end of the year Enron disclosed a reduction of $1.2 billion in the value of shareholders' stake in the company. One of the people associated with the crash was Andrew Fastow, chief financial officer, who had supported Enron in inflating profits and hiding debts [28].

On October 31, 2001, the Securities and Exchange Commission (SEC) started an inquiry into Enron. Enron subsequently ousted Fastow and announced that the SEC investigation revealed that the amount of losses for the previous five years was actually $586 million. The market reacted with a fast and sharp drop of the value of Enron's shares to levels below $1 in November 2001. Being forced to transfer stocks in order to satisfy the losses, Enron became insolvent and filed for bankruptcy in December 2001. The fallout and investigations into the Enron collapse continued throughout 2002. Lay resigned as chairman and CEO in January of 2002, and less than two weeks later from the board [1].

Long before Enron's official insolvency, Andersen had possessed knowledge of Enron's organizational situation and financial performance but did not communicate the information to the public [28]. Andersen and Enron intentionally categorized hundreds of millions of dollars of shareholders equity that were a decrease as an increase. Andersen, who did some of Enron's internal bookkeeping, advised Enron not to refer to charges against the third quarter income of 2001 as non-recurring, but did not make this information available for the public. In 2000 Andersen's internal Senior Management already had rated Enron lower then they evaluated the client publicly. Before Enron released its notice of net loss, Andersen retained a New York based law firm from handling further Enron-related issues and took over all legal matters regarding Enron. In late October 2002, Andersen instructed Enron to destroy documentation related to Enron.

Andersen was indicted for altering, destroying and concealing Enron-related material and persuading others to do the same in March 2002 [36], convicted of obstruction in June 2002, and received a probationary sentence and a fine of

$500,000 in October 2002. In 2002 Andersen got banned from auditing public companies.

Lay, Fastow and former top aid Michael Kopper appeared before Congress in February of 2002; all three of them invoked the Fifth Amendment [10]. Skilling testified twice before Congress the same month, stating that he was unaware of any accounting problems. Fastow was indicted in October 2002. Ben Glisan Jr., a former Enron treasurer, pleaded guilty to conspiracy in September 2003, and became the first former Enron executive being imprisoned [1]. Fastow pleaded guilty in January 2004 [10]. His wife, Lea Fastow, and seven former Enron executives also got charged. In February 2004 Skilling got charged with fraud, conspiracy, filing false statements to auditors and insider trading [20]. In July of 2004 Lay surrendered to the FBI and was accused of participating in a conspiracy to manipulate Enron's quarterly financial results, making false and misleading public statements about Enron's financial situation, omitting facts necessary to make financial statements accurate and fair, civil fraud, and insider trading.

In March of 2003 Enron announced a plan to emerge from bankruptcy as two separate companies. In July the company filed a reorganization plan stating that most creditors would receive about one-fifth of the $67 billion they were owed.

Research on the Enron Case

Much information is available on the Enron case[2], including some details on organizational aspects of the company that might relate to its failure, such as a certain organizational culture. However, no studies of the case have been published yet in the Organizational Science and Social Networks literature.

The Board Investigation Committee stated in February 2002 that Enron's board may have been withholding critical information and had been unable to or prevented from providing checks and balances that would have been necessary to assure ethical business practices[26]. The Congressional Commission reported that Enron's culture encouraged employees to push the limits [26].

The Management Institute of Paris (MIP) identified Enron's and Andersen's senior managers as those in charge of Enron's failure. According to them, Enron's management misled the public, lacked moral leadership and ethics, and created an organizational culture of greed, secrecy and winner-take-all mentality. In 2001 Andersen evaluated Enron's financial statements as adequate and reliable and their financial conditions as fair [22].

Based on an article in Fortunes Magazine that explains the bankruptcy of over 257 companies in 2001 with managerial errors rather than with extra-organizational factors, which are usually claimed by the management, MIP points out ten executive errors that lead to Enron's failure [23]. These factors can be grouped into three categories: misperception of reality, risk-taking organizational culture, and improper crisis management.

Misperception of reality occurred in Enron on managerial level, because a) executives ignored bad news since it did not fit into their mental models of success that they had build up previously, b) mangers blinded out perceived problems instead of tackling them, and c) employees mitigated problems they reported to their supervisors for fear of the rogue character of Enron's managers (for example, Sherron Watkins having sent her letter anonymously to Lay). Instances of Enron's risk-taking culture are the foundation of SPE's, the overdosing of risk by not providing liability for the SPE's losses, and the greedy profit taking without disclosure. Enron's improper crises involved the implementation of ad-hoc strategies, hoping for a quick solution of all difficulties and lacking a thorough analysis of the problem.

While first thoughts about the relationship between Enron's risk-pushing organizational culture in connection with managerial errors and the company's failure are being released, no network analytic studies have been published that explore the social network phenomena in Enron (with exception for the social network generated by Shetty and Adibi [33]) .

Network analysis focuses on the relations among and between entities in a social or organizational system (see for example [29][38]). In our case, the system is Enron and the entities are former Enron employees. In a social network the entities are represented as nodes, and the relations between them as edges or links. We base the research presented in this paper on the assumption that the relations among Enron's employees are represented in the exchange and content of the emails that are contained in the Enron corpus. In our study we focus on the analysis of the exchange of emails. We refer to this type of networks as communication networks because these networks represent flow of messages among communicators across space and time [24]. Since the messages are sent from one agent to one or multiple other agents, the resulting networks are directional or digraphs.

