CONTINUOUS ASSURANCE USING TEXT MINING

Glen L. Gray
College of Business and Economics
California State University at Northridge
glen.gray@csun.edu

Roger Debreceny
School of Accountancy
Shidler College of Business
University of Hawai`i at Mānoa
roger@debreceny.com

DRAFT ONLY – DO NOT CITE
Comments welcome.
23 October 2006

CONTINUOUS ASSURANCE USING TEXT MINING

ABSTRACT

Three levels of continuous assurance can be delineated, from continuous *monitoring* over business processes at the lowest level to continuous *audit* of corporate performance at the highest level. A number of research projects and commercial software products address monitoring of business processes. These projects include a variety of monitors that can be grouped under the general rubric of embedded audit modules. These modules normally are designed to monitor events and controls within accounting information systems and transaction processing subsystems such as sales systems.

However, much of the information that is relevant to continuous monitoring of business processes takes place outside the accounting information system. It may be difficult, for example, to tease out evidence of questionable related party arrangements or inappropriate revenue recognition by monitoring transactions and accounting information systems.

Increasingly email is the information DNA of the corporation. As executives drag out their Blackberries or other mobile devices at every opportunity to receive and send emails, many types of information flow not only within the entity but between the entity and its business partners and customers. Contained within those emails will be potential evidence of fraud indicators and other matters of audit concern. As such, continuous monitoring of email provides a rich opportunity for research and practice.

Emails are semi-structured data, with known fields for sender, recipient, subject, date and body. Assessment of emails requires bringing together an understanding of the social networks that underpin email exchanges, textual analysis using natural language processing, and other techniques and domain knowledge. Emails are also relatively noisy data. New metrics and tools will be required to resolve this data quality. This paper assesses the current state of practice and research, which has been given a considerable

boost by the availability of a large corpus of emails from the now defunct Enron Corporation.

Preliminary indications from this growing research domain indicate that email monitoring will be an important area for ongoing work, both in computer science and in auditing and assurance. The paper concludes with the depiction of a research agenda.

I. INTRODUCTION

This paper explores the potential role for the use of text mining and in particular mining of emails in the support of continuous monitoring over business processes and continuous assurance on the integrity of the financial reporting process.

We categorize continuous monitoring and assurance into three levels. At the highest level, continuous *audit* calls for ''a methodology that enables independent auditors to provide written assurance on a subject matter using a series of auditors' reports issued simultaneously with, or a short period" (CICA/AICPA 1998, xiii). At this level, audit reports are issued that assess a given subject matter against established criteria[1]. These audit reports might be over a particular data element such as entity revenue or customer count or on a complete set of financial reports. Whilst the term *continuous* audit is used, it is clearly a misnomer. Human judgment is required in reaching a conclusion on the outcome of the process of gaining assurance and human intervention.

At the next level, continuous *assurance* is a broader concept than continuous audit. Vasarhelyi (2002) defines continuous assurance as "an aggregate of objectively provided assurance services, derived from continuous online management information structures—the objective of which is to improve the accuracy of corporate information processes." The provision of assurance implies, as with audit, a report by an assurance provider on a subject matter against established criteria. Whether human intervention is required in the process of coming to a judgment in the process of continuous assurance is an open question. Vasarhelyi et al. (2004, 9) note that "the degree of automation of such sophisticated high level judgments is clearly limited – but not non-existent – and the likely role of the CA system will be that of facilitator." The internal focus of Vasarhelyi's definition is also worthy of note. The provision of continuous assurance over a variety of management reporting processes provides a foundation for continuous audit.

[1] We use the schema of the International Auditing and Assurance Standards Board (IAASB) in its *International Framework for Assurance Engagements.*

Finally, at the lowest level, Alles et al. (Forthcoming) introduce the concept of "continuous *monitoring* of business process controls (CMBPC)" (emphasis added) In a CMBPC environment, automated tools monitor data stores that are relevant to the conduct and completion of a variety of business processes. An example of a business processes is the interaction between the enterprise and client that culminate in the recognition of revenue and subsequent collection of cash. The focus of monitoring may be to facilitate the enterprise meeting its business goals and objectives, the detection of fraud or mitigation of other risks. An example of a class of CMBPC are embedded audit modules and related techniques (Debreceny et al. 2003; Groomer and Murthy 1989).

Alles et al. (Forthcoming) provide a case study where an enterprise leverages its investment in Enterprise Resource Planning (ERP) applications to build business process monitors that are for all intents and purposes, embedded audit modules (Debreceny et al. 2003; Groomer and Murthy 1989). These modules operate exclusively within the realm of the accounting information systems (AIS). These systems have highly structured data. Building continuous monitoring on the AIS brings the benefit of being able to directly scrutinize the transactions that result in the financial statements. The information within the AIS is, however, only a small subset of all the complete set of business-process relevant data that exists within the entity's information environment. There has been some discussion of the consideration of monitoring of qualitative data sources outside the accounting information system (e.g. Vasarhelyi and Peng 1999). There have, however, been few advances in the exploration of qualitative data sources within a continuous assurance or continuous monitoring environment. The rapid evolution of data mining techniques on unstructured or semi-structured textual data now provides many opportunities for researchers. The information outside the accounting information system may provide a rich data set for continuous business process monitoring. This paper focuses on a large and important pool of semi-structured qualitative data, namely, the organization's emails and the attachments to those emails. In many situations email communication is the primary media for communications (vs. the telephone or paper-based communications) within the enterprise and outside the organization. Email data mining has the potential to be an important instrument in the auditor's toolkit in areas such as fraud detection. Besides the obvious search for key words such as bribe or kickback, research has shown that email writers change their behavior when they compose deceptive emails—even if they know that their emails may be monitored. The social networks (links between email senders and recipients) or changes in the social networks have been also shown to help locate fraud.

Research on mining of emails has been given a considerable boost with the release by the Federal Energy Regulatory Commission of a large database of emails sent by employees of Enron Corporation. This database includes over 500K emails from 151 Enron users, comprising mostly senior managers. The database has allowed researchers in allied disciplines to better understand both the potential benefits and challenges that can arise from mining emails. Researchers in continuous assurance can benefit from this research as well as build specialized applications and approaches that respond to the particular challenges of this domain.

We discuss the issues that arise in the particular context of emails. In this paper we demonstrate how data mining tools can be used to mine the Enron data and be applied to continuous auditing. The remainder of this paper is organized as follows. Section 2 provides an overview of email data mining research. Section 3 discusses the research and industrial tools available for monitoring of email and other textual resources. Section 4 discusses our trial analysis of the Enron email corpus. Section 5 presents our conclusions and recommendations for future research.

II. EMAIL TEXT MINING

Techniques

Emails are a combination of structured and unstructured data. Each email includes structured and standard data fields such as primary recipient(s); copied recipient(s); sender; date and subject line. The text of the email is essentially unstructured. Minkov and Cohen (2006) notes that these characteristics of emails allow us to view a set of semi-structured emails as a graph, with nodes representing actors, temporality, subject matter and meetings. We can exploit these characteristics by employing a variety of techniques as illustrated in Figure 1.

