
Computational & Mathematical Organization Theory, 11, 201–228, 2005

Communication Networks from the Enron Email Corpus
"It's Always About the People. Enron is no Different"[1]

JANA DIESNER
TERRILL L. FRANTZ
KATHLEEN M. CARLEY
Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International (ISRI), School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, 1327 Wean Hall, Pittsburgh, PA 15213, USA
email: jdiesner@andrew.cmu.edu
email: terrillf@andrew.cmu.edu
email: kathleen.carley@cs.cmu.edu

Abstract

The Enron email corpus is appealing to researchers because it represents a rich temporal record of internal communication within a large, real-world organization facing a severe and survival-threatening crisis. We describe how we enhanced the original corpus database and present findings from our investigation undertaken with a social network analytic perspective. We explore the dynamics of the structure and properties of the organizational communication network, as well as the characteristics and patterns of communicative behavior of the employees from different organizational levels. We found that during the crisis period, communication among employees became more diverse with respect to established contacts and formal roles. Also during the crisis period, previously disconnected employees began to engage in mutual communication, so that interpersonal communication was intensified and spread through the network, bypassing formal chains of communication. The findings of this study provide valuable insight into a real-world organizational crisis, which may be further used for validating or developing theories and dynamic models of organizational crises; thereby leading to a better understanding of the underlying causes of, and response to, organization failure.

Keywords: Enron, email corpus, communication networks, social network analysis, dynamic network analysis, organizational crisis, organizational hierarchy

1. Introduction

In a mere fifteen years, the Enron Corporation ("Enron") grew to become the seventh-largest business organization (by revenue) in the USA. By 2001, the company employed 21,000 people in over 40 countries (Fox, 2003; Fusaro and Miller, 2002). Formed from the merger of two local gas-supply companies, Enron and its senior management were aggressively seeking growth and profit. They constructed the first nationwide natural-gas pipeline in the United States, and then promptly transformed the company's core business into global commodity and options trading. They deftly created an exceedingly successful global financial powerhouse from very simple beginnings. By taking this course, Enron quickly became a *darling* of its devoted employees, its unswerving stakeholders, and the broader stock-market community.

Nevertheless, late in the year 2001 the mammoth organization suddenly found itself insolvent, causing senior management to file for Chapter 13 bankruptcy. Financial tragedy, public outcry and scandal quickly followed. Under heavy stakeholder uproar and political pressure, the US Securities and Exchange Commission (SEC) (SEC Spotlight on Enron) and the Federal Energy Regulatory Commission (FERC) (FERC Western Energy Markets—Enron Investigation) conducted simultaneous, albeit independent, inquiries into the sudden collapse.

In May 2002, in an unprecedented action, the FERC publicly released a corpus of actual emails from 158 employees—including those involving top executives such as the company's very-public CEOs, Kenneth Lay and Jeffrey Skilling. The FERC took this unusual step in order to improve the public understanding of the various reasons for their investigation of Enron. The full corpus represents a large collection (about a half-million emails) and temporal record of email conversations over a period of 3.5 years. It is important to note that this corpus contains private communications involving numerous individuals who were not subject to legal investigation. Some emails specific to certain individuals have been removed for privacy and legal reasons.

For researchers focusing on Social Networks, Organizational Theory, and Organizational Behavior, the Enron corpus is alluring and of particular interest with much academic value because it enables the examination of social and organizational processes in a real-world organization over a long period of time. It provides researchers a rare, authentic glimpse into the social network of an actual business organization. The Enron corpus also contains a large amount of raw data on communication, knowledge, relationships, perceptions, resources and events in a company in *crisis*. We believe that scientific analysis of this data will provide information and insight leading to an understanding of the communicative relationship within and among the social and formal networks in this particular organization.

Several studies using the same core email data have already been published (see Section 3). However, the research presented in this paper is uniquely motivated by investigating the relationships among social entities and the dynamics of these relationships as the crisis escalated in the organization. We utilize data mining techniques and apply Dynamic Social Network Analysis to investigate the structure, behavior and idiosyncrasies of communication networks in Enron over time in an explorative way. In general, Social Network Analysis (SNA) is concerned with the relationships between the entities of social and organizational systems (Scott, 2000; Wasserman and Faust, 1994). We consider Enron as an organizational network and the employees as well as the hierarchical levels that constitute the formal organizational structure of Enron—positions and ranks—as entities within that network. By treating positions and ranks as entities in the network, we move beyond SNA on an individual level to analyses based on a level that is immanent to and existing in a wide range of corporations. Further on we refer to the entities we consider—employees, positions and ranks—as agents. Figure 1 shows an organizational chart of Enron that reflects the company's organizational structure, as re-modeled for our analyses, as well as the assignment of specific job titles to more general positions that we defined and further reduced to eight ranks.

This study is grounded on the premise that relations among agents in Enron are represented in the exchange of the emails that are contained in the corpus sample. Following a definition

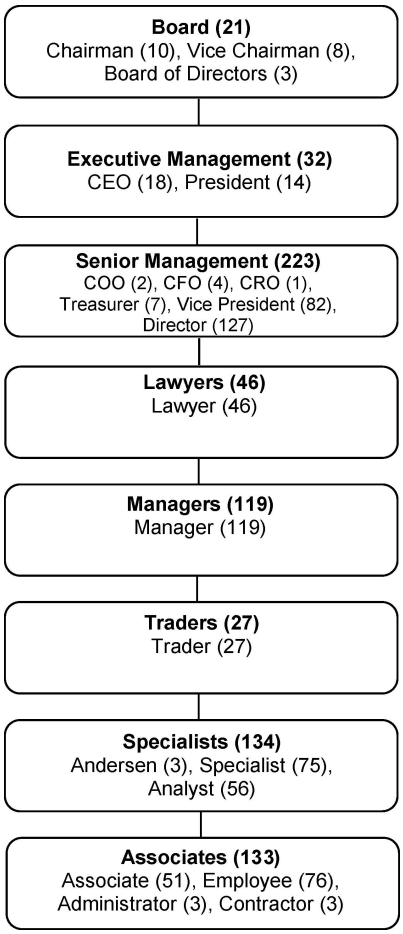

Figure 1. Job title hierarchy with number of individuals per rank and position.

by Monge and Contractor (2003), we refer to this type of network as a *communication network*, which reflects the flow of messages among communicators across space and time. We assume that the relations among and between the agents are reflected in both, the data found in email headers, indicating who had sent messages to whom, and in the email message text. SNA traditionally concentrates on examining the first type (relationships among agents) of information, whereas Natural Language Processing (NLP) is concerned with the second type (text data). In this paper we discuss research on the Enron corpus from both perspectives, point out intersections, and use SNA to analyze the data. The results from our analysis may be of further use for investigating (a) the company's adaptive capabilities to changing situations, (b) points of vulnerability within Enron, (c) indicators of failure, and (d) the modeling of the development of crisis scenarios in complex social systems. The relational data and its analysis could be utilized to further develop theories or validate hypotheses about the dynamics of communication networks.

Section 2 provides a synopsis of the Enron case in order to set up the background on the organization with which this issue is concerned. In Section 3 we give an overview on the research that uses the Enron email corpus and develop our research questions. In Section 4 we introduce the Enron email corpus and report how we refined the database and extracted relational data from it in Section 5. In Section 6 we present our results.

2. The Enron Story

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 (for greater detail on Enron's history, see Fox, 2003; Fusaro and Miller, 2002; Sanborn, 2004). The company, based in Houston, Texas, initially bought wholesale electrical power from generators and sold the electricity to industrial and retail consumers. The company quickly adapted to the deregulation of the energy market by positioning themselves as an energy broker. The management identified specific markets where energy needs where higher than available capacity and quickly built power plants in such regions, then sold the plants before their value diminished, and then pursued to move on to new areas with similar supply and demand imbalances. Later, the management applied their brokerage skills and trader mentality by expanding into new commodity markets such as TV ad time and electronic communications bandwidth.

In 1999, Enron's senior management began to separate and "distance" losses from equity and derivative trades into "special purpose entities" (SPE); these are special legal partnerships that were excluded from the company's primary financial reports. The systematic omission of negative balance sheets from SPE's in Enron's reports resulted in an off-balance-sheet-financing system. The SPE's are just one example of Enron's sweeping controversial ethics deemed as illegal accounting and business practices.

In December 2000, Jeffrey Skilling succeeded Kenneth Lay as CEO. Lay remained Chairman at a time that Enron's stock was trading at a 52-week high of $84.9. Surprisingly, in August 2001, Skilling abruptly resigned and, once again, Lay was named as Enron's COO and CEO. That same month, Sherron Watkins (2002), Enron's Vice-President of Corporate Development, wrote an anonymous letter to Lay in which she accused the company of fraud and significant improprieties such as the SPE's accounting practices. Arthur Andersen, LLP (Andersen), Enron's auditor since 1985—who provided accounting and internal and external consulting services for Enron—allegedly knew the same information contained in Watkin's letter.