[2] See material from agencies such as SEC [30], Federal Energy Regulation Commission (FERC) [12], United States Department of Justice (DOJ), Commodities Futures Trading Commission (CFTC) [6], General Accounting Office (GAO), Investigative Committee of the Board of Directors of Enron [26], and management related organizations [15].

The lack of research on Enron from a network analytic perspective motivates our research questions:

What are the structure and properties of the communication networks in Enron? How do these features relate to other networks?

Who are key players or critical individuals in the system? (On the concept of key players see [3]).

How do structure and key players change over time?

Our research questions are of an explorative nature and aim to gain a first understanding of relations between individuals in Enron. Answers to these questions will provide researchers with knowledge that can help to understand and explain this particular organization and relate this information to Enron's life cycle of success, crisis and bankruptcy. The network analytic perspective enables the investigation of vulnerabilities of the system and its adaptive capabilities to changing situations. Furthermore, the relational data that we extract and its analysis could be deployed to further develop theories or validate hypotheses about the evolution of communication networks.

3 Data

There is not *the* Enron email corpus available, but multiple instances of it. The Federal Energy Regulatory Commission (FERC) originally posted the Enron email database on the internet in May of 2002 to enable the public to understand why FERC investigates Enron [12]. The database consists of 92% of Enron's staff emails. FERC collected a total of 619,449 emails from 158 Enron employees, mainly from senior managers. Each email contains the email address of the sender and receiver, date, time, subject, body and text. Attachments were not made available. FERC's version of the database had a lot of integrity problems. Leslie Kaelbing from MIT then purchased the dataset. Later a group of people at SRI, notably Melinda Gervasio, collected and prepared the data for the CALO project [34]. The SRI group corrected most of the integrity problem and made the dataset available.

William Cohen from CMU put the dataset online for researchers in March 2004 [7]. This version of the database contains 517,431 distinct emails from 151 users. The emails are organized in 150 user folders that have further subfolders; with the total number of folders in the corpus totalling 4700. The corpus has a size of 400Mb. Some messages were deleted "as part of a redaction effort due to requests from affected employees" [7]. Invalid email addresses were converted to addresses of the form user@enron.com when a recipient was specified and to no_address@enron.com when no recipient was specified.

Andres Corrada-Emmanuel from the University of Massachusetts further explored the dataset by using the MD5 digest of the body of the emails. He found out that the corpus actually contains 250,484 unique emails from 149 people [8].

The version of the dataset that we are using was provided by Jitesh Shetty and Jafar Adibi from ISI [33]. The ISI people cleaned up the dataset by dropping emails that were blank, duplicates of unique emails, had junk data, or were returned by the system due to transaction failures. The resulting corpus contains 252,759 emails in 3000 user defined folders from 151 people. Shetty and Abidi put the information in a MySql database that contains four tables, one for each of the entities of employees, messages, recipients and reference information. We chose this version of the corpus for our work, because the process of cleaning the dataset seems very helpful to us and is well documented. Furthermore, the structure and content of the MySQL database met our needs.

The database contains many emails by individuals who were not involved in any of the actions that are subject of the Enron investigation.

4 Database Refinement and Extraction of Relational Data

In order to perform network analysis on the Enron corpus, it is necessary to extract relational data. The relations among and between the entities in Enron are reflected in a) the email exchanged between the employees (communication networks) and b) the actual content of those messages. In this paper we concentrate on the extraction and explorative analysis of the first type of data. All database work and data extraction was performed on a Linux machine with Perl modules that we wrote for this purpose.

The data in the corpus is multi-mode (e.g. work relationship, friendship), multi-link (connections across various meta-matrix entities) and multi-time period. Nodes and edges can have multiple attributes such as the position and location of an employee or the types of relationships between two communication partners (multi-mode). We refer to data that is multi-mode, multi-link and multi-time period in which both nodes and edges can have attributes that carry information on how to interpret, evolve, and impact these nodes and edges as "rich" data. In order to adequately represent and analyze the information contained in the corpus we need a data format that can handle rich social network data and can be used as input and output of multiple analysis tools that we consider to use. We chose to use DyNetML as the data format because it meets our data format requirements [35]. DyNetML is an XML based interchange language for relational data. A

DyNetML file can represent an arbitrary number of node sets and graphs. Node sets group together nodes of the same type, e.g. agents, complete with any rich data such as an agent's position or location. Each graph consists of a set of edges that connect nodes, complete with any rich data attached to the graph itself or any of its edges.

Database Refinement

DyNetML files for the representation of communication networks require data from three tables in the ISI database: The message ID, which includes time information, the sender, and the recipient. The information provided on the individuals is their first and last names and one email address. More information on properties of the individuals would enable a more thorough analysis and deeper understanding of processes in Enron. Such properties can be represented as attributes of nodes that represent agents in DyNetML. We found three additional sources of information on some of the Enron employees: A file with the positions of former employees from ISI (ISI position file) [32], a list with job information from FERC/ Aspen (FERC position file) [11], and a list from FERC/ Apsen with information on people's location (FERC location file) [13]. Note, most of the information on FERC's Western Energy Markets investigation is hosted on Aspen Corporation websites.