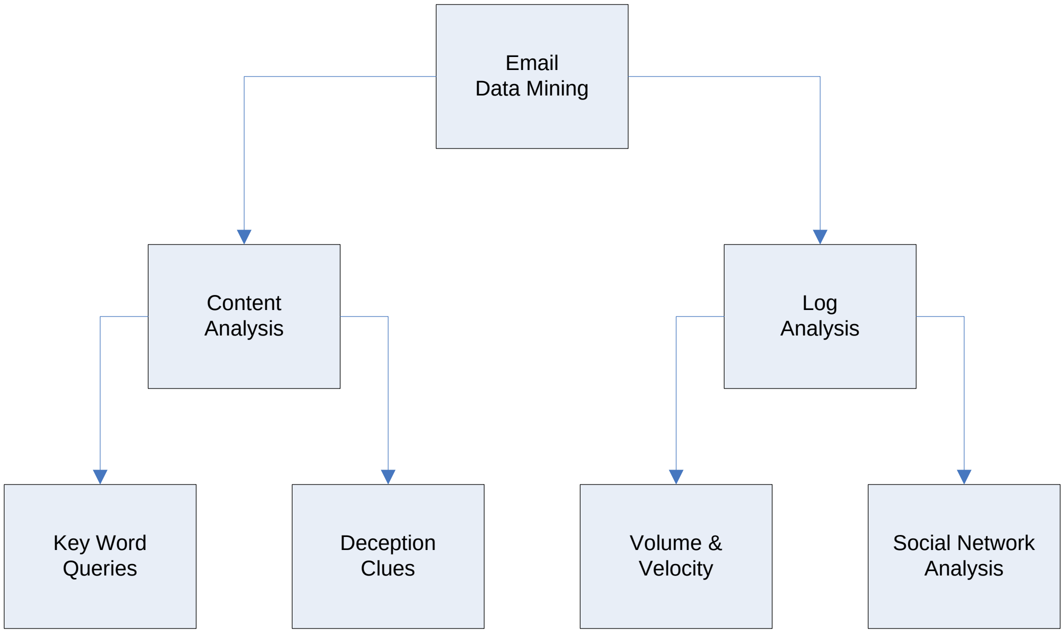


Figure 1: Email data mining framework.

The techniques, which will be discussed in the following paragraphs, can be divided into two broad categories. The first category is those techniques that data mine the content of emails and attachments to those emails. The second category focuses on the email logs as distinct from the content of the emails. These techniques are not mutually exclusive and, as the remainder of this paper discusses, blending these techniques can improve the efficiency and effectiveness of locating fraud.

Content Analysis

Content analysis of written materials has a long history in a variety of research areas including language and literature, sociology, and information systems. Applying that body of research to email contents is an obvious extension of that research. The following paragraphs summarize key word queries and identifying deceptive emails. Weaved in those discussions are comments regarding the use of continuous monitoring as it could apply to email content analysis.

Key Word Searches. Emails could be searched for key words such as finder's fee, bribe, kickback, and similar words that could indicate questionable actions or overrides of controls. This would be cherry-picking the naive fraudsters. It is hard to believe that a fraudster would use such words in the company email, but they do as found by researchers who have explored the Enron email corpus. Some of these people may falsely believe that company email has the same privacy protection as using the U.S. postal system to mail a letter.

A simple query of hundreds of suspicious key words by the auditor would probably produce an overwhelming population of false positives. Like any database query, however, the false positives could be greatly reduced by adding more parameters to the query. For example, extracting emails with the suspicious key words AND sent to domains of vendors that the company does business with AND vendors who were granted a new contact during the fiscal year. This query might extract emails where an employee requested a "finder's fee" for helping ensure that the company will win the contract. As another example, extracting emails with key words AND sent to generic domains (e.g., AOL, Hotmail, Gmail, etc.), which could indicate someone is trying to disguise his/her company affiliation.

Probably the most discussed continuous email monitoring is the *Carnivore* system developed by the FBI to scan emails in the United States. The CIA and NSA are assumed to have similar systems to monitor email traffic outside of the U.S. The difference being that the FBI needs a court order before it can monitor a specific person's email traffic in the U.S. By the way, companies do not need a court order to monitor employee emails. Currently, most companies automatically scan all emails moving through their email server for viruses and other malware. They also scan the content to capture spam. As such, scanning emails for fraud-related key words could be an extension of the current scanning process. There would be a performance hit because of the extra scanning, but the hit should be manageable even for more compound criteria (e.g., key word AND current vendor).

Deception Clues. One growing body of email mining research is deception research. Deception can take two forms. In the classic form of deception the sender is deceiving the recipient of the email. This form of deception could be an out right lie or could be the *normal* part of a negotiation strategy. The more subtle form of this type of deception is when there is collusion between the sender (e.g., an employee) and recipient (e.g., a vendor) and they are trying to deceive a third person (e.g., the employer) monitoring their emails. The second broad type of deception is when the email sender tries to disguise their identity.

Content Deception. In terms of the first type of deception introduced above, Keila and Skillicorn (2005) state that individuals who are trying to deceive generally include the following in their emails:

- Fewer first-person pronouns to dissociate themselves from their own words
- Fewer exclusive words, such as but and except, to indicate a less complex story
- More negative emotion words because of the sender's underlying feeling of guilt
- More action verbs to, again, indicate a less complex story

In addition, according Skillicorn (2005) and other researchers, even senders who suspect that their emails may be monitored will alter their emails toward "excessive blandness." According to these researchers, the deceivers are following these behaviors to reduce the cognitive demand of the deception. They want to disassociate themselves from their statements and keep their story simple because it is hard to remember all the details of a complex story.

The above parameters (first-person pronouns, exclusive words, negative emotion words, and action verbs) can be used to ranked emails based on the relative aggregate scores on those parameters. Then a specific person's emails can be compared to his/her other emails to see if the scores have changed over time. A person's emails can also be compared to his/her peers to determine if a person's parameters are significantly different than the peers.

To help conduct this type of deception analysis Pennebaker et al. (2001) developed software called Linguistic Inquiry and Word Count (LIWC)[2].

Keila and Skillicorn (2005) applied the four deception parameters to the publicly available Enron emails mentioned in the introduction to this paper. The parameters did identify deceptive emails. In addition to identifying deceptive emails, the parameters also identified potential organizational dysfunction such as "complaining, conveying information improperly, or spending organizational resources and employee time on non-work-related issues."

Sender Deception. Another form of deception is the email sender trying to disguise his/her identity—whether in written documents, emails, forum postings, or blogs. Like a person's unique fingerprint, researchers have shown that people have unique writing styles or "writeprints" that includes "vocabulary richness, length of sentence, use of function words, layout of paragraphs, and keywords." (Li et al. 2006). To develop these writeprints, some researchers have developed a body of *stylometric* research (e.g., see McEnery and Oakes 2000). A wide variety of features can be used to define writeprint characteristics. According to Li et al. (2006) these features can be divided into the four following categories:

1. **Lexical features**. These features relate to characters and words that are used by a writer. For example, the lexical writeprint features could include the lists of words, the relative frequency of the words, and lengths of sentences that appear in a person's writings. Different researchers have used specific lexical characteristics to differentiate or to identify authors: sentence length and vocabulary richness (Yule 1944); a set of 50+ high frequency words (Burrows 1992); and just focus on two- and three-letter words and "vowel word" (words that start with a vowel) (Holmes 1998).

2. **Syntactic features**. These features relate to sentence structure. Part of the analysis of syntactic features includes the use of punctuation and function words.