In October 2001, the one-month financial loss transferred from Enron's accounting books to the SPE's books totaled to over $618 million. At last, Enron began to publicly report these transactions as net losses and was subsequently obligated to announce that the SEC investigation had discovered $586 million of these losses that had occurred over the previous five years. The stock market responded with an immediate drop and incessant decrease in the price of Enron's shares. Within one month the stock was trading below $1.00 per share, equaling a reduction in company's valuation of $1.2 billion; a financial disaster from any perspective. In December 2001, Enron became insolvent and was forced to file for bankruptcy. The next month, January 2002, Lay resigned from the company.

It was widely alleged that the Andersen auditors knew about Enron's critical financial situation long before Enron went public about the problem, but the auditors did not divulge this information nor advised regulators as required (for details on the relations between Andersen and the Enron case see United States District Court Southern District of Texas, 2002). It was later shown that Enron and Andersen had fraudulently reported hundreds of millions of dollars in increased shareholder equity that were in reality a decrease. In 2000, Andersen's internal management had rated Enron lower then they were evaluating their client publicly. Before Enron publicized its true 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 October 2001, Andersen criminally instructed Enron to destroy any documentation related to the circumstance.

During February 2002, lawsuits against several now-former Enron executives were filed by various parties and Lay ultimately surrendered to the FBI. Lay was accused of several crimes: Participating in conspiracy to manipulate his company's financial results, making false and misleading public statements about Enron's financial situation, omitting information necessary to generate accurate financial statements, civil fraud, and insider trading. Skilling—who stated that he was unaware of any accounting problems—was charged with fraud, conspiracy, filing false statements to auditors, and insider trading. In 2002, Andersen was indicted for altering, destroying and concealing Enron-related material and persuading others to do the same. The company was convicted of obstruction and received a probationary sentence, leading it to being banned from auditing public companies.

In March 2003, Enron announced a plan to emerge from bankruptcy as two separate companies. In July, the company filed a formal reorganization plan stating that most creditors would receive about one-fifth of the $67 billion they were owed while equity shareholders would receive nothing.

3. Researching the Enron Email Corpus

The Enron corpus has primarily been used by the (Statistical) NLP community as a test bed and new benchmark for techniques and algorithms for the automated processing and search of a large-scale collection of textual data. One common practice across the various NLP projects is the automated identification of specific individuals, events, and communication threads and patterns from streaming data that might indicate a threat or risk. These pieces of information are then filtered and investigated in greater detail. In the following brief review of these projects, we point out links between research deploying NLP and SNA.

Separately, Shetty and Adibi (a) and Klimt and Yang (2004a) cleaned the corpus, explored the distribution of messages across users and time, and characterized discussion threads (Klimt and Yang, 2004b). Corrada-Emmanuel performed consistency checks on the emails by using the MD5 digest. Klimt and Yang (2004b) and Bekkerman, McCallum and Huang (2004) built classifiers for predicting the user's organization of emails into user-defined folders.

Berry and Browne (2005) tracked and extracted topics and then clustered messages to identify critical happenings and individuals. Priebe et al. (2005) used scan statistics, which slides a window over equally sized portions of longitudinal data, in order to search

for outlying data points that are considered as corpus-specific deviations from average communication. Network analysts could use agents or data points identified in such ways to study properties of social networks at critical or unusual states.

Keila and Skillicorn (2005) investigated the applicability of structural features of emails such as message length and the usage and frequency of certain words to detect patterns of unusual communication. They were able to identify an Enron-specific vocabulary, relate some employee's formal positions to the use of certain words, and based on this information determined relations among those positions and employees. Their results could serve as a point of comparison for formal and emergent structures identified with SNA techniques. The *communities of practice* they found based on shared *terms of art* (Brown, 1991) could be compared to communities detected with SNA methods such as graph clustering, such as shown by Chapanond, Krishnamoorthy and Yener (2005). Duan et al. (2002) identified the importance of Enron employees based on the amount of emails they sent and received, regardless of their formal position. If the authors would provide the names of the most critical people they found, which they did not do yet for privacy reasons, network analysts could compare those results to the key players identified via SNA techniques.

Beyond the NLP studies, the growing body of research into the Enron corpus includes a few publications that take a network analytic perspective. Shetty and Adibi (a) and Chapanond et al. (2005), generated an undirected social network that represents 151 and 150 Enron employees, respectively. They considered mutual exchange of at least 5 (Sheety and Adibi, a) and 30 (Chapanond et al., 2005) emails between any pair of individuals as a link. The resulting networks do not refer to a specific point in time, but show a consolidated snapshot of the dataset. Shetty and Adibi also took the people's formal positions into account, but did not further analyze the network. Chapanond et al. (2005) report network analytic measures on their graph.

McCallum et al. (2005) combined social network information extracted from sender-recipient relations with information on the topic of emails that they identified by statistical analysis of word distributions into the ART model. They extended the ART model by determining people's roles (RART model) and showed experimentally that this combination of evidence provides a better prediction of similarities among people with the same roles than traditional block modeling.

We suggest that the results from these various projects lead to complementary insights that can be further integrated into additional studies. We believe that the combination of NLP and SNA techniques yield a more holistic understanding of organizational processes in Enron.

Managerial malpractice and a certain organizational culture, two human factors that are assumed to be related to Enron's failure, are as of yet mainly neglected in the scientific analysis of the Enron case, but are discussed by lawyers who investigated the case as well as in the popular business literature. The Enron Board Investigation Committee concluded that Enron's top executives (a) did not communicate critical information to the public, (b) did not provide checks and balances that were designed to ensure ethical business practices, and (c) had established a culture that enabled personal enrichment and encouraged employees to "push the limits" (Powers, 2002). The Management Institute of Paris (MIP, 2002) also explains Enron's failure with errors made by Enron's and Andersen's senior managers rather

than with extra-organizational factors, which are usually claimed by the management. The MIP (2002) identifies a set of reasons for Enron's crash that can be grouped into three categories: Misperception of reality, a risk-taking organizational culture, and improper crisis management. Misperception of reality occurred in Enron on a managerial level when (a) executives ignored bad news since it did not fit in their mental models of success that they had previously built up, (b) managers 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 (e.g. Watkins sending a letter anonymously to Lay). Elements and consequences of Enron's risk-taking culture are a greedy profit taking without disclosure, overdosing of risk, lack of moral leadership and ethics, secrecy, and a winner-take-all mentality. Enron's improper crises management involved the implementation of ad-hoc strategies without thorough analyses of the problem, misleading the public and hoping for a quick solution of the problems. We argue that the analysis of human factors on a managerial level in Enron is an outstanding research issue.

To date, little research applying the SNA perspective has been conducted on the Enron email corpus. Applying a SNA framework to this rich dataset differentiates our research and is the lens through which we address the following questions:

1. What are the dynamics and changes to the properties and structure of the communication networks in Enron over time?
2. What are the agent specific patterns of sender versus receiver characteristics and channeling of information through the organizational structure?

Our research questions are of an explorative nature. Answers to these questions can provide researchers with knowledge that enables a better understanding of Enron's life cycle of success, crisis and bankruptcy. Note that the work presented in this paper is still research in progress; the preliminary results of our study cannot yet be generalized for the Enron corporation or other organizations, but show what knowledge we can gain from analyzing an email corpus from a social network analytic perspective and what kind of questions we can answer using this methodological framework.

4. Data

Each email in the Enron corpus contains the email address of the sender and receiver(s), date, time, subject and email body. Email attachments were not made available. FERC's version of the database had multiple integrity problems, but nonetheless, Kaelbing from MIT purchased the dataset from the regulatory commission. Gervasio and her group at SRI International (SRI International) later prepared the data for the CALO project and corrected several integrity problems. Cohen from CMU put the dataset online for research purposes. The corpus dataset is quite large (400 Mb) and contains 517,431 seperate emails sent by 151 employees. The emails originate from over 4,700 email-folders maintained by employees throughout Enron. Some messages were deleted in response to requests from affected employees (Cohen).

Corrada-Emmanuel from the University of Massachusetts enhanced the dataset by applying the MD5 digest, thus discovering that the corpus actually contains 250,484 unique emails from 149 individuals (Corrada-Emanuel). Researchers at the University of California, Berkeley developed a graphical user interface (GUI) and software for the database that enables powerful search of and retrieval of the data (UC Berkeley).

For this study, we use a version of the Enron corpus dataset provided by Shetty and Adibi (a) from ISI. The researchers cleaned up the dataset by dropping emails that were blank, duplicates of unique emails, junk data, or emails that were returned by the system because of transaction failures. The resulting corpus contains 252,759 emails in 3,000 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 because the cleansing process is well documented and the structure of the MySQL database met the needs for our work. We built a new instance of the ISI Enron database, which we call the *Enron CASOS database*.

5. Extraction of Communication Networks

5.1. Design Choices

In order to properly analyze the Enron corpus using a SNA perspective, we first must extract the relevant *relational* data from the database. The data is implicitly multi-mode (various types of relationships among people such as work relationship, friendship, kinship), multi-link (connections between agents, knowledge, resources etc.) and explicitly multi time period. Nodes and edges may have several important attributes such as the position or location of an employee and the type of a relationship between communication partners (multi-mode). These attributes carry relevant information on how to interpret, evolve, and impact the related nodes and edges. We refer to data that is multi-mode, multi-link, multi time period and enhanced with attributes of nodes and edges as "rich" data.