The ISI position file lists the names of 161 Enron employees, and for 132 of them it provides position information. ISI gathered this information from various sources, mostly from Federal Court documents which were publicly released. For 29 people no status information is provided because they, according to Shetty, were not involved in the Enron case and did not hold high posts in the company, or were employed for a only short period of time. In the social network generated by Shetty and Adibi those 29 people are assigned to the position of an employee (Table 1)[3]. The FERC position file is a list of authorized traders that contains names, positions, a few locations and trade related information on individuals from Enron and probably other companies. The FERC location file is an interoffice memorandum sent by John Lavorato to Donna Lowry from Risk Assesment and Control on October 12, 2001. In this file, people are sorted by locations – East, Central, Texas, West and Canada.

[3] The ISI position file contains two sets of names that seem semantically highly similarity: Micheal Swerzzbin/ Vice President; Mike Swerzbin/ Trader; James Schweiger/ Vice President; Jim Schwieger/ Trader. We were skeptical if Swerzzbin/ Swerzbin and Schweiger/ Schwieger were distinct individuals, therefore we matched those names against both FERC files. Based on this comparison we selected Mike Swerzbin and Jim Schwieger as unique individuals, because they appeared in the FERC files, and dropped Micheal Swerzzbin and James Schweiger, because they were not listed in the FERC files.

We added the position and location information to a new instance of the Enron database that we built. We refer to our instance of the database as the Enron CASOS database. We realized that in many cases the spelling of names did not match between the files from ISI, FERC and the database. In order to find the names in the database that are most similar to the spellings in the ISI and FERC files we used a semantic similarity algorithm [36][21] implemented in the String Similarity Perl module [19], and ran it against the database. The similarity function computes a similarity value between 0 (no similarity) and 1 (identical strings), based on how many edits are necessary to convert one string into another. We output the 25 highest scoring suggestions from the module and picked the one that we manually evaluated to represent the same name that is provided in the database. After we had identified the matching pairs we added the position and/ or location information to these names as provided in the ISI and FERC files to the database while maintaining the spelling of the names as originally defined in the database. During this process we encountered various cases of conflicting information: In 36 cases we had different position information from ISI and FERC. We assume that this is because people got promoted or changed positions. Since we did not have time information for both of the files, our default was to pick the higher position. The location information in the two FERC files was conflicting in five cases. We picked the information from the location file because it had a date on it, which was in the middle of the crisis, and seemed more focused on location information. After using the heuristics described here we enhanced our instance of the database with the position and location information.

Overall, we identified 15 unique job titles that we associated with 212 employees (Table 1), 5 unique locations that relate to 67 people, 102 employees for which we have position and location data, and 227 employees for which we have either position or location information. For five of the 29 people that ISI had no position information on we were able to identify a job title. The further data adjustment and analyses in this paper mainly concentrate on the 227 employees whose names and/ or positions we know. A file with detailed information on this subset of people such as their first and last names, position, additional information on the position, location and source of this information is available from the authors but was not included in this paper to protect the individuals' privacy. We use this subset as a point of departure for our work on the Enron data, and

will include more people who appear in the database once we obtain more information on them.

Table 1: Number of Individuals per Position

Position	ISI position file	ISI social Network	CASOS Enron database
Analyst	0	0	10
Associate	0	0	5
CEO	4	4	4
Director	23	12	27
Employee	40	85	69
Head	0	0	2
In House Lawyer	3	3	3
Manag. Director	5	3	6
Manager	13	10	31
President	4	4	4
Specialist	0	0	9
Sr. Specialist	0	0	17
Trader	12	12	9
Treasury Support	0	0	2
Vice President	28	18	29
Total	**132**	**151**	**227**

Next we normalized the email addresses for the subset of 227 people. We assumed that people might have more than the one email address specified for them in the ISI corpus. Note that the spelling of emails in the database matches the spelling in the ISI position file. Corrada [8] provides a list of 31 email addresses that mainly resemble the addresses in the ISI file, but gives two addresses for only two out of 29 individuals. We further explored this issue by using the similarity function described above to search for all email addresses ending with @enron.com for addresses similar to those specified in the employeeList table in the ISI database. The module identified the 25 highest scoring hits per address, and we manually vetted them. We found that a similarity greater than 0.7 usually indicates a match and selected these by default prior to review. Table 2 provides quantitative information on the process of email normalization.

Table 2: Statistics of Email Address Normalization

	Emails referring to 227 agents	Emails added	Emails dropped
Sum	429	92	41
min	1	0	0
max	8	3	8
Average	1.89	0.41	0.18
STD	1.18	0.71	0.72

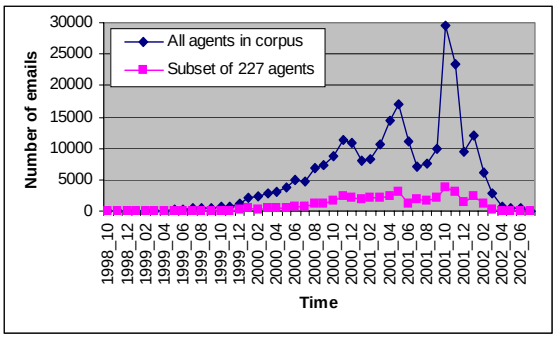
To summarize our work on the database, we have refined it by resolving name ambiguities and enhanced its information by adding the position and/ or location of 227 individuals, as well as normalizing their email addresses.

