Monitoring Corporate Password Sharing Using Social Network Analysis

Andrew S. Patrick
NRC Canada

Paper to be presented at the International Sunbelt Social Network Conference, St. Pete Beach, Florida, January 22-27, 2008.

Abstract

Corporations are facing increasing demands to monitor their compliance with policies and regulations. This research demonstrated one type of analysis tool for monitoring corporate security and privacy practices. Using the Enron email corpus as an example of corporate communications, the research explored methods to identify instances of password sharing, a practice that should be a security concern to any organization. Social network analysis was able to identify key creators and sharers of passwords, and an analysis of the passwords themselves showed that quality was clearly a problem. The network analysis was also able to reveal interesting communication patterns, such as sharing passwords with external accounts owned by the same person, which might have been useful as indicators of a problem in corporate systems or practices. The research also uncovered cases of possible policy violations, such as the sharing of internal and external accounts.

1. Introduction

Corporations are facing increasing demands to monitor their compliance with both internal and external policies and regulations. Externally, corporations are asked to comply with a variety of regulations depending on their location and areas of business. These regulations cover corporate finances (e.g., Sarbanes-Oxley), privacy (e.g., Gramm-Leach-Bliley, PIPEDA, HIPAA), national security (e.g., Patriot Act), etc. Internally, good corporate practices demand effective management, good decision making, clear accountability, effective risk management, corporate integrity, etc. Because of these increasing demands, corporations are turning to automated tools that can assist in monitoring data and processes, and in demonstrating compliance with the various requirements.

One area where compliance is especially important is security and privacy. Corporations are coming to realize that protecting their data, and the data of their customers, is essential for their business success. Here too, they are turning to automated tools to assist in the monitoring of their security and privacy practices (e.g., [1, 3]).

The purpose of the present research is to demonstrate one type of analysis tool for monitoring corporate security and privacy practices. Social Network Analysis (SNA) focuses on the ties among groups of entities, including people, groups, and even countries. SNA involves creating structures that represent the relationships (usually called ties) between entities (usually called nodes) based on some analysis of behaviors [13]. For example, the relationships between teenagers could be studied by examining the pattern of text messages sent between individuals in a given period of time.

It is our belief that SNA can provide a valuable tool for corporations interested in monitoring their security and privacy practices. SNA has proven to be valuable in other security contexts. Valdis Krebs [11], for example, demonstrated the value of relationship analysis for discovering the social structure among the hijackers involved in the 9/11 terrorist attacks. National security agencies, such as the NSA, also use social network analysis to search for current and future threats [4].

One security practice that is important to corporations is access control. The most common access control method today is to assign each individual a username and password. Based on the authentication provided by these pieces of information, different types and levels of access can be given or refused. If the authentication information is shared, however, then the access control measures fail and the corporation can loose control over their assets [17]. Even internal password sharing between colleagues can be dangerous because it can lead to a lack of accountability, errors or omissions by untrained users, inadvertent access to unintended resources, greater opportunities for malicious insider activities, and an environment that is vulnerable to social engineering attacks. As a result, many corporations have, or should have, policies against password sharing.

This study used social network analysis to examine cases of password sharing within a corporate environment. A corpus of corporate communications was searched and instances of password sharing that represent potential security problems were recorded. The purpose was to demonstrate the value of this analytic approach, and to suggest topics for future research and development in the area of automated compliance tools for security and privacy.

1.1. Background on the Corpus

Enron was a very large U.S. corporation that conducted business in the energy sector from the 1980s until the early 2000s. Enron acted as a broker, buying energy from generators and selling it to customers. With the rapid success of the business, Enron soon expanded their scope to include brokerage of a variety of commodities, including advertising time and network bandwidth. Enron grew very quickly and reported huge profits, but in 2001 stories began to emerge about improper accounting practices. The company seemed to be reporting only profits and hiding their losses in a series of related companies. When the scandal became public, the share price of the company collapsed and investors lost millions of dollars. Enron filed for bankruptcy in December 2001 [2, 6].