[2] See http://www.liwc.net/

Function words (as opposed content or lexical words) could be articles (or determiners), prepositions, pronouns, auxiliary verbs, conjunctions, and particles. There are about 300 function words in the English language.[3] A variety of studies have shown that analyzing syntactic feature was superior to analyzing lexical features alone. Holmes (1998) found that function words had good discriminating capability. Baayen (2002) found that including punctuation in the analysis also improved discriminating capability. Stamatatos et al. (2001) explored passive count and part-of-speech tags.

3. **Structural features**. These features relate the overall structure of the author's writing, which has been shown to be strong evidence of personal writing style. de Vel et al. (2001) achieved a high level of author identification performance analyzing emails.

4. **Content-specific features**. While the other three categories of features are essentially content free, how the author uses content keywords related to a specific topic has been shown to have additional discriminating power.

If a researcher attempted to operationalize these categories of features by developing metrics to measure those features, the potential lists would be almost limitless. So, a first major step to comparing emails writeprints is developing a set of features that are discriminatory, measurable, and manageable. Rudman (1998) used almost 1,000 writeprint features to analyze written materials. Zheng et al. (2006) created a taxonomy that included 270 features that broke down as follows:

- 87 lexical features
- 158 syntactic features
- 14 structural features
- 11 content-specific features

[3] See http://www.speech.psychol.ucl.ac.uk/transcription/intro.html

Adding more features to the analysis does not always improve discriminating power. de Vel et al. (2001) found that the performance of their analysis decreased when they increased the number of function words to 320 from 122. Li et al. (2006) used the generic algorithm form of heuristic search to find the optimum subset of features. Starting with 270 features introduced in the prior paragraph, they found the optimum subset for identifying message authors includes 134 features. Their finding of 134 features is not universal; instead the results will vary depending on the textual materials being analyzed and the language used by the authors. The population of textual messages that the researchers used were messages from the *misc.forsale.computers* newsgroup that involved the selling of pirated software. When the researchers applied the general algorithm to similar messages in Chinese newsgroup messages, they started with 114 features and found the optimum subset was 56 of those features.

Various forms of pattern recognition, neural networks, artificial intelligence, and other data mining techniques have been used to reduce very large feature sets down to optimal sub-sets. (See Liu and Motoda (1998) for a summary of feature selection studies and their resulting general framework.)

In terms of applying these deception detection techniques too a continuous monitoring environment there will be two major tasks. These techniques use aggregated data to form a baseline and then new emails are compared to that baseline. As such, the first task will be creating that baseline and the main issue for this task is selecting the specific (optimum) set of features that will be used. Once the aggregated baseline is developed each email passing through the email server can be scanning to develop metrics that email that, in turn, will be compared to the baseline metrics.

Temporal Analysis

The analysis of email logs (both sent and received emails) can reveal important information about employees' interests, activities, and behaviors that cannot be derived from email content analyses alone. The logs of email traffic are far more structured and less noisy than email content, which may mean that a wider variety of existing data mining techniques could be applied to those logs. The market leaders in email server software have very large market shares (i.e., Microsoft Exchange has a 31% market share, followed by IBM Lotus Domino/Notes with 20%)[4]. As such, once the auditors have developed a set of data mining routines for a specific client's email server logs, those routines will directly apply to many other clients who use the same email server.

Some of the email log analysis to find suspicious activities comes from research to identify and capture incoming spam. According to one study nearly 75% of email traffic is spam.

Volume and Velocity. Volume and velocity are two examples of metrics that can be analyzed from the temporal mining of email logs. Volume measures the number of emails a person sends and/or receives over a period of time. Velocity measures how quickly the volume changes. Gradual change would be low velocity and sudden jumps in volume would be high velocity. Changes in velocity over time for no apparent reason may also indicate suspicious activities. The first task is to create a *baseline normal* profile and then changes in volume and velocity compared to the profile for no apparent reason may indicate suspicious activities. Recognizing that the baseline profile can evolve over time, Stolfo et al. (2006) use the term "rolling histogram" to reflect the dynamic aspects of an employee's profile changing over time.[5] The concept of the rolling histogram is like other moving averages where the profile is updated based on a moving window of a time interval (e.g., a twelve-week moving average).

In terms of continuous monitoring of volume and velocity, the key issue is determining the optimum time intervals to sample the data. Continuous monitoring cannot be *continuous* in terms of sampling in real time that can be done with some accounting transactions. Comparing hourly, daily, and even weekly volumes and velocities will result in an overwhelming number of false positives. There are many legitimate reasons why volume and velocity would change from hour-to-hour and day-to-

[4] The Radicati Group, Inc., Microsoft Exchange Market Share Statistics, Palo Alto, CA.

[5] The software that these researchers use is called Email Mining Toolkit (EMT), which was developed at Columbia University and includes approximately 132,000 lines of Java code. It interfaces with relational databases. For more information on EMT and to download a copy go to http://www1.cs.columbia.edu/ids/emt/

day. Even weekly changes could be radical such as a sudden spike in volume after an employee returns from vacation. For a business that is very seasonable even quarterly variations can be easily explained.

This does not mean that continuous monitoring is not applicable monitoring volume and velocity; instead, it means some *intelligence* has to be built into the monitoring algorithms. Reasonable time intervals will have to be selected and what is reasonable may be different for different job classifications. For some job classifications changes in daily patterns may be significant and for other job classifications weekly patterns may be most applicable.

Social Network Analysis. Other email mining concepts include link discovery (LD) and social network analysis (SNA) where relationships or networks between email senders and recipients are explored based on mining email logs (Shetty and Adibi 2005). Some social network researchers trace their work back to Milgram's frequently cited 1967 article "The Small World Problem" (Milgram 1967). His paper re-popularized the concept of six degrees of separation, which was first introduced in 1929 by the Hungarian writer Frigyes Karinthy in his short story "Chain" or "Lancszemek." In the modern world of ubiquitous emails, research has shown that the flow of emails to and from an employee's can tell a lot about the employee beyond just performing a content analysis of the emails (Stolfo et al. 2006). Besides sender/recipient pairs being identified, additional analysis can identify social networks or cliques of senders and recipients. Many natural pair-wise relationships and cliques can be expected such as people who work in a specific department.

In general, these social networks or cliques are relatively stable over time. What the data miner is trying to discover is unexpected relationships and clique members. Further, in the analysis of the social network, the action of sending emails is arguably more important than receiving emails (Martin et al. 2005). The action of sending an email is markedly more active than receiving. As is well recognized by all that exist in a community facilitated by email exchanges, many emails exchanges are characterized by active interchanges between a limited numbers of participants coupled to a much longer and passive list of email recipients.

Researchers are also exploring the importance of hierarchies in the network where there are *leaders* and *followers*. In addition, the roles of *"middlemailers"* are also being explored. Middlemailers appear to be acting as a conduit between leaders and followers. Identifying these hierarchies provide much more robust results compared to merely identifying pair-wise links. It is particularly import to identify the middlemailers, because a simple *one-step* analysis (analyzing the sender and the recipients included the sender's email) could miss the scope of collusion that is occurring in the company.

In a continuous assurance setting, the roles and responsibilities of email recipients is critical. Within corporate email systems, it is relatively simple to match email recipients to corporate roles and responsibilities. As we discuss above, monitoring of email exchanges with third parties are equally important as internal exchanges. Identifying the role played by a third party may be difficult to immediately identify. Third parties may use corporate email addresses and signatures, but equally they may not. Research is beginning to link publicly available resources, such as Google and newswires, with email-based social network analysis to provide a Web-enabled social network analysis (Culotta et al. 2004). Whilst the quantity and relatively unstructured form of Web-based information is challenging, Culotta et al. (2004) show promising results in an albeit preliminary test of the ability to extract homepages of email correspondents and subsequently build a social network analysis.