Extraction of Communication Networks

Next we extracted DyNetML files that represent the communication among the subset of 227 people. Out of the 227 individuals, a union of 209 people exchanged emails amongst each other. We time sliced our data in order to enable longitudinal analysis[4]. We decided to time slice the corpus on a monthly basis from October 1998 to July 2002, as this seemed to entail time spans in which major events occurred. This resulted in 46 DyNetML files that represent the agents as nodes and exchange of emails between them as edges. The number of agents in each file can differ since the size of the population can vary from month to month. Each edge denotes a directed relation of type agent to agent. The edges are weighted by the cumulative frequency of emails exchanged between individuals per month.

Figure 1 shows the total number of emails sent by all individuals in the corpus as well as by the people in our subset across months. Both curves show peaks in the amount of communication; some of them can be related to events in the organization. The highest peaks occurred in October 2001 (29,556), the month in which the Enron crisis broke out, November 2001 (23,441), when the investigations were under way, and May (16,986) and April (14,348) 2001. The low points, which are in January and February 2000 and from August to September, might be explained as being vacation periods. The curve for the subset resembles the pattern of the curve for the entire corpus.

Figure 1: Number of Emails Sent per Month



[4] The time slicing returned 327 emails from the entire corpus with invalid dates such 2044-01 or 0001-12. Since no correct date information was given in those emails we excluded those emails from further analysis. This reduced the corpus by 0.13% to 252,432 emails.

5 Methodology

We use ORA [5] to analyze the communication networks. Since we have position information on agents available we can compare the formal and informal organizational structure. We are also able to explore changes in the network over time by comparing a network from a month during the Enron crisis with a network from a month in which no major negative happenings are reported and where the organization seemed to be on a successful path. We picked October 2000 and 2001 for this comparison. We first run an intel report in ORA that computes network analytic measures on a graph level and identifies key agents in the network. Next we run an ORA context report that compares the graph level measures from the intel report for Enron with values for real networks stored in a CASOS database as well as with numbers computed on a directed uniform random graph of identical size and density as the Enron networks. Then we run an ORA risk report that identifies critical individuals who bear risks for an organization. The risk is computed for every agent as well as the entire network with respect to the agents' communication, performance, interaction, and redundancy. This report allows researchers to explore the distribution of a particular type of risk across an organization, thus identifying systemic versus individualistic problems.

6 Results

Figures 2 and 3 show the network structure by position for Oct. 2000 (160 agents) and 2001 (174 agents). The visualizations were generated with the NetDraw software [4]. Both graphs contain only a few isolates (one in Oct. 2000, 2 in Oct. 2001), which represent individuals who are not connected to others. ORA's intel report reveals that the Oct. 2001 graph is denser than the Oct. 2000 graph: The Oct. 2000 network has a lower overall completeness, expressed in the value of density (0.018), than the Oct. 2001 network (0.031). Mathematically densities range form 0 to 1, with higher values indicating denser graphs (for more details on network analytic measures see [5][38]). Looking at the number of weak or undirected components (2 in Oct. 2000, 3 in Oct. 2001) we learn that in both graphs all individuals, except for the isolates, are in one component. This means that in both networks each person can reach each any other person. Components are maximally connected subset of nodes, also referred to as subgraph. Weak components do not consider directionality of a link, whereas strong components take a link's directionality into account. The existence of components indicates that a graph is disconnected. The number of directed components is higher for Oct. 2000 (96) than for Oct. 2001 (39). This result suggests that during the crisis there are fewer disconnected subgroups of people who mutually exchange emails than in a normal month. The values of density and number of strong components indicate that during the crisis the communication among Enron employees has been intensified and spread out through the network in comparison to a month before the crisis.

Figure 2: Communication Network October 2000

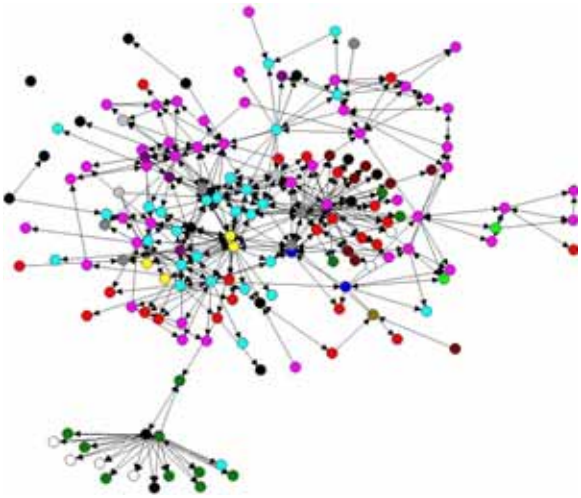
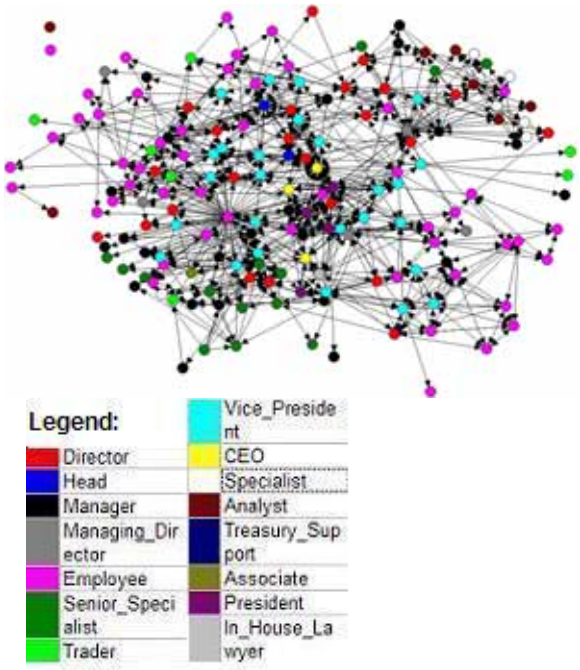