During the investigations that followed the collapse of the company, the Federal Energy Regulatory Commission collected a large number of corporate email messages. This corpus was made available to the public and it was cleaned by researchers at SRI and Carnegie Mellon University [8]. Message integrity issues were resolved and some of the emails were deleted at the request of some of the affected employees. The result was a database of 517,431 email messages from 151 users distributed in 3500 folders [14]. This database is available for download from CMU, and it has frequently been used by researchers interested in corporate communications, natural language understanding, and social networks (e.g., [7, 12, 15]).

2. Method

2.1. Corpus Cleaning

For the present study, the Enron email corpus was downloaded from Carnegie Mellon University[1] and unpacked on a PC running Windows XP. A series of Perl scripts were developed to clean the data and search the corpus files. The first cleaning pass involved parsing each message for the date (using the Date: header), the sender (using the From: header), and the recipients (using the To:, Cc:, and Bcc: headers). In cases where there were multiple recipients, each recipient was counted as a separate message for our analyses.

There were a number of email address aliases in the email corpus. For example, employees could be addressed using the pattern
firstname.lastname@enron.com
as well as
firstname.middle-initial.lastname@enron.com
and
first-initial..lastname@enron.com.

A list of aliases was developed by examining similarities in the email addresses. This list was also refined as the social network analysis was conducted because it sometimes became clear that two closely related entities were actually the same person. The result was a list of 49 aliases that were converted to a canonical form
firstname.lastname@enron.com
for the analyses that follow.

Following the alias substitutions, duplicate messages were removed from the corpus by looking for repetitions of the date, sender, and recipient. The result was a total of 250,641 unique messages, which is similar to the 252,759 unique messages reported by [14] who dropped messages based on duplicate messages automatically saved by the email system in some of the corpus folders (e.g., "sent_mail" folders). There was a total of 31,718 different email addresses found in the database, including both Enron employees and outside correspondents. Most of the addresses (63%) appeared only once in the database, but some addresses (those of prolific Enron employees) were very frequent.

2.2. Searching for Passwords

A simple method was used for searching for passwords within the cleaned email corpus. Each message was searched using two case-insensitive regular expressions: "password:" and "password is". When a match was found, the next whole word in the message line was recorded as the password. This method did result in some false matches. For example, the sentence "Your password is case sensitive" would result in a recorded password of "case." A list of common false matches was developed (see Table 1) and these passwords were ignored in the analyses. In all other cases, the sender, recipient, and password was recorded and used for later analyses.

Table 1: Password False Matches Not Used in Analyses

a	also	as
birthdate	blank	can
case-sensitive	case	completed
currently	different	first
for	in	included
issued	necessary	needed
not	noted	now
only	our	required
same	the	their
upper	what	your
travelocity		

2.3. Social Network Analysis

The instances of password sharing were counted to produce a dataset that was suitable for social network analysis. The direction and frequency of password sharing were tallied, and the results were saved as a Pajek net-