Once the networks and cliques are identified, continuous monitoring techniques could be used flag emails that fall outside the established patterns. For example, an email from a high-level executive to a warehouse worker, where no similar emails existed before, could be suspicious. Emails from employees to customer domains where the employee does not hold a position that normally communicates with customers would be suspicious. Unlike the monitoring of volume and velocity, link and network monitoring could be performed on a near real-time basis.

The Challenges of Email Data Mining

Besides being free form and unstructured, emails are noisy, which makes them challenging for data mining. Although email is a written form of communications, senders rarely subject their emails to the same editing scrutiny that they do to formal written (paper-based) communications. As such, there are a variety of reasons for that noise, for example, including:

- Inconsistent use of abbreviations.
- Inconsistent capitalization of words.
- Misspelled words.
- Numbers sometimes spelled-out (e.g., one, two, thirteen) and other times numerical representations are used (e.g., 1, 2, 13).
- Missing and incorrectly used words.
- Incorrect grammar.
- The sender's message frequently includes the prior sender's email as a part of their replying email. Sometimes emails could have the complete discussion tread that includes several generations of replies and replies to replies in the same email.
- Inability to identify the identities of email participants and their relative roles and responsibilities (Elsayed and Oard 2006).

Therefore, in general, any content analysis of email, as opposed to analyzing email logs, will have to be proceeded by significant email data cleaning, which will be a major challenge to attempts to create a system for the *continuous* monitoring of emails. The quoted text (replies to replies) within the emails will have to be removed so only the sender's "new" material in the email is analyzed. Non-text information (e.g., line breaks, extra space, and other control characters) will have to be filtered out. The remaining text will need to be normalized. Tang et al. (2005) recommend a four-pass cascade approach. The first pass is non-text filtering, which is then followed by paragraph normalization, sentence normalization, and word normalization. Then the subsequent content analysis will be both more efficient and effective.

This email cleaning that appears to be a required prerequisite to analyzing email content probably means that email content (as opposed to email logs) will not be analyzed on a true real-time basis. Instead, like the FBI's *Carnivore*, copies of emails will be temporarily stored offline. Offline emails will be then be cleaned and analyzed. This will not be a major problem, meaning that a few minutes will elapse before a suspicious email is flagged and reported to an administrator or auditor.

Textual Data Sources other than Email.

These techniques that are maturing for email data mining can also be applied to other unstructured data external to the company such as blogs, financial forums (e.g., Yahoo Financials), interviews appearing in newspapers, magazines, etc.[6]

Companies should be concerned that employees are inadvertently or purposely distributing proprietary information or information that could violate Reg FD (Regulation Full Disclosure), which is supposed to curb selective disclosure of material nonpublic information by public companies. There is also concern that employees or non-employees are stating false information about the company either to hurt the company's image or to artificially manipulate the company's stock price.

Continuous monitoring of the various external data sources will be challenging. Many companies are already manually monitoring these sources on a periodic basis. There are also services that provide this monitoring. For example, Google Alerts automatically sends emails to a person when there are new Google results for the search terms the person selected.

[6] See Agrawal et al. (2003) for a discussion of data mining of newsgroups.

To experiment with textual mining of newsgroups groups.google.com provides a rich source of messages with 20 years of Usenet archives that includes over 700 million messages.

III. Email Mining Tools

In this section, we introduce tools developed in the research community as well as introduce industrial solutions to provide a flavor of current and future solutions to continuously monitor emails.

Email Content Monitoring

A number of commercial solutions have been developed in recent years to allow corporate security managers to monitor the content of emails. The first class of tools is general purpose network monitoring tools that have been developed for the purposes of security monitoring and assessment. Increasingly vendors are providing email content monitoring as a by-product of spam or spyware assessment. For example, eSoft Corporation's[7] ThreatWall manages virus and spam coming into the entity but also undertakes content filtering "by scanning all emails for admin-defined keywords, phrases or regular expressions." The software emails violations to administrators. A similar feature is found in Symantec's Symantec Mail Security 8x00 Series appliances, which combine hardware and software in a single device. As can be imagined, given the provenance of Symantec, the primary focus of the appliance is on virus defense and spam avoidance, additional plug-ins for content monitoring are also available.

The second class of tools explicitly monitors emails and other Internet communications. At present these tools are largely designed to prevent losses of IP or breaches of compliance requirements. The latter issue is particularly important for organizations that are subject to intensive privacy compliance requirements. The requirements under HIPAA for health care providers to maintain the privacy of their clients is an example of such an important compliance requirement. Vericept

[7] www.esoft.com

Corporation's[8] Vericept Content 360º tool is an example of this relatively new class of software. According to Vericept, "Vericept's Content 360° visibility provides early detection of impending threats and insider risk by monitoring all Internet-based communication and identifying areas of immediate financial, reputation and legal risk. By correlating events and analyzing patterns of behavior, Vericept can help organizations to identify an incident before it occurs and take immediate action to protect against serious security breaches that can cause irreversible brand and reputation damage." An interesting tool within the Vericept suite is "Email Vericept Self-Compliance," which allows the sender of the email to identify the email as being appropriate. Taking a somewhat different approach, Reconnex Corporation[9] has developed "iGuard Appliance" that scans networks, including emails, for sensitive data. The focus of the product is the protection of intellectual property and compliance.

A relatively new startup that has specifically addressed email monitoring is InBoxer, Inc.[10] Whilst the first products of the corporation were in the anti-spam domain, more recently the corporation has developed its "Anti-Risk Appliance." According to InBoxer the appliance draws from the corporation's "proprietary, sophisticated "language models" based on the way words are commonly used in order to identify patterns in text. Our technology comes from years of experience in the speech recognition industry. We learned how to distinguish between words that sound alike by analyzing the entire message." Interestingly, InBoxer archives and indexes all email within the corporation, allowing subsequent searches for forensic or other purposes straightforward.

Social Network Analysis and Email Monitoring

In an earlier section, we discussed the important role that social network analysis develops plays in email monitoring. Whilst the products have made progress in identifying email textual patterns, they do not yet seem to have moved to understand the social networks that underpin the emails that flow within the corporation. Indeed, in the

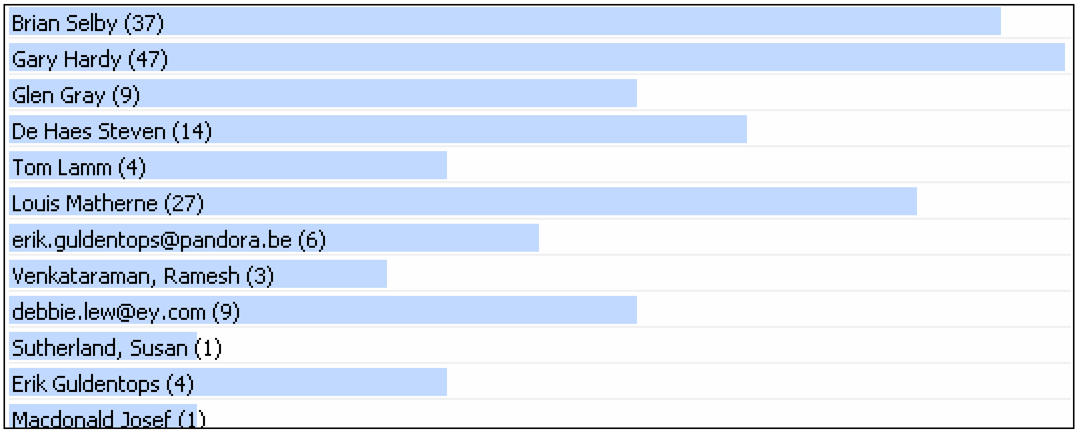
[8] www.vericept.com/solutions/control.asp

[9] www.reconnex.net

[10] www.inboxer.com

research community there are only a handful of tools that explicitly links social network analysis and emails.