To set up an infrastructure for adequately representing and analyzing the information contained in the corpus, we need a data format that handles rich social network data and can be used as input and output of multiple analysis tools. We chose the DyNetML format (Tsvetovat, 2003) because it meets these requirements. DyNetML is an XML-based markup specification designed for complex social-network data. A DyNetML document can fully represent an arbitrary number of node sets, edges and their corresponding adjacency matrices, i.e. network graphs.

In contrast to the social networks extracted from the Enron corpus by other researchers (Shetty and Adibi, a; McCallum et al., 2005), we represent the networks as directed graphs (also referred to as digraphs), because email exchange does not necessarily transpire in a strictly reciprocal fashion, but frequently is directed from one agent to multiple others in the form of a broadcast message. Another of our design choices regarding relationship representation—which is another deviation from others' methodology—is the relaxation of the limitation of the number of senders to the 151 people on which the dataset had been collected. Our rational behind this decision is that the 151 people have emails in their inboxes received from individuals who are not a member of the sample. Furthermore, in accordance

with other social network analyses on Enron, we decided to assign a weight to each edge that reflects the cumulative frequency of emails exchanged between the corresponding pair of communicators.

5.2. Data Enhancement

We strengthened the power of the Enron corpus sample by adding important information to the database in a three-step enhancement process. We will refer to this enhanced database as the *Enron CASOS Generation II* data. As the first step, we added the full names of 525 previously unaccounted employees of Enron and Andersen. Next, recognizing that people often have multiple valid email addresses, we increased the number of email-address assignments directly mapped to individuals from 1.0 to an average of 2.2. We achieved this by applying sophisticated text matching techniques to the unmatched email addresses and individuals' name(s) in the database. By increasing the *address-assignment ratio* (the number of valid email addresses matched to an individual) we effectively improved the "personalization" of the data so that extracted communication networks are linking *people* rather than merely linking *email addresses*. In the third step of the enhancement process, we added a career history for 676 individuals. We assigned a rank to each of the unique job titles in order to track people's job positions and their career progression through the organizational hierarchy. With the enhanced and categorized job information we can move beyond analyses that are based on individuals or groups that were identified by clustering techniques to analyses based on the actual organizational structure.

As the basis for these enhancements we gathered and systematically reviewed data pertaining to the hundreds of employees in the database that was obtainable from several public sources. We collected published documents from FERC (FERC/Aspen) consisting of employee's names, jobs, locations, names of their supervisors, business units and trade relations. We also utilized a data file generated by ISI (Shetty and Adibi, b) from Federal Court documents with individual's job titles (ISI position file),[2] gleaned Enron's annual reports, books on Enron (Fox, 2003; Fusaro, 2002) and pieces of data from media coverage. Previously, we reported on a dataset that contained 227 people on whom we had job and location information (Diesner and Carley, 2005a). The data at this stage, which we refer to as *Enron CASOS Generation I*, was integrated into the Enron CASOS Generation II data and further enhanced with information on 449 more people.

During the process of integrating information from various sources, we encountered a plethora of conflicting information with consistency issues which we resolved: The original ISI database contains personal data for 151 people, consisting of their full name and a single email address. In order to match and merge various spellings of names we used a semantic similarity algorithm (Ukkonen, 1985; Meyers, 1986) implemented in the String Similarity Perl module (Lehmann). 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 confirmed, or rejected, each manually.

As the second step to upgrade the existing Enron corpus data, we increased the assignment ratio of email-address to employees. The ISI Enron database associates each individual with

one email address; however, we posit that people commonly have and are truly affiliated with more than one email address. Corrada-Emmanuel provides two email addresses per person for 31 individuals in the Enron corpus. To further increase the address-assignment ratio, we explored this issue in Enron CASOS Generation I by using the similarity function described to search for all email addresses ending with @enron.com for addresses similar to those specified in the employee list table of the database. The module identified the 25 highest scoring hits per address, which we manually vetted. We found that a similarity greater than 0.7 usually indicates a match and selected these by default prior to review (for statistics on the outcome of this process see Diesner and Carley, 2005a, Table 2). Moving from Enron CASOS Generation I to II we dropped 1 email address of 1 person since it had been mismatched. In Enron CASOS Generation II, due to technical constraints, we applied a different process whereby we automatically suggested an assignment for yet unassigned addressees containing @enron.com or containing andersen somewhere after the @ to names that exactly mirror or logically reflect an employee's name. The suggestions were manually evaluated and either confirmed or rejected. The Enron CASOS database contains 252,759 emails and 2,019,847 instances of email addresses ending with @enron.com or containing andersen after the @. Instances represent the cumulative frequency of the occurrence of each unique email address in any of the email header fields (from, to, cc, bcc). Overall, using the described routines, we managed to:

- Increase the number of people associated with at least one email address from 151 in the original database to 557.
- Increase the total number of email addresses assigned to specific individuals from 151 to 1,234, ranging from 1 to 17 addresses per person, with an average of 2.2 and a standard deviation of 1.9.
- Increase the number of emails that have both a sender and at least one receiver associated with persons whose full name and job title is contained in the database (*known individuals*) from 21,254 to 52,866.
- Assign 24,825 of the emails (9.8%) in the Enron CASOS database and 797,569 instances of email addresses (39.5%) to known individuals.

In order to further automate the process of dissolving ambiguities of email addresses, machine learning approaches such as Malin's (2005) for unsupervised learning could be applied. However, that is a point of further research.

Another issue that complicated data cleaning was the simultaneous and multiple job titles identified for individuals. We refer to the initial set of position information on people as *job data.* As a first step we disambiguated the job data by collecting time information on the positions a person held. Based on this information, we built a *career history* for each person. The career history specifies the time range(s) for a person's position(s). In order to eliminate and avoid ambiguities in the network data with respect to multiple positions per time for a person we normalized cases with various positions per person at one time. Our default was to pick the higher position (for a comparison of the positions given in the ISI file and in the Enron CASOS database see Diesner and Carley, 2005a). As a result,

a person's career history contains one job title per time span. Note that the time span is an individual feature that can vary from person to person. The number of employees per position and rank can vary between time spans, depending on an individual's career history. Overall we identified 110 unique job titles that we associated with 676 unique employees. Figure 1 shows the assignment of the 110 specific positions to 13 more general positions that we defined and further reduced to 8 ranks. In that chart the sum of individuals across all ranks is 739, which exceeds the number people that we posses position information on (676). This is because several people had multiple positions.

In the following results section we show how the additions to the database allow for richer SNA. Instead of analyzing Enron's communications network merely from the perspective of *email addresses,* we are now in the position to analyze the social network from the perspective of specific *persons* communicating as well as from the perspective of their career histories in the organization.

5.3. Extraction of Communication Networks

We extracted communication networks in DyNetML format in which agents are represented as nodes and the exchange of email instances between them as edges. Each edge represents a directed relation from one sender to one recipient as captured in the email data. Recipients were retrieved from the *to*, *from*, *cc* and *bcc* fields in the emails, with the latter three fields representing an email recipient without further distinction. Edges reflect contacts, meaning that if one agent sent one email to three other agents, three edges would be formed. The edges are weighted by the cumulative frequency of emails exchanged between any pair of agents in a particular time range.

Email addresses that do not refer to actual individuals, e.g., announcements.enron@enron.com and group.enron@enron.com, were removed from the dataset. This includes removal of the address "pete.davis.@enron.com", which is known to have served as a proxy for automatically generated broadcast emails by Enron processes.

In order to enable longitudinal analysis we time sliced our data[3] on a calendar-month basis between October 1998 and July 2002, as this seemed to entail time spans in which major events happened in Enron. Since the number of emails and people involved in email communication were very low for the months before May 1999 and after March 2002, we decided to disregard these time periods in further analyses. The remaining 35 months constitute the time range of our sample.

The Enron CASOS Generation II database contains 119 individuals who are associated with a job title but no email address. This is either a result of the collection of additional data, which had not been limited to people on who we had an email address for, or due to the fact that those people's names were not yet matched to an email address. These people were disregarded in the communication networks.

For our sample we initially considered 557 people for whom we at least had a full name, email address and career history. Email communication exclusively to or from addresses outside the Enron or Andersen domains were removed from the sample. Of the 557 people we studied, 535 were directly involved in at least one email exchange within the sample, which defines the sample size. No board member is included in that sample. The number

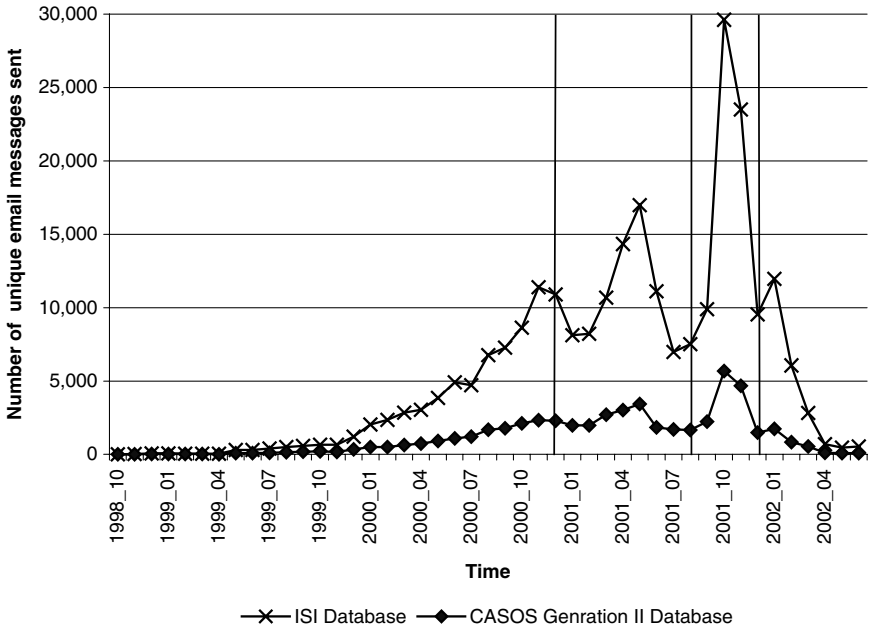

Figure 2. Comparison of number of sent emails in original and study's database.

of people participating in email communication varies between 16 and 412 (figure 3) per month.