[1] http://www.cs.cmu.edu/~enron/

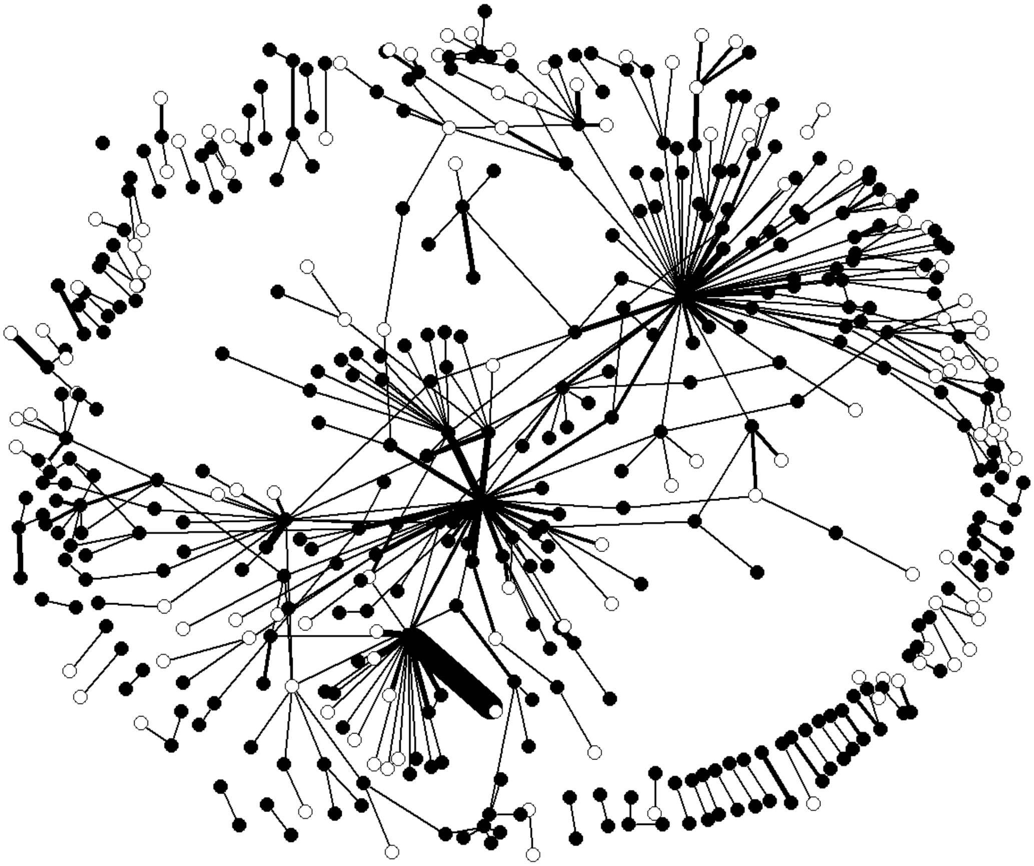

Figure 1: The total network of Enron password sharing. Black nodes represent internal Enron addresses, while white nodes represent external addresses.

work file.[2] The Pajek program was used to analyze and visualize the resulting network. In addition, the NetDraw program[3] was used for analyses that were not available in the Pajek program (e.g., faction analysis).

3. Results

For the social network analysis, each email address was treated as a node and the frequency of password sharing was used to create weighted directional arcs. A total of 642 instances of password sharing involving 500 different email addresses were observed. This data resulted in a network of 500 nodes and 500 arcs, as is shown in Figure 1. This network was very sparse, with 418 of the nodes (addresses) only appearing once. The network density was 0.2%, showing that very few of the possible network connections actually occurred. There were also 9 instances of network loops, representing cases where people emailed passwords to themselves.

There are a number of interesting characteristics observable in the network shown in Figure 1. First, there is clearly a group of relatively well connected nodes at the centre of the network and a large collection of isolated pairs at the periphery. Second, Enron internal addresses are shown as black nodes in the network, while external addresses are shown in white. As might be expected, external addresses were most commonly found in the network periphery, but there were some cases of well connected external nodes. A review of the messages involving those nodes revealed that they were most commonly associated with subscription services, such as the New York Times, that were popular with a variety of Enron employees. In fact, Enron employees were frequent subscribers to a variety of third party news and research services, and a majority of the cases of external password sharing were new subscription passwords being sent to the Enron subscribers. We will return to this finding later.

[2] http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[3] http://www.analytictech.com/Netdraw/netdraw.htm