A research group within Microsoft Corporation has created a prototype tool, SNARF, to allow individuals to assess their read and unread email using social network analysis. Running as an add-in to Microsoft Outlook, SNARF[11] undertakes filtering and mining of emails based on a variety of social metrics. These metrics include emails sent as well as copied to the email user from recipients in the database and the reverse and the read status of emails. Email from a sender sent rather than copied to a user and which are consistently read will be more important in the social network than those that are copied and unread. SNARF builds and maintains the social network database and adds emails to the database in real-time. The email user can display multiple panes or views of the database. The user can adjust the pane to screen and sort the email according to a variety of criteria. Figure 2 shows a pane that displays emails in the database that were sent or copied to a user within the last thirty days sorted by emails sent or copied over the last year.



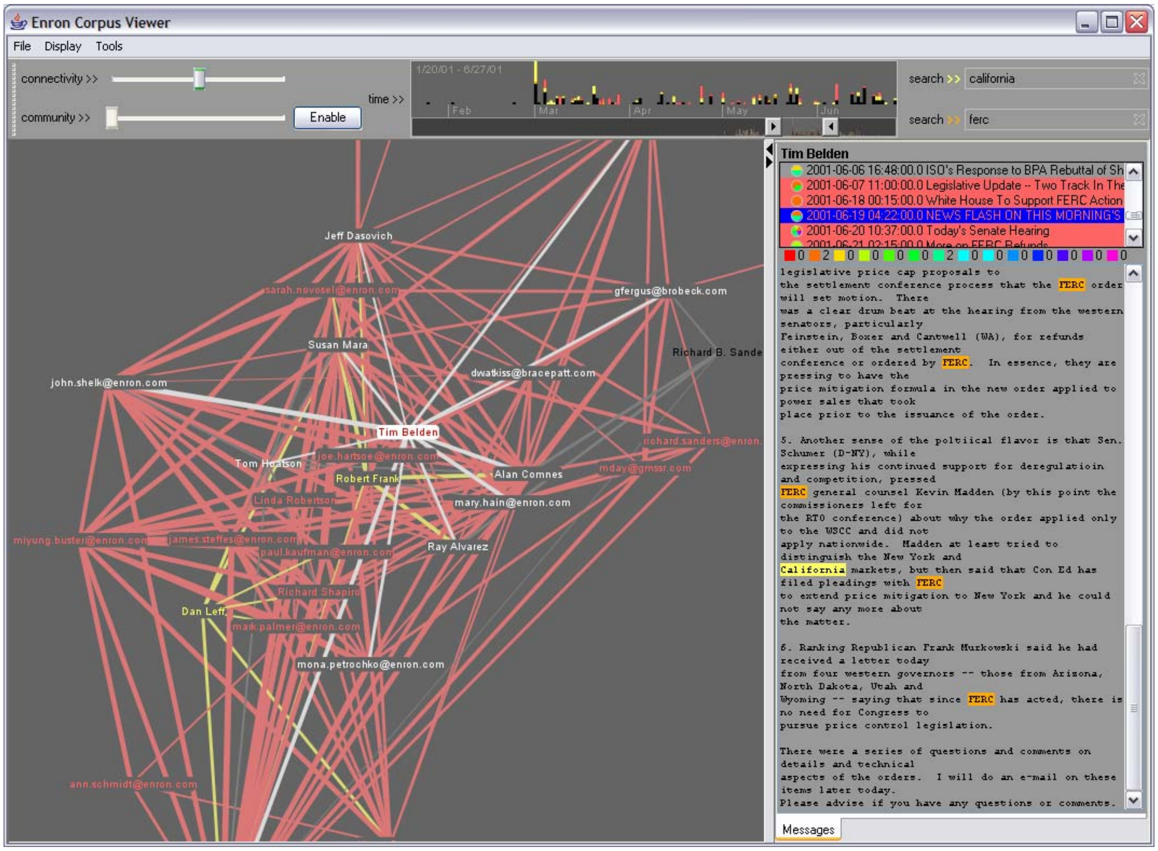
Brian Selby (37)
Gary Hardy (47)
Glen Gray (9)
De Haes Steven (14)
Tom Lamm (4)
Louis Matherne (27)
erik.guldentops@pandora.be (6)
Venkataraman, Ramesh (3)
debbie.lew@ey.com (9)
Sutherland, Susan (1)
Erik Guldentops (4)
Macdonald Josef (1)

Figure 2: SNARF Snapshot

SNARF is a simple email tool that builds on the social networks embedded in the email. The tool demonstrates how social networks can be relatively simply developed from a corpus of emails.

[11] See http://research.microsoft.com/community/snarf/

A rapidly growing area of research in social network analysis is the development of tools for visual analysis of network relationships, allowing users to interactively determine relations between email recipients in an email corpus. Heer[12] has built a variety of tools that allow visual representations of several types of social networks including Friendster relationships (Vizter) and relationships expressed in Enron emails (Enronic). These tools are based on natural language processing techniques. Figure 3 illustrates a visual representation of the social networks embedded in the Enron corpus.



Tim Belden

2001-06-06 16:48:00.0 ISO's Response to BPA Rebuttal of Sh
2001-06-07 11:00:00.0 Legislative Update -- Two Track In The
2001-06-18 00:15:00.0 White House To Support FERC Action
2001-06-19 04:22:00.0 NEWS FLASH ON THIS MORNING'S
2001-06-20 10:37:00.0 Today's Senate Hearing
2001-06-21 02:15:00.0 More on FERC Refunds

🟥 0 🟧 2 🟨 0 🟩 0 🟩 0 🟩 0 🟩 2 🟦 0 🟦 0 🟦 0 🟦 0 🟪 0 🟪 0 🟪 0

```
legislative price cap proposals to
the settlement conference process that the FERC order
will set motion.  There
was a clear drum beat at the hearing from the western
senators, particularly
Feinstein, Boxer and Cantwell (WA), for refunds
either out of the settlement
conference or ordered by FERC.  In essence, they are
pressing to have the
price mitigation formula in the new order applied to
power sales that took
place prior to the issuance of the order.

5. Another sense of the poltiical flavor is that Sen.
Schumer (D-NY), while
expressing his continued support for deregulatioin
and competition, pressed
FERC general counsel Kevin Madden (by this point the
commissioners left for
the RTO conference) about why the order applied only
to the WSCC and did not
apply nationwide.  Madden at least tried to
distinguish the New York and
California markets, but then said that Con Ed has
filed pleadings with FERC
to extend price mitigation to New York and he could
not say any more about
the matter.

6. Ranking Republican Frank Murkowski said he had
received a letter today
from four western governors -- those from Arizona,
North Dakota, Utah and
Wyoming -- saying that since FERC has acted, there is
no need for Congress to
pursue price control legislation.

There were a series of questions and comments on
details and technical
aspects of the orders.  I will do an e-mail on these
items later today.
Please advise if you have any questions or comments.
```

Figure 3: Enronic Interface

[12] jheer.org

IV. MINING THE ENRON DATABASE

The Enron Database

A large corpus of emails from Enron was put into the public domain by the Federal Energy Regulatory Commission following its investigation of the corporation in relation to alleged manipulation of “electricity and natural gas markets in California and other Western states in 2000 and 2001.” The original corpus contained 0.6m emails from 158 users (Klimt and Yang 2004). These users included key participants in the events that brought down the corporation in 2001 including Messrs. Lay, Skilling, Fastow and Causey. Klimt and Yang (2004) note that there were many repeated emails in the corpus. Elimination of these duplicates reduced the corpus to 0.2m messages; an average of 757 messages per user. They also worked to remove quotations of previous emails in subsequent emails in that thread.

