
Using PLSI-U to Detect Insider Threats by Datamining Email

J.S. Okolica

G.L. Peterson

R.F. Mills

Air Force Institute of Technology

AFIT/ENG, BLDG 641 RM 220

2950 Hobson Way

Wright Patterson AFB, OH 45433-7765

E-mail: {james.okolica, gilbert.peterson, robert.mills}@afit.edu

Abstract: Despite a technology bias that focuses on external electronic threats, insiders pose the greatest threat to an organisation. This paper discusses an approach to assist investigators in identifying potential insider threats. We discern employees interests from e-mail using an extended version of PLSI. These interests are transformed into implicit and explicit social network graphs, which are used to locate potential insiders by identifying individuals who feel alienated from the organisation or have a hidden interest in a sensitive topic. By applying this technique to the Enron e-mail corpus, a small number of employees appear as potential insider threats.

Keywords: Probabilistic Latent Semantic Indexing (PLSI); Insider Threat; Datamining; Social Networks; Large Dataset.

Reference to this paper should be made as follows: Okolica, J.S., Peterson, G.L., and Mills, R.F. (2006) 'Using PLSI-U to Detect Insider Threats by Datamining Email', International Journal Of Security and Networks (IJSN), Vol. 1, Nos. 1/2/3, pp.64-74.

Biographical Notes: James Sean Okolica is a Captain in the United States Air Force currently stationed at the Air Force Institute of Technology earning his Masters of Computer Science Degree. His research interests include digital forensics, machine learning and pattern recognition.

Gilbert L. Peterson is an Assistant Professor of Computer Engineering at the AFIT. Dr. Peterson holds a BS degree in Architecture, an MS in Computer Science, and a PhD in Computer Science from the University of Texas at Arlington. He teaches and conducts research in digital forensics, and artificial intelligence.

Robert F. Mills is an assistant professor of electrical engineering at the Air Force Institute of Technology (AFIT). He holds a BS degree in electrical engineering from Montana State University, an MS in electrical engineering from AFIT, and a PhD in electrical engineering from the University of Kansas. He teaches and conducts research in computer security, network management, and communications systems.



1 Introduction

Insiders are members of an organization who often have a legitimate right to the information that they are accessing. However, they abrogate the trust they have been given by using the information for illegitimate reasons. Once an insider attack has occurred, finding the culprit as quickly as possible is critical. From a population that can number in the tens of thousands, investigators must quickly reduce the suspects to a number for which they have sufficient investigators.


One of the best indicators of a person's interests in today's organizations is their email traffic. Through datamining the organization's email, topics of interest can be extracted and people categorized by those topics they are most interested in. By finding those individuals who have shown an interest in the relevant topics, the number of investigative leads is reduced. Especially likely suspects are people who have shown an interest in the topic but have never communicated that interest with anyone within the organization. These people either have a secret interest in the topic or generally feel alienated from the organization and so communicate their interest only outside of it.

A second method for identifying investigative leads is finding individuals who have shown previously undetected warning signs of becoming insider threats. One warning sign is when an individual begins to separate himself from the organization and feels alienated by it. When this occurs, individuals will cease socializing with others within the organization and instead look for social opportunities externally. What these two methods have in common is that in both cases, investigators are looking for individuals who have hidden their interests from their co-workers. In the first case, this is an interest in the sensitive or classified topic and in the second case this is an interest in socializing.

In this paper, Probabilistic Latent Semantic Indexing (PLSI) Hoffman (1999) is expanded to include users and then used on the Enron email corpus to test its applicability on generating insider threat investigative leads. The resulting PLSI-U (PLSI with users) model performs well, creating 48 clear categories and extracting a small number of individuals with clandestine interests as potential insider threats investigative leads.

2 Motivation: Datamining Email to Detect Insider Threats

During a RAND workshop on the Insider Threat Herbig (2002), the first priority for improving the detection of insider's misuse was "developing [user] profiling as a technique" RAND (1999). To develop these profiles, the workshop participants proposed using: files and processes normally accessed, periods of time that a user is logged in, and keystroke patterns. By comparing old profiles with current ones, anomalies (e.g. use of administrator or logging commands) are better detected RAND (1999). While this is successful if there is historical data to compare to, the amount of history that is needed is overwhelming. One alternative to the



use of these audit logs is to develop these profiles using existing sources of data. One such data source is email.