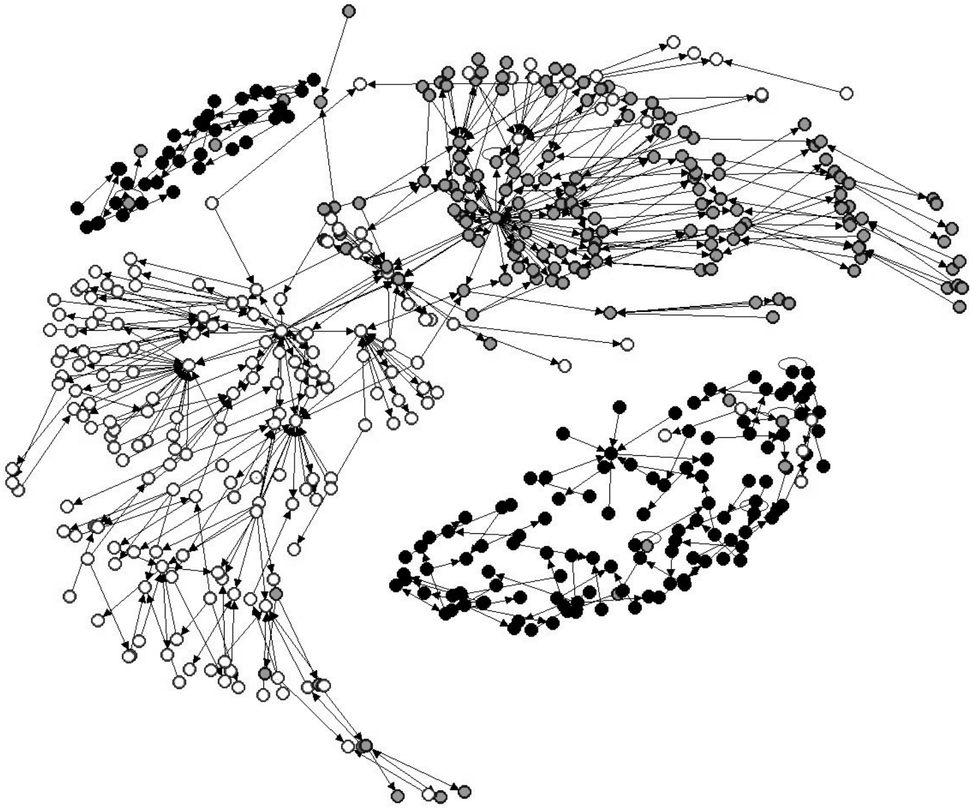

Figure 2. Faction analysis of the password sharing network.

A third observation from the password sharing network shown in Figure 1 is the scarcity of repeated password sharing. The width of the arcs in the diagram represents the frequency of password sharing between the respective nodes. In most cases, a password was only shared once between nodes (thin lines), and very frequent password sharing was very rare. The highest frequency of password sharing between two nodes was 14 instances in one direction and five in the other. We will examine this unusual case below.

3.1. Network Factions

To further characterize the network of password sharing, a faction analysis was conducted using the NetDraw tool. This analysis breaks the network into factions based on the number of links between nodes. Nodes with strong connections to each other are placed within the same faction, and a unique color is used to represent each faction in the network. An analysis using three factions resulted in the cleanest separation of the nodes, and this is shown in Figure 2. There are clearly two large factions represented by the white and grey nodes, with frequent connections within the factions and very little overlap. It is obvious that an analysis focusing on the differences between these two groups is required. The black nodes are a residual faction that represents people that are relatively unconnected, often involving pairs of people and small groups. NetDraw has chosen to draw these nodes in two groups, but they really represent one faction of isolated nodes.

3.2. The Network Core

To analyze the network core, the NetDraw program was used to remove isolated nodes and pairs (pendants). This was done in two passes, and the resulting core network is shown in Figure 3. This analysis shows the reason behind the two main factions discussed above. There appear to be two nodes (Nodes 1 and 12 in the figure) that are heavily involved with password sharing to nodes that, in turn, are not well connected to each other. For the entire network, Node 12 had the highest degree centrality score of 57 (5.7% of nodes), while Node 1 had the second highest score of 31 (3.1%). The average degree centrality for all nodes was 0.2%.