Data analyses can be performed based on three different definitions of a sample: (1) the size of the sample could be kept static (535 people) across all time points. With this approach the 22 people who were not involved in any email exchange throughout all time points would be kept in the sample. An argument supporting this strategy is that those people had the chance to communicate with others in the sample, but we did not find evidence of such communication in the Enron CASOS database. (2) In contrast one could argue that people who were not involved in any email exchange throughout all time points should be removed from the network data because they did not impact the network or network analytic measures. In this case the number of agents would still remain constant (513) across all time periods. (3) Extending the argument made in strategy 2, all people who were not engaged in email communication at a certain time period (isolates) could be excluded from the communication networks. With this strategy, the size of the sample potentially varies from one time point to the next. We decided to use the third approach because the number of isolates exceeded the number of people participating in communication in 17 out of the 35 months that we considered for analyses (figure 2) and therefore significantly impacted network analytic measures. The strategy we chose facilitates analyses based on agents who were part of the actual communication flow as opposed to analysis of a social network where some agents were engaged in communication and others were not.

Figure 2 shows the total number of emails sent by all people in the corpus and in our sample. Given the resemblance of the curve for the entire corpus by the curve for our sample we believe that our sample is a representative subset of the corpus with respect to the frequency distribution of sent emails. The peaks in the amount of communication can be partially related to events in the organization: The peaks in the curve occur in October 2001, the month in which the Enron crisis fully broke out and November 2001, when the investigations were under way. The low points around the end of the year and during July and August might be explained as being vacation periods. The three vertical lines in figures 2 and the following figures mark key events in the organization: In December 2000 Skilling took over the position of CEO from Lay. In August 2001 Skilling resigned and Lay was named as COO and CEO again. In December 2001 Enron filed for bankruptcy.

6. Analysis and Results

To explore the dynamics and changes of the properties and structure of Enron's communication network (research question 1), we analyzed the data at various levels according to our definitions of agents: By employees (based on individual's job titles), by positions, and by ranks (with the latter two levels being based on the rank of a job in the corporate hierarchy). We used the ORA statistical software, version 1.5.5 (Carley and Reminga, 2004), to calculate the network measures used throughout the study. ORA is a dynamic network analysis toolkit for assessing and comparing complex network data that changes over time (Carley, 2003; Carley et al., forthcoming). We focus our discussion on the time periods during which Enron faced its major crisis.

The extent to which all agents in a network are connected to each other can be measured as density. Density ranges from 0 to 1, with higher values indicating more interconnections of agents (for details on network analytic measures see Carley and Reminga, 2004; Wasserman and Faust, 2004). The densities of the employees' networks are very low (ranging from 0.01 to 0.05) and lower than the densities of the rank's networks (ranging from 0.32 up to 0.90) (figure 4). At either level, however, the patterns of evolution in the density do not appear to coincide with the Enron scandal. In contrast, when viewed by position, density rose as the scandal escalated. An exception can be found in November 2001, where the cohesion across positions strengthened considerably. It may be the case that at the individual level the density measure is too finely grained and the rank level too condensed to capture this trend.

The communicative reachability of individuals in an organization can be measured by examining the number of components in the communication network. Components are maximally-connected subsets of a graph in which each node can communicate with any other node either directly or by passing messages through others. The existence of more than one component in a network indicates that some actors are severely disconnected from others. Given that we consider the directionality of a link, we focus on *strong* components rather than the *weak* components measure, which does not consider directionality.

In August 2001, when the Enron crisis surfaced, the number of strong components on the employee level reached its peak, and the number of people involved in email exchange also increased (figure 3). This suggests that during the breakout of the crisis the interpersonal communication had intensified and spread throughout the network in a way that seemingly

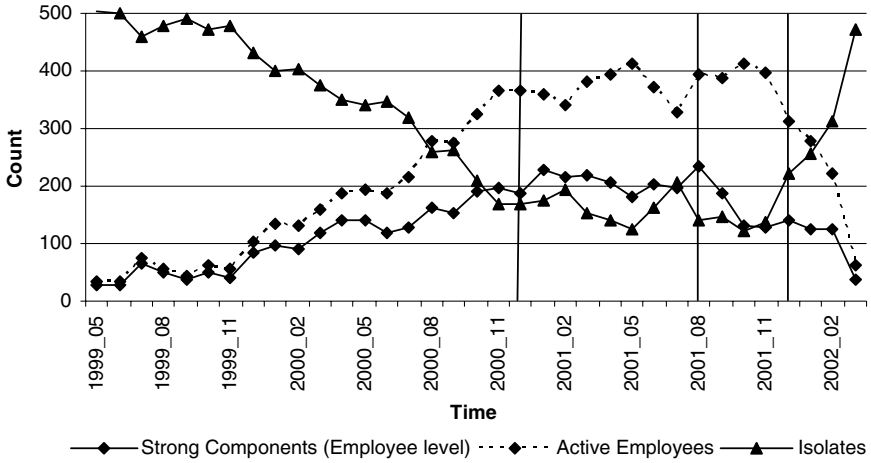

Figure 3. Temporal summary of employee-level network: Size of strong component, active actors and isolates.

fractured the organization into distinct sub-groups. Over the subsequent two months, the number of components dropped stronger than ever before, while the number of communication participants changed less and without a clear trend. This finding indicates that during the escalation of the crisis, previously disconnected people began to engage in mutual communication, thus strengthening the cohesion of the system. Looking at positions, we do not observe a clear trend regarding components. For ranks, there is only one strong component from April 2001 on, indicating that mutual communication occurred within and across all ranks.

The degree to which single agents have high importance in a network while others are of low prominence can provide information about the disparity of a network. This inequality is represented in centralization measures, computed on graph level. Betweenness centralization captures an agent's control over communication flow by measuring how often an agent is positioned as the connection point on the shortest path between any pair of non directly adjacent agents. Degree centralization is an indicator for an agent's activity by counting how often an agent is directly linked to others. Since the Enron networks are directed, we split up degree centralization into outgoing (outdegree) and incoming (indegree) links. Mathematically, degree and betweenness centralization range from 0 to 1, with higher values indicating the stronger inequality of importance in a network. As can be seen in figure 4, during the crisis the disparity in employee's communication control raised and reached its peak in December 2001, the month that Enron filed for bankruptcy. Regarding the inequalities in communication activity, we observe on the employee and position level that the disparities in distributing information are greater than for receiving information (figure 5). In the months preceding the crisis the disparity in sending activity jumped up, and subsequently fell until November 2001; further supporting the insight that during the crisis the agent's established roles and contacts got altered in a way that resulted in a more diverse network.

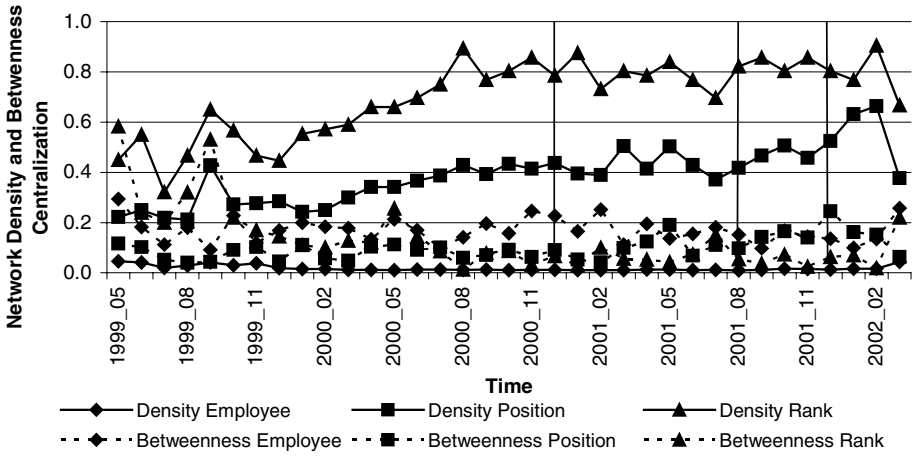

Figure 4. Network density and betweenness over time by level of analysis.