Figure 3: Communication Network October 2001



In order to put graph level measures for Enron into a broader context we run an ORA context report. Graph level centralization measures express the degree to which single actors have high

Table 3: Graph Level Measures in Comparison

Measure	Oct. 2000	Oct. 2001	Social Networks	Interpretation: On average ...
Betweenness Centrality	0.008	0.012	0.047	there are fewer paths by which information can flow from any one person to any other person in this group compared to other groups.
Closeness Centrality	0.031	0.253	0.380	it takes more steps for information to get from any person in this group to any other person in this group compared to other groups.
Eigenvector	0.046	0.055	0.165	this group is less cohesive than other groups.
Total Degree	0.018	0.031	0.284	each person in this group has fewer connections to others than people in other groups.
Strong Components	96	39	8.455	there are more components in this group than in other groups: i.e. it is more disconnected.

importance or prominence in a network and others have low centrality. Thus, graph centrality represents the heterogeneity or dispersion of the agents' centralities in a network[5]. The ORA results (Table 3) show that both Enron networks are less centralized than other networks, and that the Oct. 2000 graph is less centralized than the Oct. 2001 graph. These findings suggest that during the crisis the inequality of the importance of the employees, the amount of communication, and the group cohesion increased. The results also suggest that a highly segmented workforce with little cross communication may have been a factor that supported the frauds in Enron.

In order to identify the most important people in the network centralization measures can be computed on an individual level. Table 4 shows the 5 individuals who score highest in the Oct. 2000 and Oct. 2001 network with respect to the following centrality measures: Closeness centrality describes how close an actor is to all other actors. Betweenness centrality measures how often an actor is positioned on the shortest path between any other pair of actors. Eigenvector centrality tells us how close an actor is to other actors who are important with respect to degree centrality, and an actors' degree is the number of other actors directly linked to him or her. Since the Enron networks are directed, we split up centrality into outdegree (actors adjacent from an actor) and indegree (actors adjacent to an actor). Table 4 contains a union of 21 distinct people (13 distinct ones in Oct. 2000, 14 distinct ones in Oct. 2001), and 6 of them appear in both months. The intersection of individuals per measure in Oct. 2000 and Oct. 2001 is low and varies between 0 and 3. For the people who appear in both months their position in the ranking changes as often as it remains the same (4 times) from Oct. 2000 to Oct. 2001. Looking at the key players' formal positions the results show that for closeness centrality people with lower positions appear more often among the most central individuals in Oct. 2001 than in Oct. 2000. This observation does not apply to the other measures, but in general people with higher positions are more likely to be key players in this organization. Analyzing the values for closeness centrality for all 209 agents across all 46 months (Figure 4) reveals that the values per individuals are less different from each other than for other

Table 4: Key Players per Centrality Measures

October 2000			October 2001		
Value	**Name**	**Position**	**Value**	**Name**	**Position**
Closeness Centrality					
0.07	W. Stuart	Manager	0.21	S. Beck	Employee
0.07	D. Delainey	CEO	0.20	L. Kitchen	President
0.07	C. Dorland	Manager	0.19	S. Kean	VP
0.07	J. Derrick	Lawyer	0.19	S. White	Employee
0.07	T. Belden	Mang. Dir.	0.18	J. Dasovich	Employee
Betweenness Centrality					
0.11	D. Delainey	CEO	0.24	L. Kitchen	President
0.10	R. Sanders	VP	0.16	S. Beck	Employee
0.08	T. Belden	Mang. Dir.	0.13	T. Belden	Mang. Dir.
0.08	J. Lavorato	CEO	0.10	J. Lavorato	CEO
0.08	J. Dasovich	Employee	0.07	M. Grigsby	Head
Eigenvector Centrality					
0.60	J. Dasovich	Employee	0.69	J. Dasovich	Employee
0.54	J. Steffes	VP	0.52	J. Steffes	VP
0.41	M. Hain	Lawyer	0.40	R. Shapiro	VP
0.31	R. Shapiro	VP	0.23	S. Kean	VP
0.19	R. Sanders	VP	0.13	B. Tycholiz	VP
In Degree Centrality					
0.80	J. Steffes	VP	0.77	R. Shapiro	VP
0.46	R. Shapiro	VP	0.76	J. Lavorato	CEO
0.42	T. Belden	Mang. Dir.	0.66	B. Tycholiz	VP
0.36	M. Taylor	Employee	0.66	J. Steffes	VP
0.33	R. Sanders	VP	0.49	L. Kitchen	President
Out Degree Centrality					
1.08	J. Dasovich	Employee	1.63	D. Delainey	Employee
1.01	M. Hain	Lawyer	1.51	M. Grigsby	Head
0.96	T. Jones	Employee	1.04	B. Williams	Analyst
0.81	D. Delainey	CEO	0.90	S. Beck	Employee
0.48	T. Belden	Mang. Dir.	0.76	J. Steffes	VP