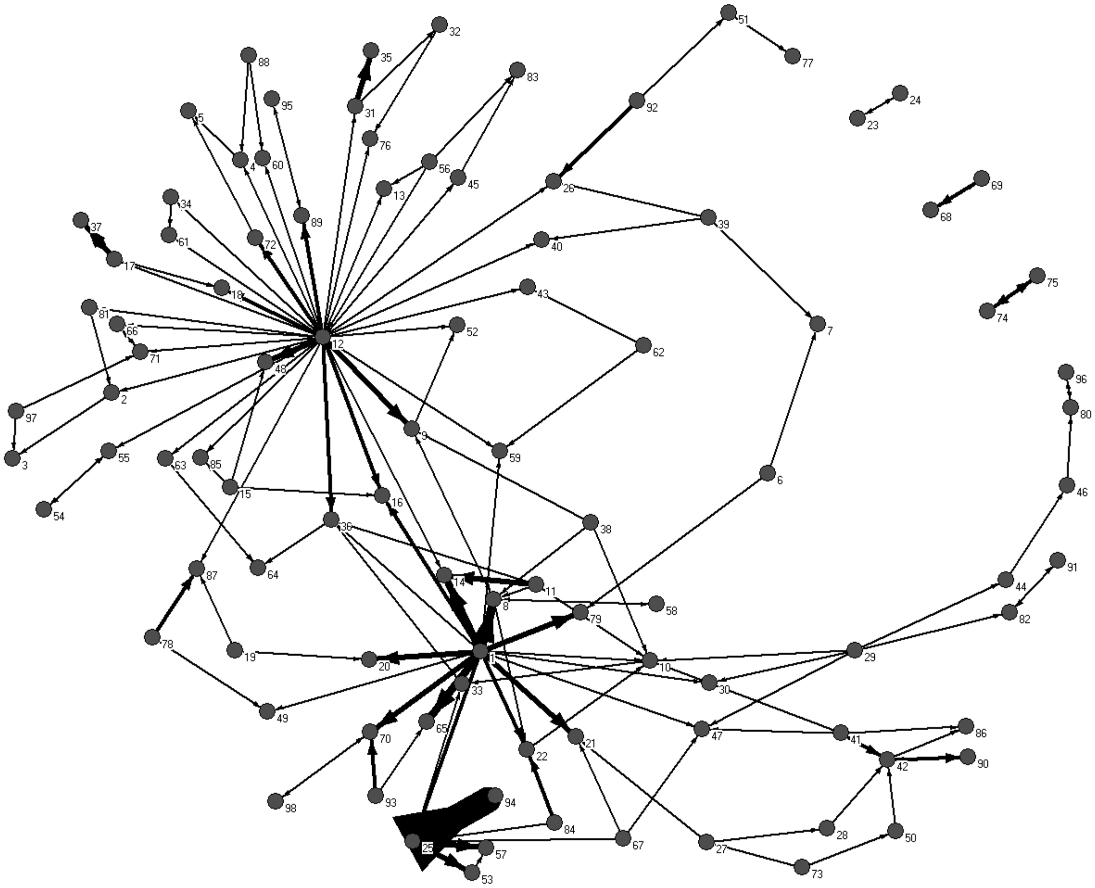

Figure 3: The core network after isolated nodes and pairs were removed in two passes.

Node 1 in the core network shown in Figure 3 represents an email address used by the employee performance management system within Enron. This was an online system that was used to periodically collect performance appraisals about employees. During each of the evaluation periods, a username and password was sent from the Node 1 address to the senior managers of the corporation. In addition, if the managers did not use the system, they were sent reminders and their username and password were repeated. The network surrounding Node 1, then, represents senior managers of the corporation. They received passwords from the performance management system, and they also sometimes shared passwords among themselves.

Node 12 in the core network represents a senior manager at Enron Online, the virtual trading arm of the corporation. Even though her personal email folders were not included in the corpus, her messages were often found in other employees' mailboxes. This manager appears to have been responsible for setting up accounts for traders who used the online trading system. The network surrounding Node 12, then, represents traders who used the Enron Online service. In most cases, these traders did not share passwords with the corporate managers who are clustered around Node 1.

It is also interesting to note the frequent password sharing between Nodes 25 and 94 in Figure 3. These nodes represent an internal (Node 25) and external (Node 94) email address used by the same person, a senior manager of research within Enron. The external address was with the AOL Internet service, and this person sent a password from his Enron address to the AOL address 14 times, and received a password from the AOL address 5 times. In most cases, the passwords were to subscription services provided by third parties (e.g., news and research services). A possible reason behind this frequent sharing of passwords to an external address was revealed by a message from the email system administrators sent in July 2001. This message announced that external access to the corporate Outlook email system was now possible. So, it is likely that the pattern of sending passwords to an external address may have been necessary in order to work remotely. It is interesting to note that this person continued his self-sharing practices after this announcement, perhaps out of habit or convenience.

This pattern of sending passwords to an external address demonstrates one possible value of social network analysis for security and privacy compliance. Had this social network analysis been done during the life of the corporation, the pattern of sending passwords to external addresses could have been detected and remedial measures (e.g., policies, training, message blocking) put in place. Moreover, the root cause of the behavior, an apparent inability to access the internal mail system, could have been addressed sooner.