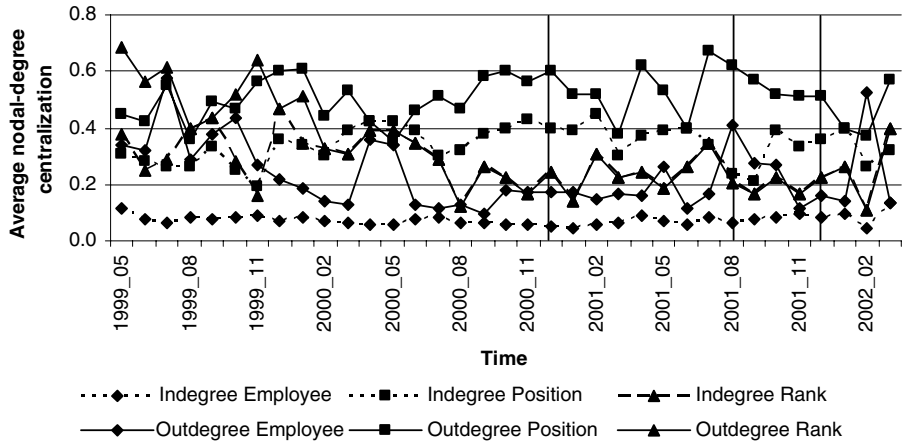

Figure 5. Average nodal-degree density over time by exchange relationship and level of analysis.

Analyzing the mutual communication in small groups of two to three employees reveals that during the crisis the mutual communication between pairs of employees slightly increased (figure 6). This suggests a movement toward communication only among trusted others, possibly due to accountability. In contrast, the frequency of cycles of communication among triads decreased (figure 6), possibly due to a decrease in communication to support tasks, which is more common in a chain or triad.

To explore the communication patterns within the formal organizational structure in more detail, we investigated sender versus receiver characteristics and ways of channeling information through formal hierarchies (research question 2). Those analyses are performed

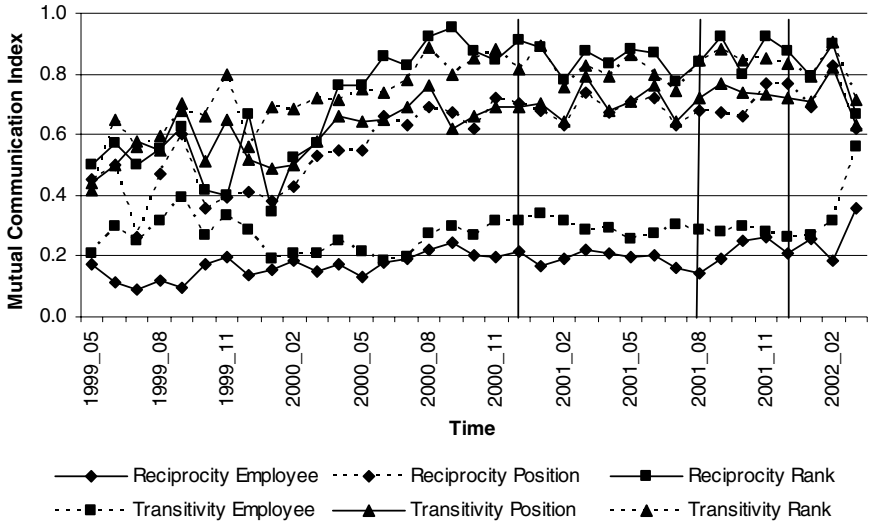

Figure 6. Mutual communication over time among dyads (reciprocity) and triads (transitivity) by level of analysis.

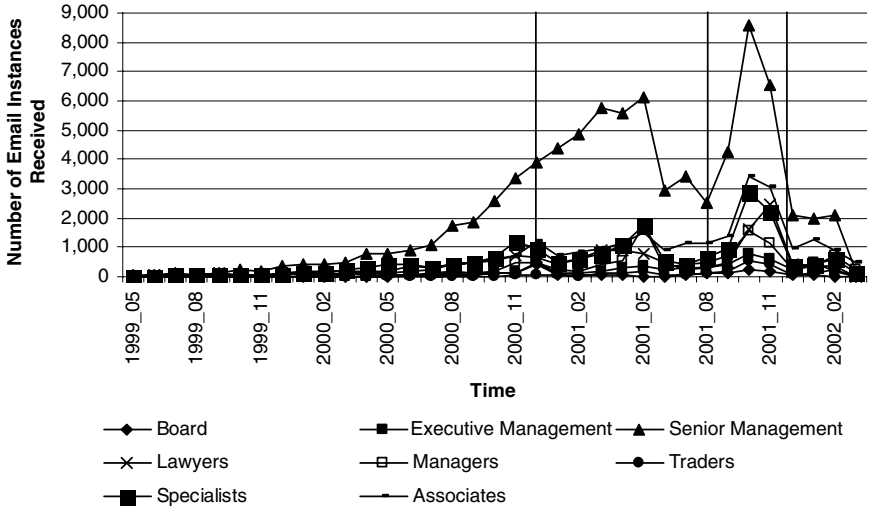

Figure 7. Number of email instances received per rank.

on the rank level. The distribution of the number of *received* email instances shows similar patterns for all ranks, with the Senior Management being contacted more often than any other rank (figure 7). In contrast, the number of *sent* emails is volatile and indicates that different ranks apply different communication patterns (figure 8). The Senior Management and the Associates were the most active message senders. The distribution for both sent and

received email shows a peak in the frequency of communication between September and November 2001, which corresponds to the period in which Enron weakened from being a corporation in organizational crisis to being one in legal bankruptcy. Figure 8 shows that during the crisis period, employees of all ranks engaged in more outgoing communication relative to the prior pre-crisis period, which might be explained by an increased motivation for information seeking and knowledge sharing, perhaps in order to reduce uncertainty. From the end of 2000 on—when Skilling became Chairman—until the middle of 2001, a second, yet lower, peak in the number of incoming emails of the Senior Management occurred, as it did for outgoing emails for the Senior and Executive Management, the Associates and the Lawyers. This indicates that between the time of Enron's first-ever change in CEO (and COO) and the time just prior to the breakout of the crisis, several ranks, including counsels and upper management, were motivated to increase their communication.

Figures 9–12 show the ratio between the number of email instances that each rank had sent and received per month. This ratio is normalized to range from 0 to 1. A value of 0.5 indicates a balance between outgoing and incoming emails, values lower than 0.5 represent *receivers* and values higher than 0.5 *senders*. In August 2001, when Skilling abruptly resigned, leading to Lay's reappointment, the Board showed strong sender characteristics, and became a heavy receiver in the following months. This may be a function of explaining the resignation or requesting it to respond to queries about the implications of the resignation. The ratio for the Executive Management is still volatile, but less so than for the Board. Both ranks show a common and significant peak in May 2001, indicating that in this month both ranks contacted others more often than ever before. The Senior Management, who since March 2000 had stronger characteristics of a receiver than of a sender, during the crisis had a balanced ratio between incoming and outgoing information. Mirroring the curve for the

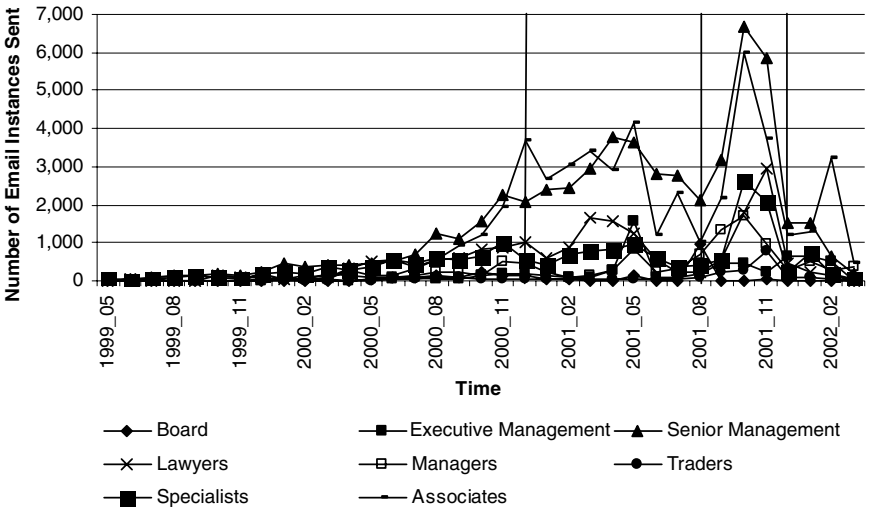

Figure 8. Number of email instances sent per rank.