[5] On graph and node level, betweenness and closeness centrality vary between 0 and 1. Eigenvector and degree centrality can reach values higher than 1. The higher the value the more central is a network or an agent in a network.

measures (for example eigenvector centrality Figure 5). These results suggest that in 2000 Enron had a segmented culture with directives being sent from on-high and sporadic feedback. By 2001, the VP's and other executives had formed a tight knit clique supporting each other and whose interactions with the rest of Enron are highly brokered.

Table 5: Emails Exchanged per Month

	October 2000		October 2001	
Position	**sent**	**received**	**sent**	**received**
CEO	71%	29%	27%	73%
President	58%	42%	53%	47%
VP	38%	62%	44%	56%
Man. Dir.	43%	57%	57%	43%
Director	8%	92%	41%	59%
Head	57%	43%	79%	21%
Manager	53%	47%	42%	58%
Lawyer	72%	28%	52%	48%
Sr. Specialis	27%	73%	45%	55%
Specialist	0%	100%	29%	71%
Analyst	20%	80%	61%	39%
Associate	20%	80%	50%	50%
Employee	55%	45%	57%	43%
Trader	62%	38%	32%	68%

To further explore the relationship between positions and different situations in the company as well as the correspondence of the formal position network with the informal one we compared the amount of emails exchanged between positions (Tables 5, 6) for October 2000 and 2001.

Table 5 indicates that in contrast to Oct. 2000 in Oct. 2001 the CEOs, Heads, Managers and Traders sent more emails than they received, whereas the Managing Directors and Analysts received more messages than they sent. The major shift from 2000 to 2001 is that in 2000 higher rank positions tended to be directive (send more than receive) whereas by 2001 they became consumers (receive more than send). The major exception here are the VP's who have always been consumers and if anything became more directive.

The results in Table 6 show that high ranking positions (1 to 6 and 8 in Table 5) perform more top-down communication than the send information to higher ranks. In contrast, lower ranks send more communication up the hierarchy or within the same rank. Table 6 suggests that during the crisis 9 out of 12 positions communicate less with the same position or rank than they did in Oct. 2000. The differences of the percentages of emails sent to higher and lower ranks are less in Oct. 2001 than in Oct. 2000. Those findings indicate that during the crisis the communication has been more diverse with respect to formal positions than during a normal month. Furthermore, in contrast to Oct. 2000 in Oct. 2001 the Heads tended to communicate more often with lower ranks than with higher ranks and the Sr. Specialists more often sent messages to higher ranks than to lower ranks.

7 Limitations

The main limitation of our study is that we have not validated the relation data we have extracted and analyzed yet. In order to perform validation we will compare our data and findings against material from reliable sources such as reports and press articles on the Enron case, letters from and interviews with former Enron employees, and information from other people with direct insight into the company. Once we have such material we also will evaluate the extracted networks by analyzing what portion of the relevant links we have captured (recall) and what portion of the captured links is actually relevant (precision).

We note that the results presented herein cannot be generalized for the Enron organization or other corporations since we analyzed only two time points and a subset of 227 people.

8 Conclusion and Future Work

In this paper we have described how we enhanced and refined the Enron database. We have reported on the extraction of relational data from our instance of the database. Our initial results, which are based on snapshots of Enron's communication network at 2 time points, suggest that in Oct. 2001 the network had been denser, more centralized and more connected than in Oct. 2000. We also learned that about half of the people who were key players in Oct. 2000 were also the most important in the network of Oct. 2001. Our data suggests that during the crisis the communication among Enron's employees had been more diverse with respect to people's formal positions and that the top executives had formed a tight clique with mutual support and highly brokered interactions with the rest of organization.

In our future work we will consider all points of time that we extracted network data for and a larger set of people in order to learn more about this network and how its properties and entities relate to various phases of the company's life cycle of success, crisis and failure.

In the future we will analyze the actual content of the emails via Network Text Analysis [25][9] in order to explore the perception of the company's situation on an individual and group level, as well as across time. We will extract these perceptions as mental models, which are representations of the reality that people use to make sense of their surroundings [14][27].