It is also apparent in Figure 3 that Node 25 was also involved with other password sharing (sending to Nodes 57 and 53 and receiving from Nodes 84 and 67). In one case, this person shared a password for a third party service with an internal colleague, and in another case he sent software installation instructions, including a download password, to a group member. These might represent breaches of internal and external security policies, so automatic detection of these messages may have been valuable.

3.3. Password Quality

Since Nodes 1 and 12 represent key sources of passwords within the corporation, it may be useful to examine the actual passwords that they sent for repetition and strength [5]. The Node 1 address (the performance management system) was used to send passwords 62 times, and in 39 of those cases the same weak password ("WELCOME") was sent to different people. This means that it would have been very easy for an Enron employee to guess another employee's password, and the password could have been easily cracked using dictionary searches. In other cases, however, a stronger unique password (e.g., "KTDVWCCH") was sent to the managers. These tended to occur in messages that were identical in format and they may have been generated by an automated process.

Node 12 in the core network (the Enron Online manager) sent out passwords 74 times, but only 20 different passwords were used. One password ("q#9M#npX"), although apparently strong, was sent 30 times to 30 different recipients. Node 12 also used the password "WELCOME!" 20 times for 15 different recipients, and the password "enron1" 4 times for 4 different recipients. Again, this means that an insider might be able to guess another employee's password based on the one that they received.

These password quality findings also demonstrate the value of social network analysis for identifying possible security and privacy issues. They key originators of passwords were identified using the communication patterns found in the email corpus, and a content analysis showed poor security and privacy behaviors. Had such an analysis been done during the life of the corporation, remedial training programs and better account management tools and practices could have been put in place to ensure that high-quality, unique passwords were being used.

3.4. The Network Periphery

The email corpus also contained many instances of password sharing between isolated node pairs and small groups. It may be useful to examine these instances of password sharing to look for any unusual behaviors. As a demonstration, Figure 4 shows the entire password sharing network again, but the coloring of the nodes has been changed. Nodes that are well connected are colored white, and isolated nodes and pairs are colored black. Small isolated groups, representing more than 3 nodes, are shown in a grey color in this figure, and these clusters are the focus of this analysis.

This first cluster of interest is the group of 7 nodes colored grey appearing at the 12 o'clock position in Figure 4. These instances of password sharing centre around one Enron employee who stored his passwords in a specific email folder. In most cases, this person received passwords from third party services, usually personal services used to purchase auto parts or auto loans. In one case a password was sent from a personal external address (on the MSN service) to the internal Enron address, and this involved a password for an online discussion forum about trucks. These instances might be in violation of policies against conducting personal business at work. Also, in one instance, this employee shared his password to the Enron Online service with an internal colleague, and this might represent a breach of corporate security policies.

The second cluster of interest is the group of 5 nodes colored grey appearing at 11 o'clock in Figure 4. This cluster was involved in sharing a collection of password-protected memos and documents. It is not clear why these documents were being protected, but the behavior appears to be unique to this group and might represent a concern about insider activities or a breach of corporate policies.

The third interesting cluster is the group of 4 nodes colored grey appearing at 9 o'clock in Figure 4. This cluster is centered on an Enron employee who received a password for a third party service and shared that password with an internal colleague. This person also twice sent account information for a third party service to a personal address on a local Internet service. Again, this password sharing behavior might be of interest from a corporate security perspective.

The fourth cluster of interest is the group of four grey nodes appearing at 8 o'clock in Figure 4. This cluster is centered on an Enron employee who sent her password to an internal help desk so they could diagnose a system problem. She also shared the password to an airline reservation service with an internal group address, and she sent an employee ID number and password to a new employee. These may or may not represent activities that are of interest from a security and privacy perspective.

4. Discussion

The purpose of this research was to demonstrate the value of social network analysis for the monitoring of security and privacy issues within an organization. Using the Enron email corpus as an example of corporate communications, we were able to develop simple methods to identify instances of password sharing, a practice that should be a security concern to any organization.