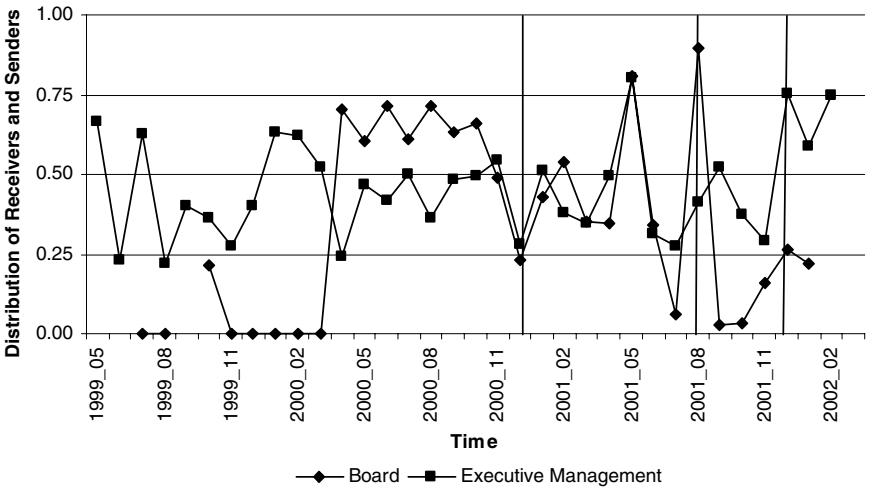

Figure 9. Ratio between sender and receiver characteristics for Board and Executive Management.

Figure 10. Ratio between sender and receiver characteristics for Senior Management and Lawyers.

Senior Management, the Lawyers show more active communication behavior throughout the crisis. Since March 2000, they were more a point of contact for others than they themselves distributed information. The Senior Management and the Lawyers had the most balanced ratios of incoming and outgoing contacts across time and ranks. The Managers and Traders were on a path of changing from being a sender to being a receiver, but starting from June 2001 both of them became more active again. The Specialists were mainly moderate senders. The Associates, bouncing around a balance of outgoing and incoming information in moderate, reoccurring waves, expose no clear pattern.

Figure 11. Ratio between sender and receiver characteristics for Managers and Traders.

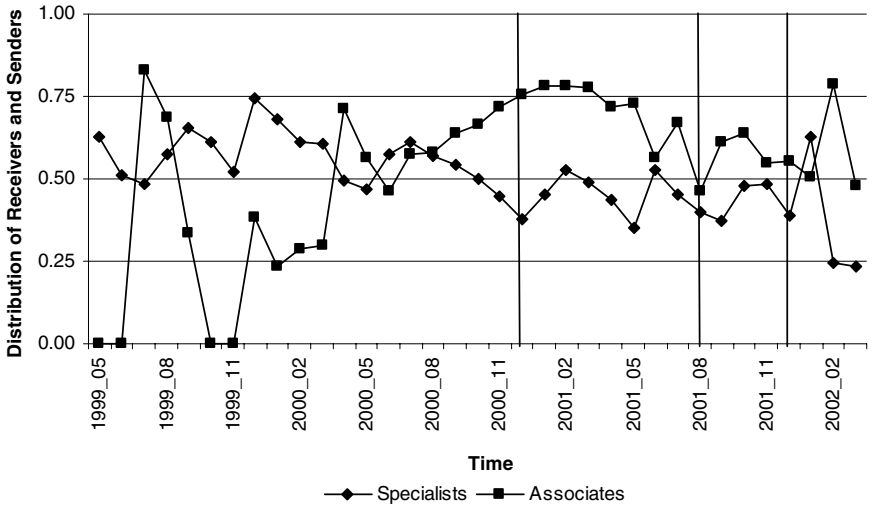

Figure 12. Ratio between sender and receiver characteristics for Specialists and Associates.

In order to analyze the hierarchical communication among and between ranks in more detail we investigated, for each rank, the ratio between email instances that had been sent to higher, equal and lower ranks (figures 13–20). In general and in agreement with what one would expect, we observe that higher ranks typically performed more downwards communication, thus fulfilling their directive roles. In contrast, lower ranks sent out more upward communication, as their responsibility of reporting to their supervising-manager imposes.

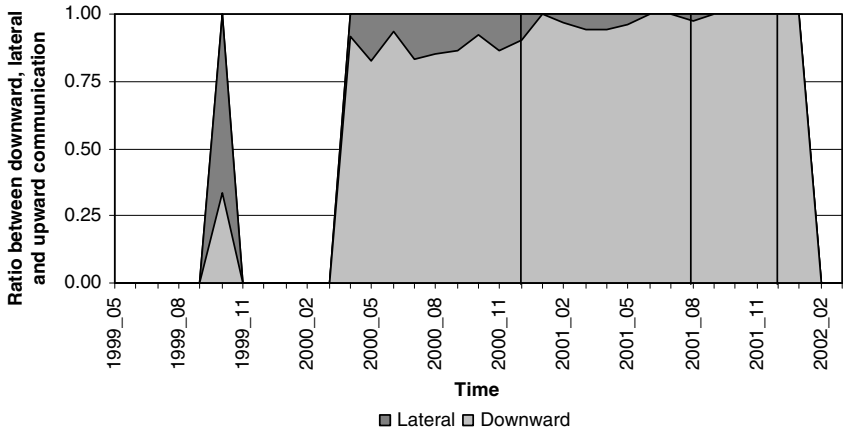

Figure 13. Ratio between downward, lateral and upward communication for Board.

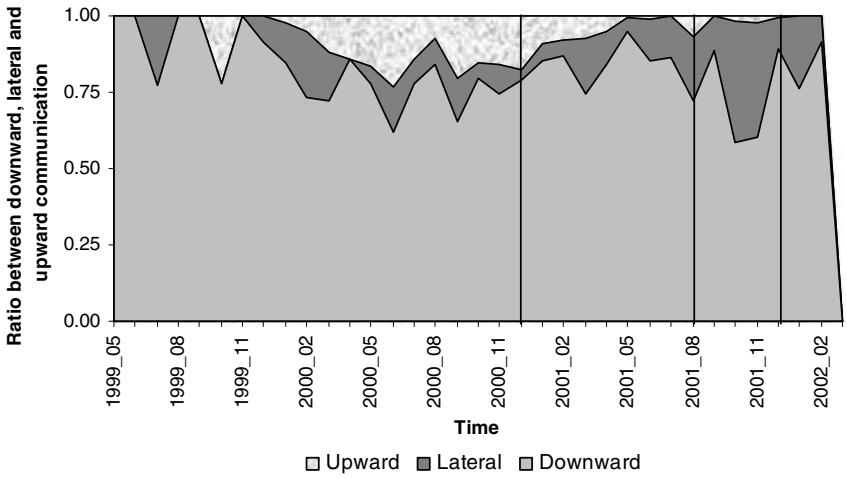

Figure 14. Ratio between downward, lateral and upward communication for Executive Management.

Looking at the different ranks in detail we notice some interesting peculiarities: The Board during the crisis diminished communication within this rank. In contrast, the Executive Management as well as the Lawyers intensified their lateral contacts and decreased upward reporting throughout the crisis. We thus learn that those two ranks throughout the crisis enhanced their intra group relations. These patterns may reflect the need to move communication to a style that reflects concerns with legal issues.

The Senior Management is the group that performed the most lateral communication (about 50% of their emails) across time, indicating that this rank always had a culture of strong interconnections. Even through the crisis, when the Senior Management increased

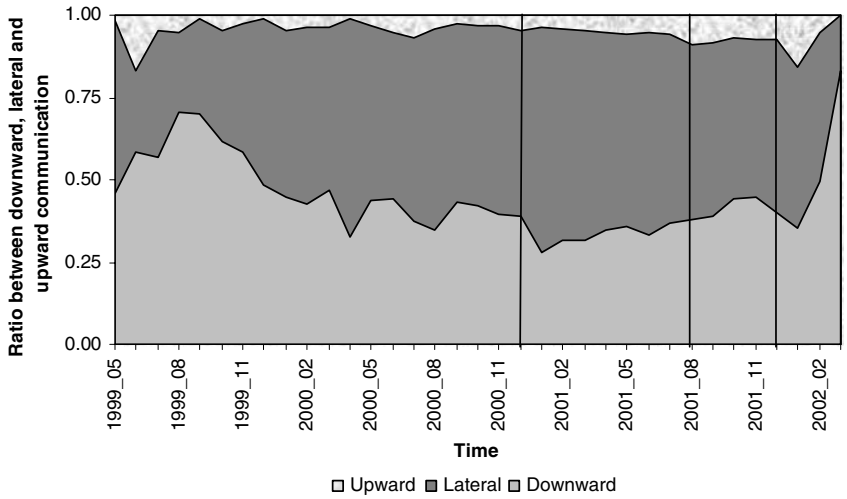

Figure 15. Ratio between downward, lateral and upward communication for Senior Management.