Investigating the Enron Database

The Enron email corpus has subsequently been transformed into a relational database format and published in MySQL format (Shetty and Adibi 2004). Figure 4 shows the database schema structure of the Enron database.

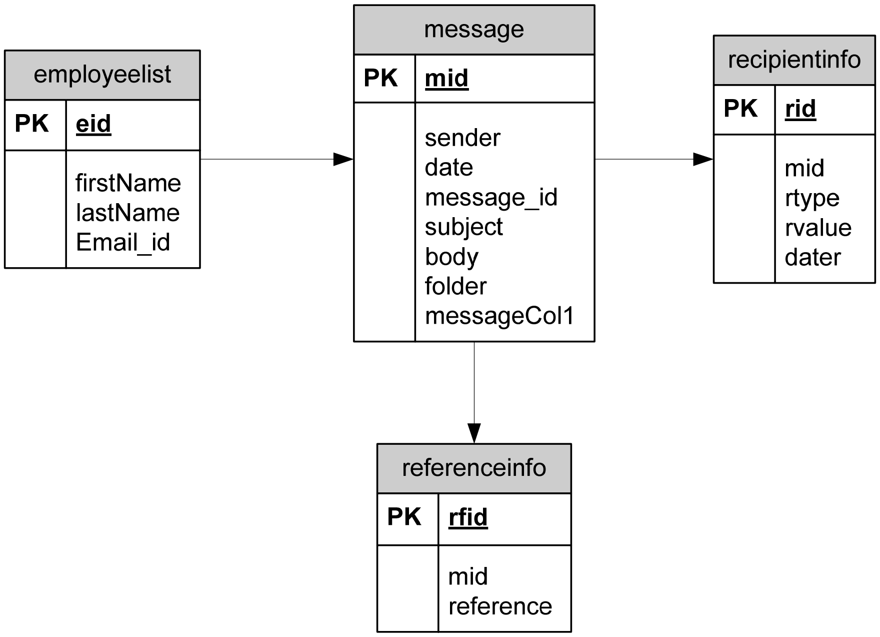


Figure 4: Enron Email Database

Exploring the emails may be undertaken by targeted analysis based upon known patterns, or by automated techniques. Some examples of the former approach generate interesting results, albeit undertaken with hindsight. For example, emails to Richard Causey, the Enron Controller, relating to Raptor, the codeword for Special Purpose Entities associated with Andrew Fastow, the CFO, provide a fascinating insight into the debates within the corporation. The SQL query was:

```
SELECT m.date, m.sender, r.rvalue, m.subject, m.body FROM recipientinfo
r, message m WHERE r.mid = m.mid AND r.rvalue LIKE '%causey%' AND
m.body LIKE '%raptor%';
```

On 31 July 2000, David Delainey, the CEO of Enron North America and Enron Energy Services emailed Richard Causey about Raptor:

> Rick, we have re-examined the portfolio as a result of our last meeting and have an expected portfolio of $500M (gross) or $408M (net-ENA s share) to be placed into Raptor on September 1. We have a second tranche of $608M (gross) or $396M (net ENA s share) to be placed into Raptor 2 on December 1. Kafus, Ecogas, Brigham, Crown Energy and Merlin CLO Equity Option will not be part of either Raptor 1 or Raptor 2. Please confirm that this is consistent with you view of capacity and timing. Regards Delainey

In December of that year, Delainey again emailed Richard Causey about Raptor, pointed to some of the problems with both Raptor I and III:

> Please note that the Raptor I credit capacity is at $(6.2) million due mainly to Catalytica which is now a publicly traded stock. Raptor 3 also has a negative credit capacity.

And then, as Enron began to unravel, in a thread entitled "Raptor Debris,' Delainey once again communicated with Causey about Raptor in October 2001:

> Now that Raptor is blown up, should we begin valuation efforts of assets and include in merchant portfolio. I am not familiar with details of how it unwound so I don t know what we are left with. Is this worth a meeting with you or your designers? Rick

Later, Delainey once again communicated with Causey at the end of the thread, which discussed what was left with as the Raptors unraveled:

> The quarterly valuations for the assets hedged in the Raptor structure were valued through the normal quarterly revaluation process. The business units, RAC and Arthur Andersen all signed off on the initial valuations for the assets hedged in Raptor. All the investments in Raptor were on the MPR and were monitored by the business units and we prepared the Raptor position report based upon this information. The MPR report now reflects that the hedged assets are not in Raptor. Please contact me if you have any questions. Thanks, Gordon
>
> -----Original Message-----From: Baker, Ron Sent: Tuesday, October 09, 2001 12:55 PMTo: Glisan, Ben; Causey, Richard; Buy, Rick; Butts, Bob; Colwell, WesCc: Gorte, David; Mckean, George; Mckillop, GordonSubject: RE: Raptor Debris Ben, I met with Rick Buy and several members of his group last week to discuss the buyout and what the remaining exposures to Enron were. As the Raptor DPR s were generally taking hedged values directly from ENA s MPR, they were going to discuss the valuation on those assets with Dick Lydecker. I haven t heard anything else from RAC since that meeting. Let me know if there is anything I can do to further this process. Thanks, Ron
>
> -----Original Message-----From: Glisan, Ben Sent: Tuesday, October 09, 2001 12:46 PMTo: Causey, Richard; Buy, Rick; Baker, Ron; Butts, Bob; Colwell, WesCc: Gorte, David; Mckean, George; Mckillop, Gordon Subject: RE: Raptor Debris George & Gordon Please work with RAC (Dave Gorte) to ensure that all of the Raptor Investments are being monitored. Ben
>
> -----Original Message-----From: Causey, Richard Sent: Wednesday, October 03, 2001 11:35 AMTo: Buy, Rick; Glisan, Ben; Baker, Ron; Butts, Bob; Colwell, WesCc: Gorte, David Subject:

> RE: Raptor Debris I think that all investments that were hedged in Raptor are on th MPR and are being monitored and reported against. Is this true? Ron, will you take the lead in making sure we have a smooth transition on this? Thanks
>
> -----Original Message-----From: Buy, Rick Sent: Wednesday, October 03, 2001 9:38 AMTo: Causey, Richard; Glisan, BenCc: Gorte, David Subject: Raptor Debris Now that Raptor is blown up, should we begin valuation efforts of assets and include in merchant portfolio. I am not familiar with details of how it unwound so I don t know what we are left with. Is this worth a meeting with you or your designers? Rick