Figure 4: Closeness Centrality of 209 Agents over Time

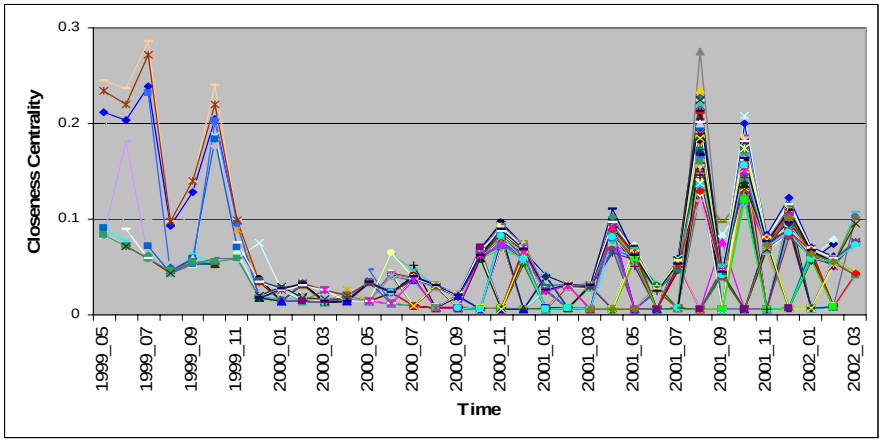


Figure 5: Eigenvector Centrality of 209 Agents over Time

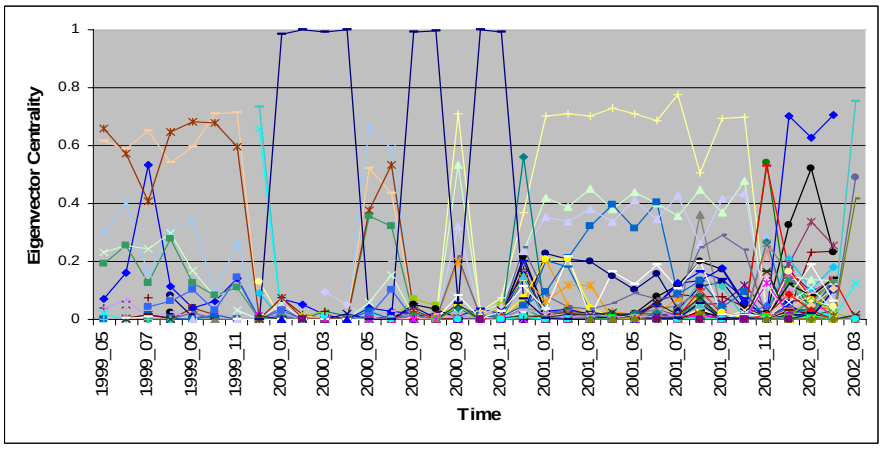


Table 6: Emails Sent to Positions

		October 2000			October 2001		
Rank	**Position**	**higher rank**	**lower rank**	**same pos. & same rank**	**higher rank**	**lower rank**	**same pos. & same rank**
1	CEO	NA	83%	17%	NA	100%	0%
2	President	12%	85%	3%	26%	54%	20%
3	VP	9%	58%	33%	14%	45%	41%
4	Man. Dir.	20%	75%	5%	30%	69%	1%
5	Director	27%	64%	9%	35%	43%	22%
6	Head	56%	31%	13%	43%	50%	7%
7	Sr. Specialist	6%	28%	66%	54%	30%	16%
8	Lawyer	89%	10%	1%	87%	13%	0%
9	Manager	18%	24%	59%	20%	49%	31%
10	Specialist	0%	0%	0%	17%	34%	49%
11	Analyst	40%	NA	60%	60%	NA	40%
11	Associate	100%	NA	0%	100%	NA	0%
11	Employee	40%	NA	60%	53%	NA	47%
11	Trader	NA	NA	98%	38%	NA	62%
11	Treas. Support	0%	0%	0%	100%	NA	0%

Mental models can be conceptualized as cognitive constructs that help researchers to gain an insight into how knowledge and information are represented in people's minds [16]. Since organizational culture is also represented in messages [24], we also will analyze the mental models to learn about Enron's culture.

References

[1] *A Chronology of Enron Corp.* (2004). NewsMax Wires. Retrieved October 13, 2004, from http://www.newsmax.com/archives/articles/2004/7/8/110332.shtml

[2] Bekkerman, R. (n.d.). Retrieved November 4, 2004, from http://www.cs.umass.edu/~ronb/

[3] Borgatti, S. P. (2004). The Key Player Problem. In R. Breiger, K. M. Carley, & P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis: 2002 Workshop Summary and Papers* (pp. 241-52). Washington, DC: National Academies Press.

[4] Borgatti, S.P. (2002). *NetDraw1.0. Graph Visualization Software*. Harvard: Analytic Technologies.

[5] Carley, K.M., & Reminga, J. (2004). *ORA: Organization Risk Analyzer*. Technical Report, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, URL: http://www.casos.cs.cmu.edu/projects/ora/publications.html

[6] *CFTC Enron Information Link Page*. (2003). Retrieved October 9, 2004, from http://www.cftc.gov/enf/enron/enfenrondefault.htm

[7] Cohen, W.W. (n.d). *CALD, CMU*. Retrieved October 5, 2004, from http://www-2.cs.cmu.edu/~enron/

[8] Corrada-Emmanuel, A. (n.d.). *Enron Email Dataset Research*. Retrieved October 5, 2004, from http://ciir.cs.umass.edu/~corrada/enron/

[9] Diesner, J., & Carley, K.M. (2005). Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. In V.K. Narayanan & D.J. Armstrong (Eds.), *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations*, (pp. 81-108). Harrisburg, PA: Idea Group Publishing.