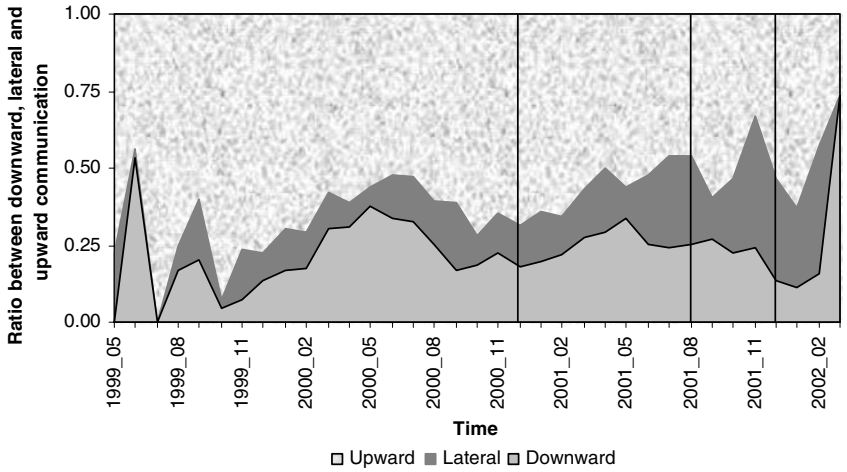

Figure 16. Ratio between downward, lateral and upward communication for Lawyers.

their information sharing with both, higher and lower ranks, they still distributed most information within their rank. The Managers typically sent most of their emails to lower ranks, and sent about equally to lateral and upward ranks. During the crisis, the Managers increased their communication to higher ranks on costs of contacts within this rank. The Traders show the most dramatic changes in their hierarchical communication pattern across time. While they at the outset almost exclusively contacted subordinates, from March 2001 on they switched to a balanced upward and downward communication. Furthermore, they expose the lowest amount of lateral contacts, which might be explained with the fact that

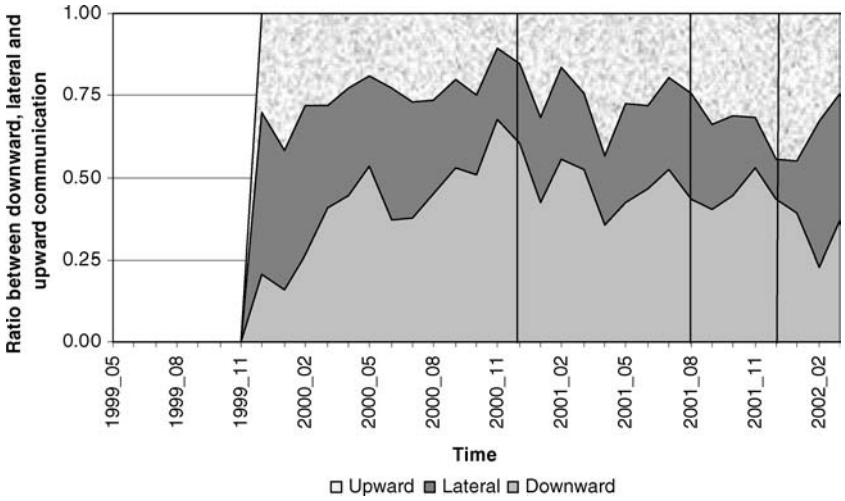

Figure 17. Ratio between downward, lateral and upward communication for Managers.

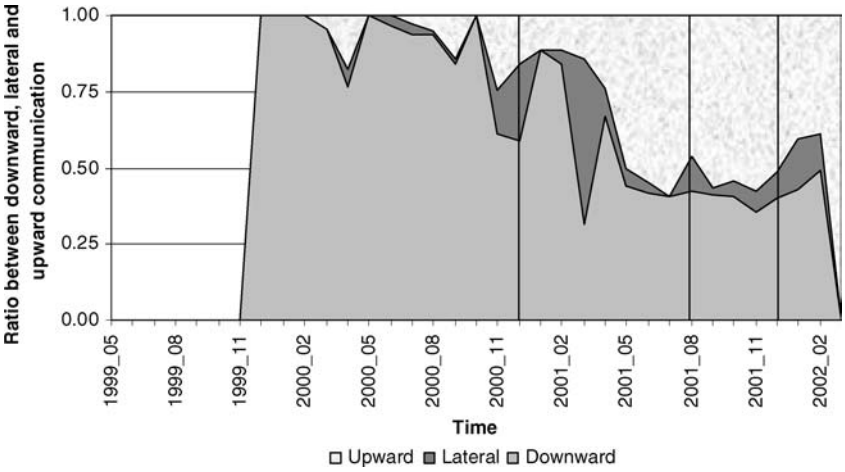

Figure 18. Ratio between downward, lateral and upward communication for Traders.

most of the Traders were located in Houston and therefore possibly talked face to face to each other, or used phones for coordinating power trading related tasks. While the Specialists basically show unchanged curves across time, the Associates clearly started talking to each other more often and almost as much as they did with higher ranks.

Summarizing the insights gained about patterns in the communication behavior of the ranks, we conclude that during the crisis higher ranks tended to be directive and the communication had been more diverse with respect to formal positions than during a normal month. We also see changes in communication style that reflect increased concerns with

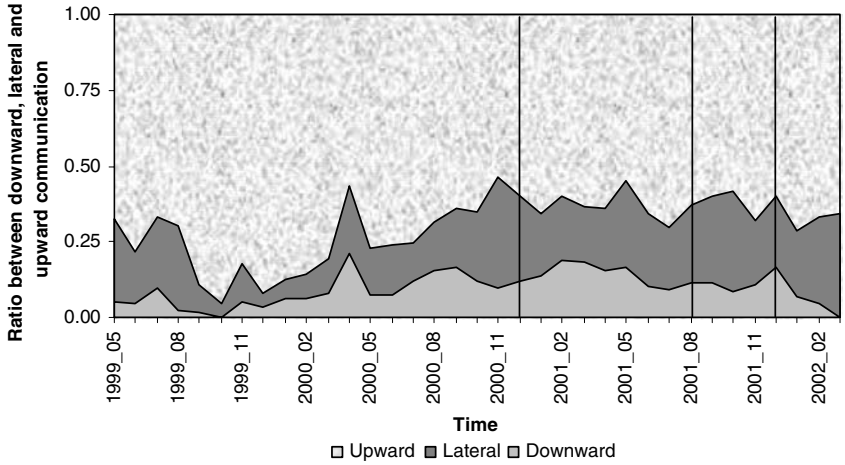

Figure 19. Ratio between downward, lateral and upward communication for Specialists.

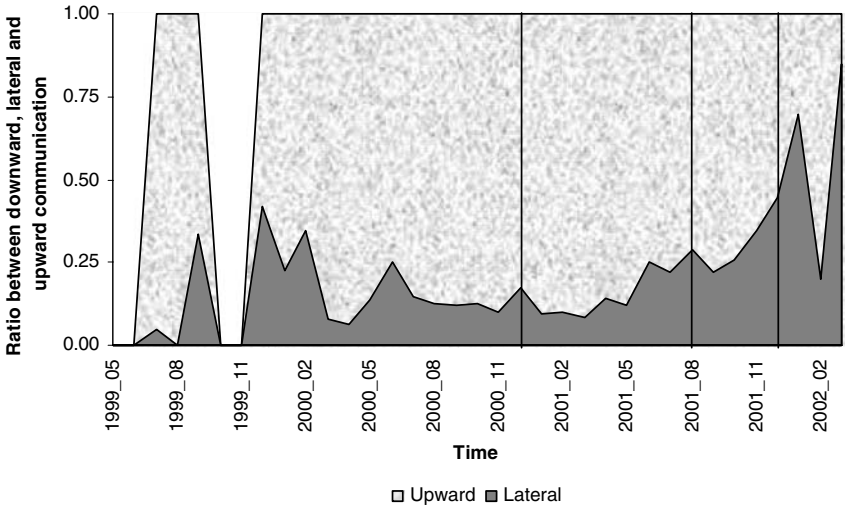

Figure 20. Ratio between downward, lateral and upward communication for Associates.

legal issues, possibly due to a realization that correspondence could be used in courts of law, and increased correspondence with lawyers perhaps to discuss legal matters.

7. Discussion

Clearly, we investigated the email communications of only a small fraction of the 21,000 Enron employees. This might bias the graph analytic results we presented. To overcome

this shortcoming we included all Enron employees who are contained in our instance of the Enron corpus, whose full name we know and on who we performed email address normalization in graph generation. People on who we do not have position information yet were considered as employees by default.

Another limitation of our methodology is that we could not confirm the relations that were extracted. In order to perform validation we plan to 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).

8. Conclusion and Future Work

We have described how the Enron email corpus database was enhanced and refined for this study, and how rich communication-network data was extracted from it, leading to two main conclusions about the corpus as a resource for research. First, it is difficult and time consuming to clean this real-world email data and put it into a form that is analyzable with SNA or DNA tools. Better automated and semi-automated techniques are needed. It is debatable whether we would have gleaned the similar results with data that was not so purposefully cleaned. While we have not systematically investigated this question, we believe it is likely that the full set of data in the corpus would reflect similar over-time trends. Regardless, cleaning and normalizing the database is necessary to extract subtle peculiarities of communication on all levels that we investigated.

The second point is theoretical. We find substantial evidence that organizational communication, even the sample that exists in email, changes not only in volume but in who is communicating with whom during periods of organizational change and crisis. Gross level changes reflect the changing role and importance of specialized functions such as the legal personnel. To fully understand the role of these functions, it is important to look at the over time aspects of the agent's interactions with the rest of the organization. Gross level changes in communication also reflect changes in communication norms and the way in which groups present themselves. Over time, the communication structures appear to reflect changing environmental, legal, and social conditions in which business is done.

We have shown that SNA and the analysis of the frequency and direction of email communication reveal communication patterns that correspond with the organizational life span. We have learned that during a crisis period: (a) communication among employees becomes more diverse with respect to established contacts and formal roles, (b) previously disconnected employees begin to engage in mutual communication, and (c) interpersonal communication intensifies and spreads more widely throughout the network. We also showed how network analytic measures are sensitive to different definitions of agents.