Electronic mail is fast becoming the most common form of communication and in 2006 is expected to exceed over 60 billion messages daily Martin (2005). It is one of the best electronic sources of personal information available, especially due to its ease of accessibility in an organization making it an ideal data source for user profiling. While there has been a large amount of research in preventing incoming mail that is deemed suspicious Stolfo (2003), the idea of reading the outgoing mail has not received a lot of activity. This is due in large part to privacy concerns and the lack of large-scale email datasets.

Semantic analysis, i.e. extracting meaning from text, has been directly applied to countering insider threats by Symonenko, et al. Symonenko (2004). They investigated the effectiveness of using natural language processing (NLP) to discover intelligence analysts who were accessing information outside of their community of interest. By using interviews with analysts to acquire significant domain specific knowledge, the researchers were able to use clustering to determine when an analyst was looking at (or producing) reports on areas other than the ones assigned to his group.

While Symonenko, et al.'s success is impressive, it requires a significant amount of up front work to develop the domain specific knowledge. Furthermore once this knowledge is acquired, the resulting model is only applicable to one domain. By contrast, the model described in this paper works without any specific domain knowledge in a much more generalized setting. Probabilistic clustering is applied to email in order to extract an individual's interests. By comparing the interests an individual shares with his co-workers with those he only shares with individuals external to the organization, investigators can uncover individuals who are hiding things from their co-workers. If they are hiding an interest in information that has been stolen or if they are hiding an interest in socializing (i.e. they feel alienated from their co-workers), they are promising investigative leads for potential insider threats.

3 Methodology

This paper examines the potential use of constructing social networks from email activity to generate insider threat leads. The first step is developing user "interest profiles". These profiles are generated through probabilistic clustering algorithms derived from the PLSI-U model. Individuals are considered to have an interest in a topic if their probability of selecting the topic ($p(topic|user)$) is greater than 95% of the population. The profiles are then used in generating an implicit social network between people for each topic. Individuals are connected in the implicit social network for a topic if they have an interest in a topic. A second explicit social network for each topic is then constructed strictly based on the presence of email activity associated with that topic between pairs of individuals (emails are considered associated with a topic if their conditional probability for that topic ($p(topic|email)$) is greater than 95% of the email corpus). If an email is associated with a topic, then the sender and recipients of that email will all be linked together in the explicit network for that topic. Observe that using the subject line of an

email to determine the topic that email is about is problematic for several reasons including missing or vague subject lines (e.g. RE: Hi”) and emails that contain multiple topics only one of which is referenced in the subject line (the issue is avoided in this research by using probabilistic clustering to discover the topic of an email). These two networks are then compared for discrepancies. People who fail to communicate via email for a specific topic (i.e. not connected to anyone according to the explicit social network) but who have shown an interest in that topic (i.e. connected according to an implicit social network) are then considered as possibly having a clandestine interest and worthy of additional investigation. Consider the example in Figure 1 (Implicit Interest Network). By examining Susan’s emails, it emerges that she has an interest in football. However, none of the emails she sent or received *within the company* (Explicit Interest Network) have included anything about football. Therefore, for Susan football is a clandestine interest. By varying the subset of interests that generate the networks (e.g. limiting it to suspicious interests), these clandestine connections become more relevant.

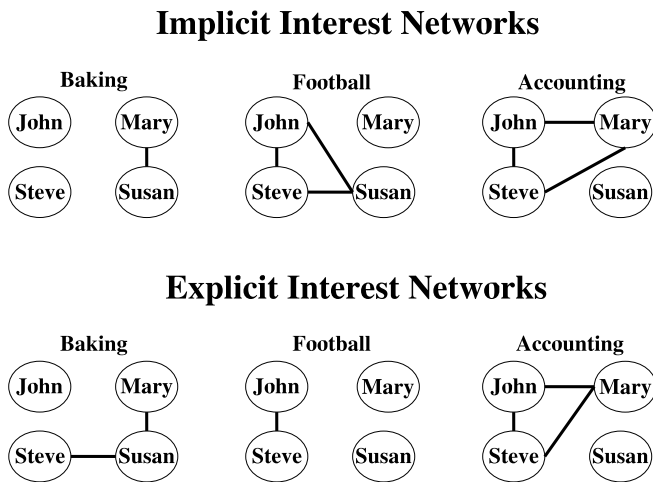


Figure 1 An Example of Clandestine Interests (implicit network = external; internal email explicit network = internal email only).