[10] *Enron Scandal at a Glance.* (2002). BBC News. Retrieved October 13, 2004, from http://news.bbc.co.uk/1/hi/business/1780075.stm

[11] *FERC position file*. (n.d.). Retrieved October 10, 2004, from http://ferc.aspensys.com/FercData/Miscellaneous%20CD's/Box005/Response%20to%20Request%2015/RAC/Compliance/Authorized%20Trader%20Lists/Authorised%20Traders%20List5_11_01.pdf

[12] *FERC Western Energy Markets - Enron Investigation, PA02-2*. (n.d.). Retrieved October 18, 2004, from http://www.ferc.gov/industries/electric/indusact/wem/pa02-2/info-release.asp

[13] *Ferc/ Apsen Location file*. (n.d.). Retrieved November 4, 2004, from http://ferc.aspensys.com/FercData/Miscellaneous%20cd's/Box005/Response%20to%20Request%2015/RAC/Compliance/Authorized%20Trader/Authorized%20Trader%20Memos%20Dtd%2010-01/North%20American%20Natural%20Gas%2010-01.pdf

[14] Johnson-Laird, P. (1983). *Mental Models*. Cambridge, MA: Harvard University.

[15] Kefgen, K., & Kogen, M. (2002). *Enron Anyone?* HVS International. Retrieved November 4, 2004, from http://www.hvsinternational.com/emails/execsearch/hospitality/7-26.htm

[16] Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management* 20, 403-437.

[17] Klimt, B., & Yang, Y. (2004). Introducing the Enron Corpus. First Conference on Email and Anti-Spam (CEAS), Mountain View, CA. Retrieved October 14, 2004, from http://www.ceas.cc/papers-2004/168.pdf

[18] Klimt, B., & Yang, Y. (2004). The Enron Corpus: A New Dataset for Email Classification Research. *European Conference on Machine Learning*, Pisa, Italy.

[19] Lehmann, M. (n.d.). *String Similarity*. From http://search.cpan.org/~mlehmann/String-Similarity-1/Similarity.pm

[20] *Lights out at Enron.* (2003). CBSNews.com. Retrieved October 13, 2004, from http://www.cbsnews.com/stories/2003/02/06/60minutes/main539719.shtml

[21] Meyers, E.W. (1986). An O(ND) Difference Algorithm and its Variations. *Algorithmica*, 1(2).

[22] MIP. (2002). Enron: *Who is really to blame?* Retrieved November 11, 2004, from http://www.mip-paris.com/knowledge/article.asp?id=21.

[23] MIP. (2002). *Fortune Magazine's List of 10 Corporate Sins*. Retrieved November 11, 2004, from http://www.mip-paris.com/knowledge/article.asp?id=132

[24] Monge, P.R., & Contractor, N.S. (2003). *Theories of Communication Networks*. New York: Oxford University Press.

[25] Popping, R. (2000). *Computer-assisted Text Analysis*. Thousand Oaks, CA: Sage Publications.

[26] Powers, W.C. (2002). *Report of Investigation, By the Special Investigative Committee of the Board of Directors of Enron Corp*. Retrieved November 4, 2004, from http://news.findlaw.com/hdocs/docs/enron/sicreport/sicreport020102.pdf

[27] Rouse, W.B., & Morris, N.M. (1986). On looking into the black box; prospects and limits in the

search for mental models. *Psychological Bulletin* 100, 349-363.

[28] Sanborn, R. (n.d.). *Enron*. Retrieved November 4, 2004, from http://www.hoylecpa.com/cpe/lesson001/Lesson.htm

[29] Scott, J. (2000). *Social Network Analysis*. London: Sage, 2nd edition.

[30] *SEC Spotlight on Enron*. (n.d.). Retrieved November 4, 2004, from http://www.sec.gov/spotlight/enron.htm

[31] *Sherron Watkins eMail to Enron Chairman Kenneth Lay*. (2002). Retrieved November 11, 2004, from www.itmweb.com/f012002.htm

[32] Shetty, J., & Adibi, J. (n.d.). *Ex employee status report*. Retrieved November 4, 2004, from http://www.isi.edu/~adibi/Enron/Enron_Employee_Status.xls

[33] Shetty, J., & Adibi, J. (n.d.). *The Enron Dataset Database Schema and Brief Statistical Report*. Retrieved November 4, 2004, from http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf

[34] *SRI International*, CALO (Cognitive Assistant that Learns and Organizes). (2004). Retrieved November 4, 2004, from http://www.ai.sri.com/project/CALO

[35] Tsvetovat, M., Reminga, J., & Carley, K.M. (2003). *DyNetML: Interchange Format for Rich Social Network Data*. CASOS Technical Report, Carnegie Mellon University, School of Computer Science, Institute for Software Research International, URL: http://reports-archive.adm.cs.cmu.edu/anon/isri2004/abstracts/04-105.html

[36] Ukkonen, E. (1985). Algorithms for approximate string matching. *Information and Control*, 64, 100-118.

[37] *United States District Court Southern District of Texas, Indictment*. (2002). Retrieved October 8, 2004, from http://news.findlaw.com/hdocs/docs/enron/usandersen030702ind.pdf

[38] Wasserman, S., & Faust, K. (1994). *Social Network Analysis*. New York: Cambridge University Press.