Emails also provide a flavor of the relationship between the external auditors and Enron. A total of 466 emails in the corpus were sent from Enron email addresses to Andersen email addresses. A surprising proportion of these emails were comprised of jokes often in poor taste, event announcements and personal mail. There were 33 emails from Andersen email addresses to Enron email addresses. Again, many of these emails were related to matters other than the audit. An example of a more substantial email is a 2001 email from Andersen staffer Tatiana Waxler to an Enron staffer Stacey White, setting out the key directions for the audit of the Power Trading Unit, one of the more important elements of Enron:

> Stacey, As per our conversation today, enclosed please find a list of topics we would like to discuss for our Power Trading Audit: 1. Understand Nature of Portfolio? Nature of portfolio, risk factors, correlation between curves, calculation of correlation factors, off-peak trading, liquidity of portfolio, etc. 2. Understand Performance Metrics? Performance metrics utilized for traders, benchmarking trader performance, results of inadequate performance, bonus determination process. 3. Organizational Structure 4. Overall Market/ Newly Traded Markets? Traders? assessment of the overall market, including most active regions, liquidity of peak versus off-peak positions, new markets, etc? Extent of autonomy of the traders with implementing trading strategy (i.e., whether traders can implement own trading strategies)? Reports that are available regarding the number of trades executed on a monthly or quarterly basis? Changes in management reporting process from prior year (i.e., book administrator rolls, Schedule B & C, etc.)? Extent of use of ExPower (if still utilized)? Extent of exotic deals valued outside of EnPower and the related controls surrounding deal capture, valuation, management reporting, and financial reporting? Limits placed on trader activity (including position limits, VAR limits, etc.) by the commodity group and the monitoring of those limits, if applicable. 5. Ancillary Services? Type of ancillary services provided to customers and the extent of ancillary service trading. Discuss what governmental regulations or ISO regulations exist for ancillary services (i.e., specified rates,

> market caps, etc.), and the impact on the curve. Understand why the curve typically does not change for ancillary services. 6. Monitoring of Long-Term Deals? Long-term deals monitoring by risk management, and reports provided to management regarding the status of long-term deals? Status of NSTAR and other significant long-term deals of which we have reviewed for proper booking. 7. Monitoring of Requirements Basis Deals (UI, NSTAR, etc.)? Monitoring requirements of basis deals with risk management and document? Nature of changes made, and the reasonableness of the current booking of the deal compared to current available information. Please let us know at your earliest convenience if you are the right person to talk about these topics. If not, please direct us to the correct individual(s).Thank you for all your help. Tatiana

The emails shown above provide a flavor of the rich dataset that the Enron email corpus provides and which is no doubt replicated throughout organizations. These emails were retrieved by drawing on known relationships (e.g. Andersen to and from Enron staffers, emails to senior administrators using known keywords etc.) In a continuous assurance environment it will be unrealistic to expect that manual searches of the type described above could be efficient or effective.

Research on automated understanding of the emails within the Enron corpus is still at a relatively early stage. A considerable amount of research has gone into understanding social network relationships within email data sets. Whilst this is an important first step in identifying key relationships that can be used for assurance purposes, social network analysis does not on its own allow identification of emails that may be critical to an assurance issue. As discussed above, there are also a variety of techniques to analyze the text within a corpus including natural language processing and machine learning. The next step is to match content analysis and social networks. McCallum et al. (2005) introduces the Author-Recipient-Topic (ART) which is a Bayesian model that links sender, recipient and text. McCallum asserts that the nature of relationship between sender and recipient can be determined from an analysis of the nature and volume of email interactions and from the text. McCallum have taken a subset of the Enron corpus and further cleaned the database to identify individuals with multiple email addresses.

The ART model is capable of identifying clusters of emails within subsets of the overall social network. For example, McCallum et al. (2005) identifies groups of emails clustered around what the authors describe as “Legal Contracts,” “Operations” etc.

Particular action words are identified within each of these natural groups. For example, the terms “section,” “party and “notice” appear in emails within the Legal Contracts grouping and “gas,” “business” and Houston” within the Operations grouping. The authors show that top emailers within each of these groups have some natural affinity to the topics.

McCallum et al. (2005) notes that individuals, such as executives and their personal assistants, may have similar social networks but are likely to have very different roles and employ markedly different communication styles. The email monitor must distinguish these two classes of users.

The research on Social Network Analysis and Content Analysis has yet to fully address the particular needs of continuous assurance. In assurance models, we are interested in outliers as much as we are with general directions. The reality of many types of fraud, for example, is that they are conducted by parties that may be bound by normal patterns of interaction but are also conducting fraudulent co-operation. The assurance monitor that links social network analysis and text must also identify atypical patterns of conversation.

V. CONCLUSIONS

In this paper, we open a new thread of research in continuous assurance—continuous monitoring of emails within the corporate environment. As the Enron email corpus demonstrates, emails are likely to contain exchange of information between parties that will provide evidence and context for matters that are subject to assurance. Notably these matters include evidence of a wide range of managerial fraud including inappropriate related party transactions, channel stuffing, and improper revenue recognition. There are other matters of interest to assurance providers, particularly internal auditors that may be embedded in emails. These include avoiding breaches of compliance requirements such as HIPAA as well as the protection of corporate assets such as the intellectual property embedded within valuable research in a knowledge intensive corporation (e.g. pharmaceutical industry).

Emails are forms of semi-structured data, with key fields for sender, recipients, dates, topics, text and attachments. As emails pass through email gateways, it is a relatively simple matter to archive, index, and categorize emails, and monitor key words. As vendors of software that monitor emails for breaches of compliance requirements demonstrate, it is feasible to monitor incoming and outgoing email in something approaching realtime. Early research results demonstrate the considerable potential of email monitors to identify key emails of interest. The availability of commercial monitoring software also suggests appropriate functionality, although this has not been tested by researchers. It would be interesting to learn of the extent of Type I and Type II errors for this class of software.

There are many issues with monitoring email. As the Enron emails also demonstrated, emails are noisy. The Enron database we used was the third generation of that database. The first version that was released by Federal Energy Regulatory Commission was almost unusable to do any deep analysis beyond simple key-word searches. Researchers at MIT made the first cut at cleaning the Enron email. The database we used was a subsequent cleaning and structuring of the emails by two researchers at USC. There are other versions of the corpus that have been subjected to manual taxonomic analysis. These human interventions do not provide a foundation for continuous assurance. Automated data cleaning would seem to be an essential pre-requisite for email monitoring.

It may be difficult to uniquely identify email parties if users employ aliases or different email addresses. Links of email addresses to corporate positions and the interrelationships of those positions will often not be readily able to be incorporated into a social network analysis. New email text will be added to existing text as a thread is developed. Content analysis tools must recognize threads and eliminate the inherent double counting that come from repeated texts. Key information will often be in attachments, which will need to be content analyzed in a different fashion than the more casual nature of email text.

However, these commercial software packages do not perform any of the deeper analysis that is still being performed in research settings, such as, changes in volume and velocity, deception parameters (first-person pronouns, exclusive words, negative emotion words, and action verbs), and comparing writeprint features. Continuous monitoring of these types of metrics is also going to be particularly challenging because these metrics require cleaned data, which will slow the overall detection process.