Social network analysis was able to identify key creators

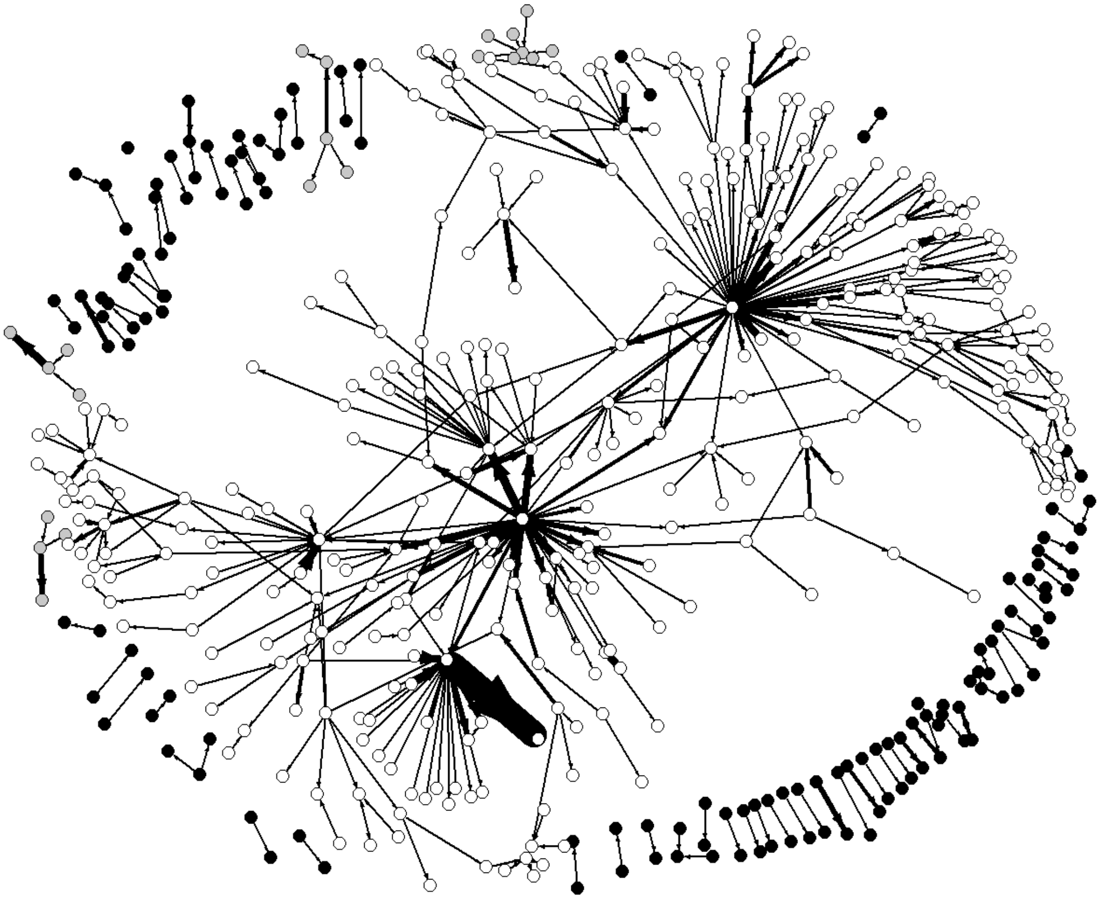

Figure 4: The full password sharing network highlighting isolated clusters involving more than 3 nodes (colored grey). The well connected nodes are colored white, while isolates and pairs are colored light black.

and sharers of passwords, and an analysis of the passwords themselves showed that quality was clearly a problem. The network analysis was also able to reveal interesting communication patterns, such as sharing passwords with external accounts owned by the same person, which might have been useful as indicators of a problem in corporate systems or practices. We were also able to uncover cases of possible policy violations, such as the sharing of internal and external accounts. In addition, one case of possible insider collusion was found in a pattern of sharing password-protected documents.

Although password sharing may be contrary to the goals of keeping information secure, it may be necessary in some situations. If people are asked to work together on documents or systems, for example, and the organization does not provide methods for shared access, then it is inevitable that people will share whatever passwords are necessary to get the job done.