However, these investigative techniques do not enable us to evaluate the context and intent of individual messages, namely communications for the purpose of information seeking, clarification, understanding details, seeking advice, or cover-up. In the future, we intend to compare and combine the findings herein with an analysis of the email content by applying

Network Text Analysis (NTA) techniques (Popping, 2000; Diesner and Carley, 2005b). NTA enables the exploration of the perception of the company's situation across time on an individual and organizational level. We will extract these perceptions as mental models, which are representations of the reality that people use to make sense of their surroundings (Johnson-Laird, 1983; Rouse and Morris, 1986). 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 (Klimoski and Mohammed, 1994; Carley and Palmquist, 1992). Since mental models can and have also been used to study culture at the team (Carley, 1997), organization (Monge and Contractor, 2003), and community (Carley and Palmquist, 1997) level we plan to verify assumptions about Enron's culture such as the features discussed in Section 3 via the analysis of mental models extracted from the content of emails.

For Enron, and possibly for any organization, a change in the usual pattern of communication accompanies major corporate events. There may be a statistical signature to the email communication patterns in an organization in the midst of a crisis. Whether this is in fact true will require more real-world email datasets to be examined.

Acknowledgments

This paper is part of the Dynamics Networks project in CASOS (Center for Computational Analysis of Social and Organizational Systems) at Carnegie Mellon University. This technology described herein 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 and analysis was provided by the DOD and the NSF under the MKIDS project, IIS-0218466-NSF and the NSF under the IGERT program in CASOS. 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 ONR, DOD, 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. An earlier version of portions of this paper was presented at the Workshop on Link Analysis, Counterterrorism and Security, held at the SIAM Data Mining Conference in Newport Beach on April 23rd, 2005.

Notes

1. Alex Gibney in an interview conducted by with Mark Leibovich for The Washington Post: In Enron, Filmmaker Saw 'Ultimate Morality Tale'. April 29, 2005; C01.
2. The ISI position file lists the names of 161 Enron employees, and for 132 of them it provides position information 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 only a short period of time. In the social network generated by Shetty and Adibi those 29 people are assigned to the position of an employee.
3. 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.

References

Bekkerman, R., A. McCallum, and G. Huang (2004), "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora," CIIR Technical Report IR-418 2004. Retrieved June 16, 2004, from http://www.cs.umass.edu/∼ronb/papers/email.pdf

Berry, M.W. and M. Browne (2005), "Email Surveillance Using Nonnegative Matrix Factorization," *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, SIAM International Conference on Data Mining 2005. Newport Beach, CA, April 2005, 45–54.

Borgatti, S.P. (2004), "The Key Player Problem," in R. Breiger, K. M. Carley, and P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis: 2002 Workshop Summary and Papers*. Washington, DC: National Academies Press, 241–252.

Brown, J. S. and P. Duguid (1991), "Organization Learning and Communities-of-Practice: Toward a Unified View of Working, Learning, and Innovation," *Organization Science*, 2(1), 40–57.

Carley, K.M. (2003), "Dynamic Network Analysis," in R. Breiger, K.M. Carley, and P. Pattison (Eds.), *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, Committee on Human Factors, National Research Council, National Research Council. Washington, DC, 133–145.

Carley, K.M., J. Diesner, J. Reminga, and M. Tsvetovat (forthcoming), "Toward an Interoperable Dynamic Network Analysis Toolkit," *Decision Support Systems Journal*, Special Issue on Cyberinfrastructure for Homeland Security: Advances in Information Sharing, Data Mining, and Collaboration Systems.

Carley, K.M. and J. Reminga (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

Carley, Kathleen M. (1997), "Extracting Team Mental Models Through Textual Analysis," *Journal of Organizational Behavior*, 18, 533–538.

Carley, Kathleen M. and M. Palmquist (1992), "Extracting, Representing and Analyzing Mental Models," *Social Forces*, 70(3), 601–636.

Chapanond, A., M.S. Krishnamoorthy, and B. Yener (2005), "Graph Theoretic and Spectral Analysis of Enron Email Data," *Proceedings of Workshop on Link Analysis, Counterterrorism and Security*, *SIAM International Conference on Data Mining 2005*. Newport Beach, CA, April 2005, 15–22.

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

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

Diesner, J. and K.M. Carley (2005a), "Exploration of Communication Networks from the Enron Email Corpus," *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*. Newport Beach, CA, April 2005, 3–14.

Diesner, J. and K.M. Carley (2005b), "Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis," Chapter 4 in V.K. Narayanan and D.J. Armstrong (Eds.), *Causal Mapping for Information Systems and Technology Research: Approaches, Advances, and Illustrations*. Harrisburg, PA: Idea Group Publishing, pp. 81–108.

Duan, Y., J. Wang, M. Kam, and J. Canny (2002), "A Secure Online Algorithm for Link Analysis on Weighted Graph," *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*. Newport Beach, CA, April 2005, 71–81.

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.

Ferc/Apsen web site. (n.d.), Retrieved between November 4, 2004 and June 22, 2005, from http://ferc.aspensys.com.

Fox, L. (2003), *Enron. The Raise and Fall*. Wiley & Sons: Hoboken, N.J.

Fusaro, P.C. and R.M. Miller (2002), *What Went Wrong at Enron*. Wiley & Sons: Hoboken, N.J.

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

Keila, P.S. and D.B. Skillicorn (2005), "Structure in the Enron Email Dataset," *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*. Newport Beach, CA, April 2005, 55–64.

Klimoski, R. and S. Mohammed (1994), "Team Mental Model: Construct or Metaphor?" *Journal of Management* 20, 403–437.

Klimt, B. and Y. Yang (2004, a), "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

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

Lehmann, M. (n.d.), "String Similarity," Retrieved from http://search.cpan.org/∼mlehmann/String-Similarity-1/Similarity.pm.

Malin, Bradley (2005), "Unsupervised Name Disambiguation via Social Network Similarity," *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*. Newport Beach, CA, April 2005, 93–102.

McCallum, A., A. Corrada-Emmanuel, and X. Wang (2005), "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks, with Application to Enron and Academic Email," *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*. Newport Beach, CA, April 2005, 33–44.

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

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

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

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

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

Priebe, C.E., J.M. Conroy, D.J. Marchette, and Y. Park (2005), "Scan Statistics on Enron Graphs," *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining 2005*. Newport Beach, CA, April 2005, 23–32.

Rouse, W.B. and N.M. Morris (1986), "On Looking into the Black Box; Prospects and Limits in the Search for Mental Models," *Psychological Bulletin*, 100, 349–363.

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

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

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

Watkins, S. (2002), *eMail to Enron Chairman Kenneth Lay*, Retrieved November 2, 2004 from www.itmweb.com/f012002.htm

Shetty, J. and J. Adibi (n.d., a), "Ex employee status report," Retrieved November 4, 2004, from http://www.isi.edu/∼adibi/Enron/Enron_Employee_Status.xls

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

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

Tsvetovat, M., J. Reminga, and K.M. Carley (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

UC Berkeley (n.d), Enron Email Analysis Project. Retrieved from http://bailando.sims.berkeley.edu/enron_email.html

Ukkonen, E. (1985), "Algorithms for Approximate String Matching," *Information and Control*, 64, 100–118.

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

Wasserman, S. and K. Faust (1994), *Social Network Analysis*. New York: Cambridge University Press.

Jana Diesner is a Research Associate and Linguistic Programmer at the Center for Computational Analysis of Social and Organizational Systems at the School of Computer Science (CASOS), Carnegie Mellon University (CMU). She received her Masters in Communications from Dresden University of Technology in 2003. She had

been a research scholar at the Institute for Complex Engineered System at CMU in 2001 and 2002. Her research combines computational linguistics, social network analysis and computational organization theory.

Terrill L. Frantz is a post-doc researcher at the Center for Computational Analysis of Social and Organizational Systems (CASOS) in the School of Computer Science at Carnegie Mellon University. His research involves studying the dynamics of organization social-networks and behavior via computer modeling and simulation. He is developing an expertise in workforce integration strategy and policy evaluation during organization mergers. He earned his doctorate (Ed.D. in Organization Change) from Pepperdine University, a MBA from New York University and a BS in Business Administration (Computer Systems Management) from Drexel University. Prior to entering academic research, for nearly 20 years he was a software applications development manager in the global financial services and industrial chemicals industries; most recently as a Vice President in Information Technology at Morgan Stanley in Hong Kong, New York and London.

Kathleen M. Carley is a professor at the Institute for Software Research International in the School of Computer Science at Carnegie Mellon University. She is the director of the center for Computational Analysis of Social and Organizational Systems (CASOS) <http://www.casos.cs.cmu.edu/>, a university wide interdisciplinary center that brings together network analysis, computer science and organization science (www.casos.ece.cmu.edu) and has an associated NSF funded training program for Ph.D. students. She carries out research that combines cognitive science, social networks and computer science to address complex social and organizational problems. Her specific research areas are computational social and organization theory, group, organizational and social adaptation and evolution, social and dynamic network analysis, computational text analysis, and the impact of telecommunication technologies and policy on communication, information diffusion, disease contagion and response within and among groups particularly in disaster or crisis situations.