The continuous monitoring of textual resources in general and email in particular is a new area of inquiry in the field of continuous assurance and audit. With email being at the center of corporate DNA, mining and monitoring of emails will be an important area for future research and there are many possible areas for future research. Here we suggest a few areas for productive investigation. First, there is a need to bring together the existing work on social networks with the examination of textual patterns. There are several areas of interest to auditors that are of perhaps lesser interest in other domains. Auditors are often interested in outliers rather than in the overall direction of an email corpus. The emails of interest are the relatively few emails that may hypothetically go between members of senior management and related parties that are indicative of high level financial statement fraud. Identifying such emails will require developing sophisticated understanding of social relationships that

Second, there will be a need to link social network analysis with the control environment. Auditors are often interested in control overrides. Roles and responsibilities will not be embedded in the email corpus and linking of control databases to the email corpus will be important. Further, going beyond the corporation, it will be important to understand the social network relationships that may presage kickbacks or collusion between executives and key third parties. Research on understanding social networks in something approaching realtime would seem to be a challenging research area.

Third, much investigation will be required to understand the particular email of footprint of typical frauds or control overrides. What textual patterns are indicative of fraud? What tools are the most productive to assess these textual patterns? Can existing tools operate in the continuous arena?

Fourth, what are the impediments to cleaning email? Are emails dumps such as that originating from the FERC investigation of Enron significantly more messy than those that would be found within a corporate environment?

Fifth, what are the privacy and policy issues of continuous assessment of emails?

Sixth, what are the views of internal and external auditors and compliance officers on monitoring of emails? What work is already going on in organizations? How do they see this potential tool assisting them in their work?

Seventh, at a more prosaic level, what lessons can be learned from the existing commercial investment in monitoring tools? It is interesting that there appears to be much more activity in the area of continuous monitoring of emails in the vendorspace than in the academy.

Research in this area is also relatively new, and there are few corpi that allow systematic assessment of alternative approaches and tools.

REFERENCES

Agrawal, R., S. Rajagopalan, R. Srikant, and Y. Xu. 2003. Mining newsgroups using networks arising from social behavior. Paper read at 12th international conference on World Wide Web, at Budapest, Hungary.

Alles, M., G. Brennan, A. Kogan, and M. Vasarhelyi. Forthcoming. Continuous monitoring of business process controls: a pilot implementation of a continuous auditing system at Siemens. *International Journal of Accounting Information Systems*.

Baayen, H., Halteren, H., Neijt, A., and Tweedie, F. 2002. An experiment in authorship attribution. In *6th International Conference on the Statistical Analysis of Textual Data (JADT 2002).*

Burrows, J. F. 1992. Word patterns and story shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing* 2 (1):61-67.

CICA/AICPA. 1998. *Continuous Auditing*. Toronto: Canadian Institute of Chartered Accountants.

Culotta, A., R. Bekkerman, and A. McCallum. 2004. Extracting social networks and contact information from email and the Web. In *First Conference on Email and Anti-Spam - CEAS 2004*. Mountain View, California USA.

de Vel, O., A. Anderson, M. Corney, and G. Mohay. 2001. Mining E-mail content for author identification forensics. *SIGMOD Record* 30 (4):55-64.

Debreceny, R. S., G. L. Gray, W.-L. Tham, K.-Y. Goh, and P.-L. Tang. 2003. The Development of Embedded Audit Modules to Support Continuous Monitoring in the Electronic Commerce Environment. *International Journal of Auditing* 7 (2):169-185.

Elsayed, T., and D. W. Oard. 2006. Modeling Identity in Archival Collections of Email: A Preliminary Study. Paper read at Third Conference on Email and Anti-Spam CEAS 2006, July 27-28, at Mountain View, California USA.

Groomer, S. M., and U. S. Murthy. 1989. Continuous Auditing of Database Applications: An Embedded Audit Module Approach. *Journal of Information Systems* 4 (1):53-69.

Holmes, D. I. 1998. The evolution of stylometry in humanities. *Literary and Linguistic Computing* 13 (3):111-117.

Keila, P. S., and D. B. Skillicorn. 2005. Detecting Unusual Email Communication. Belfast: Queens University.

Klimt, B., and Y. Yang. 2004. Introducing the Enron Corpus. In *First Conference on Email and Anti-Spam - CEAS 2004*. Mountain View, California USA.

Li, J., R. Zheng, and H. Chen. 2006. From Fingerprint to Writeprint. *Communications of the ACM* 49 (4):76-82.

Liu, H., and H. Motoda. 1998. *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA: Kluwer Academic Publishers.

Martin, S., A. Sewani, B. Nelson, K. Chen, and A. D. Joseph. 2005. Analyzing Behaviorial Features for Email Classification. In *Second Conference on Email and Anti-Spam CEAS 2005*. Berkeley, CA: University of Caiifornia at Berkeley.

McCallum, A., A. Corrada-Emmanuel, and X. Wang. 2005. A Probabilistic Model for Topic and Role Discovery in Social Networks and Message Text. Amherst, MA: University of Massachusetts.

McEnery, A., and M. Oakes. 2000. *Authorship Studies/Textual Statistics*: Marcel Dekker.

Milgram, S. 1967. The small world problem. *Psychology Today* 2 (60-67).

Minkov, E., and W. W. Cohen. 2006. An Email and Meeting Assistant using Graph Walks. In *Third Conference on Email and Anti-Spam CEAS 2006*. Palo Alto: ACM, 14-20.

Pennebaker, J. W., M. E. Francis, and R. J. Booth. 2001. *Linguistic inquiry and word count (LIWC)*: Erlbaum Publishers.

Rudman, J. 1998. The state of authorship attribution studies: some problems and solutions. *Computers and the Humanities* 31:351–365.

Shetty, J., and J. Adibi. 2004. The Enron email dataset database schema and brief statistical report. Los Angeles, CA: University of Southern California, Information Sciences Institute.

———. 2005. Discovering important nodes through graph entropy the case of Enron email database. Paper read at 3rd International Workshop on Link Discovery, Conference on Knowledge Discovery in Data.

Skillicorn, D. B. 2005. Beyond keyword filtering for message and conversation detection. Paper read at IEEE International Conference on Intelligence and Security Informatics (ISI2005), May.

Stamatatos, E., N. Fakotakis, and G. Kokkinakis. 2001. Automatic text categorization in terms of genre and author. *Computational Linguistics* 26 (2):471–495.

Stolfo, S. J., G. Creamer, and S. Hershkop. 2006. A temporal based forensic analysis of electronic communication. Paper read at 2006 national conference on Digital government research.

Tang, J., H. Li, Y. Cao, and Z. Tang. 2005. Email Data Cleaning. In *5th International Conference on Knowledge and Data Discovery KDD'05*. Chicago, Illinois, USA: ACM Press, 489-498.

Vasarhelyi, M., A. M, and A. Kogan. 2004. Principles of Analytic Monitoring for Continuous Assurance. *Journal of Emerging Technologies in Accounting* 1 (1-21).

Vasarhelyi, M. A. 2002. Concepts in Continuous Auditing. In *Research Accounting as an Information Systems Discipline*, edited by S. G. Sutton and V. Arnold. Saratoga: American Accounting Association.

Vasarhelyi, M. A., and J. Peng. 1999. Qualitative corporate dashboards for corporate monitoring. *IS Audit and Control Journal* 5 (Fall):45-48.

Yule, G. U. 1944. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Zheng, R., J. Li, Z. Huang, and H. A. Chen. 2006. A framework of authorship identification for online messages: writing style features and classification techniques. *The Journal of the American Society for Information Science and Technology* 57 (3):378–393.