Australian researchers recently examined the issue of people sharing bank passwords with friends or family members [16]. Although banks explicitly stated that customers should not share their passwords and they risk losing all financial protections if they do, this study found many cases where password sharing was considered acceptable and necessary. For example, married couples often found it necessary to share access to bank accounts, and such sharing was often seen as an expression of trust in one's partner.

Singh et al. [16] also documented password sharing within remote communities where banking services were scarce. Often only one member of a group is able to travel to the bank, so that person might carry many cards and passwords to do the banking for many people. Elderly and disable people may also have to share their banking passwords with family, friends, or caregivers because they cannot visit the bank themselves. Thus, uncovering examples of password sharing is only a first step. Understanding the reasons behind the sharing and the appropriate response is also very important.

The major shortcoming of this work is that it has all been done after the fact. In order to be of significant value to a company or organization, the possible security issues would have to be discovered as they occur. Moreover, due to the large amount of data that would have to be monitored, the identification of items of interest would have to be largely automatic. We are currently conducting research and development in these areas [9, 10].

5. References

1. Ashley, P., Powers, C., & Schunter, M. (2002).From privacy promises to privacy management: A new approach for enforcing privacy throughout an enterprise. *Proceedings of the 2002 Workshop on New Security Paradigms*, 43-50.
2. Benston, G.J., & Hartgraves, A.L. (2002). Enron: What happened and what we can learn from it. *Journal of Accounting and Public Policy, 21,* 105-127.
3. Bhattacharya, J., Dass, R., Kapoor, V., & Gupta, S.K. (2006). Utilizing network features for privacy violation detection. *Proceedings of the First International Conference on Communication System Software and Middleware*, Delhi, India, Jan. 8-12, 1-10.
4. Cauley, L. (2006). NSA has massive database of Americans' phone calls. *USA Today*. URL: http://www.usatoday.com/news/washington/2006-05-10-nsa_x.htm
5. Cisneros, R., Bliss, D., & Garcia, M. (2006). Password auditing applications. *Journal of Computing Sciences in Colleges, 21*(4), 196-202.
6. Diesner, J., & Carley, K.M. (2005). Exploration of communication networks from the Enron email corpus. *Proceedings of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining*, 3-14.
7. Keila, P.S., & Skillicorn, D.B. (2005). Structure in the Enron email dataset. *Computational & Mathematical Organization Theory, 11*(3), 183-199.
8. Klimt, B., & Yang, Y. (2004). Introducing the Enron corpus. *First Conference on Email and Anti-Spam (CEAS)*. URL: http://www.ceas.cc/papers-2004/168.pdf
9. Korba, L., Song, R., Yee, G., & Patrick, A.S.. (2006). Automated social network analysis for collaborative work. *Proceedings of the Third International Conference on Cooperative Design, Visualization and Engineering* (CDVE 2006). Palma de Mallorca, Spain. September 17-20.
10. Korba, L., Song, R., Yee, G., Patrick, A..S., Buffett, S., Wang, Y., & Geng, L. (2007). Private data management in collaborative environments. To appear in *Proceedings of the Third International Conference on Cooperative Design, Visualization and Engineering* (CDVE 2007). Shanghai, China. September 16-20.
11. Krebs, V.E. (2002). Uncloaking terrorist networks. *First Monday, 7*(4). URL: http://firstmonday.org/issues/issue7_4/krebs/index.html
12. McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2005). Topic and role discovery in social networks. *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, July.

13. Nooy, W. and Mrvar, A. and Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge University Press.
14. Shetty, J. & Adibi, J. (2004).The Enron email dataset database schema and brief statistical report. Technical report, Information Sciences Institute, 2004. URL: http://wittas.info/diplomka/soubory/Enron_Dataset_Report.pdf
15. Shetty, J., & Adibi, J. (2005). Discovering important nodes through graph entropy: The case of Enron email database. *Proceedings of the 3rd International Workshop on Link Discovery*, 74-81.
16. Singh, S., Cabraal, A., Demosthenous, C., Astbrink, G., & Furlong, M. (2007). Password sharing: Implications for security design based on social practice. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 895-904.
17. Wilson, S. (2002). Combating the lazy user: An examination of various password policies and guidelines. SANS Institute. URL: http://www.sans.org/reading_room/whitepapers/authentication/142.php