The first step is to use PLSI-U to cluster the email activity into relevant group interests, or topics. Once the data has been clustered, building the social networks is straightforward. First, an implicit network is constructed from the PLSI-U data. If two people both have an interest in a topic that exceeds a threshold, specifically 95% of the population, a link is created between those two people. Mathematically, if $p(z = Z_1|u = U_1) > \mu + 1.64\sigma$ and $p(z = Z_1|u = U_2) > \mu + 1.64\sigma$ where Z_1 is a topic, U_1 and U_2 are individuals, and μ and σ are the mean and standard deviation of $p(z = Z_1|u)$ for all u , then the link U_1U_2 is created for the implicit PLSI network for category Z_1 . This process is repeated for every pair of people for every topic.

Once the implicit network is formed, an explicit network is created based on email data. If there is at least one email message for a specific topic between two people, a link is created between them. Mathematically, if $p(z = Z_1|d = D_1) > \mu + 1.64\sigma$ where D_1 is an email and μ and σ are the mean and standard deviation of $p(z = Z_1|d)$ for all d , then $\forall U_1 \in D_1 \forall U_2 \in D_1$ the link U_1U_2 is created for the explicit

network for category Z_1 . This process is repeated for every topic and every pair of people.

The final step is to examine the implicit and explicit social networks for each topic. If a person has an interest in a topic (i.e. there are links between that person and others in the implicit network) but has no links to anyone in the explicit network for that topic, that individual is considered to have a clandestine interest in that topic. Figure 2 provides a summary of the Potential Insider Threat Detection Algorithm.

1. Perform Probabilistic Clustering on the email dataset
2. Describe the topics. Topics are described by the 50 most likely words ($p(z|w)$) for this topic.
3. Determine the topics an individual is interested in. Topics must have a probability ($p(z|u)$) greater than 95% of the population.
4. Determine the topics an email contains. Topics must have a probability ($p(z|d)$) greater than 95% of the emails in the corpus.
5. Develop an implicit social network and an explicit social network for each topic. The first network links individuals who share an interest in the topic. The second links individuals who pass an email about that topic between them.
6. Determine who has a clandestine interest in each topic. These individuals have links in the implicit network but do not have any links in the explicit network.
7. Discover potential insider threats. These individuals have an interest in at least one of two types of topics:
 - a. Sensitive Topics – for Enron this is off-book partnerships.
 - b. Socializing – clandestine interests in this topic suggest individuals who feel alienated.

Figure 2 Potential Insider Threat Detection Algorithm

4 Generative Model

This section describes the theoretical background used in developing the statistical model which is then used to predict the likelihood that a specific email is constructed from a specific topic, and consequently is a member of a particular topic.

Notationally, M is the number of emails, $d_{i=1..M}$, in the corpus. There are V words in the vocabulary and each email, d_i , is composed of N_i words, $w_{j=1..N_i}$. Furthermore, there are K topics. For simplicity, each email is considered to have a non-zero probability of each topic, $z_{r=1..K}$. Finally, each email has exactly one sender and one or more recipients. For this paper, the roles of these people are not distinguished (for models where roles are distinguished, see McCallum (2004)) and so each email, d_i , is considered to have L_i people, $u_{s=1..L_i}$, associated with it, drawn from a population of P people.

For simplicity, we use the naive bayes assumption that each topic in an email is conditionally independent of every other topic and that every word and person is conditionally independent of every other word and person conditioned on the topic. Although this assumption is obviously wrong (e.g. “the cat ate the mouse”



is different than “the mouse ate the cat”), techniques that make this assumption still produce good results.

PLSI is a generative model for the creation of a document within a corpus. However, it does not include the concept of people. Therefore, to use PLSI as a generative model for email, the concept of people is incorporated, generating a new model, PLSI with users (or PLSI-U). PLSI-U assumes an email is constructed by first adding a user at a time and then adding a word at a time. Before each word or user is added, a topic is selected from a multinomial distribution and then the word or user is selected conditionally given the topic from a multinomial distribution. What is most desired is the joint probability of a word w_i and user u_s occurring in email d_j which contains topic z_r . However, given the size of the vocabulary, the number of people in the population, the number of words and people in the emails and the number of topics, determining this full joint probability is unrealistic. However, it is sufficient to determine the probability of topic z_r for a specific email. Then by looking at the probabilities for all of the topics, one can determine which topics the email contains (since they will have the greatest probabilities). Therefore, the goal is to determine $p(z_r|d_j)$. However, given the generative model, there is no direct relationship between topics and emails. A topic “produces” words and the collection of words creates the emails. Therefore, in order to determine $p(z|d)$, it is first necessary to consider $p(z|d, w, u)$. Through the use of Bayes Rule, the following equations are derived:

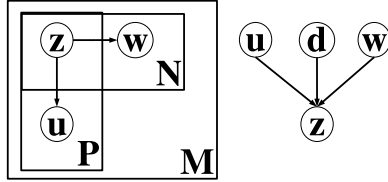


Figure 3 PLSI-User Mixture Model

$$\begin{aligned}
 (1) \quad p(z|u, d, w) &= \frac{p(u|z)p(d|z)p(w|z)p(z)}{\sum_{z' \in Z} p(u|z')p(d|z')p(w|z')p(z')} \\
 (2) \quad p(w|z) &= \frac{\sum_{u \in U} \sum_{d \in D} p(z|u, d, w)n(d, w)}{\sum_{u \in U} \sum_{d \in D} \sum_{w' \in W} p(z|u, d, w')n(d, w)} \\
 (3) \quad p(d|z) &= \frac{\sum_{u \in U} \sum_{w \in W} p(z|u, d, w)n(d, w)}{\sum_{u \in U} \sum_{d' \in D} \sum_{w \in W} p(z|u, d', w)n(d, w)} \\
 (4) \quad p(u|z) &= \frac{\sum_{d \in D} \sum_{w \in W} p(z|u, d, w)n(d, w)}{\sum_{u' \in U} \sum_{d \in D} \sum_{w \in W} p(z|u', d, w)n(d, w)} \\
 (5) \quad p(z) &= \sum_{u \in U} \sum_{d \in D} \sum_{w \in W} p(z|u, d, w)
 \end{aligned}$$

where $n(d, w)$ is the number of times a word occurs in an email. For a derivation, refer to Okolica (2006).

These equations can now form the expectation (eq. 1) and maximization (eq. 2, eq. 3, eq. 4, eq. 5) equations for Expectation-Maximization (EM). EM alternates two steps:

1. Assign random probabilities to $p(d|z)$, $p(w|z)$, $p(u|z)$, and $p(z)$ such that they produce probability distributions (i.e. the probabilities are all non-negative and sum to one).
2. Calculate all of the values for $p(z|u, d, w)$.
3. Using the values from step 2, calculate the new values of $p(d|z)$, $p(w|z)$, $p(u|z)$, and $p(z)$.
4. Repeat steps 2 and 3 until convergence.

5 Results

For this paper, the Enron corpus was used as data. It is the only large corpus of real-world email traffic that is available. As part of their investigation into Enron, the Federal Energy Regulatory Commission (FERC) seized Enron’s email and made a portion of it publicly available. While it only includes the email folders of 151 employees, it still contains over 250,000 email messages. Furthermore, due to the number of individuals the emails were sent to, the resulting corpus has sufficient data on over 34,000 Enron employees. In addition to being valuable for the prosecution of the case against Enron’s senior management, this data has become a touchstone of research into email data mining techniques. These particular experiments do not use Enron as a case study; instead it is simply used as a “proof of concept”. As such, the Enron email corpus is used as data and only a small effort is made to uncover the principal actors involved in the Enron scandal. Due to the large size of the data, each iteration of the EM algorithm is implemented in parallel with each topic occurring at the same time. Because only 16 machines are available on the server cluster used and memory on each server is only sufficient to run three topics, the total number of topics selected a priori is 48. This number concurs with previous research done by McCallum, et al. McCallum (2004) who found 50 as the appropriate number of topics. After running the algorithm, the data consistently converged to a mean square error (MSE) of less than 1×10^{-5} percent prior to 80 iterations. As a result, 80 was selected as a sufficient number of iterations.

Two separate experimental runs are performed. The first only included words that are in the dictionary. The second includes all words, allowing organization-specific slang and acronyms to be included as well as proper names. To reduce the number of words in the corpus, all of the words are stemmed (e.g. baking, baker, and baked are all combined with the stemmed word, bake). Some of the words from the resulting categories are shown in Figures 4 and 5. The words shown are those that had the highest conditional probability given the topic (i.e. $p(w|z)$). Although complete words are shown, they have been extrapolated from the word stems actually produced. Despite initial concerns that stemming might make some of the words difficult to determine (e.g. trying to determine the original word family that stemmed to ‘thi’), the stemmed words that distinguish categories prove easy to identify. In order to produce a list that exclude common, non-distinguishing words, only words that appear in at most 5 categories are used to define a category. The first topic, Senior Mgmt, was generated by observing the preferred topics of Ken Lay (Enron’s Chairman), Jeff Skilling (Enron’s CEO) and Andy Fastow (Enron’s

CFO) and selecting the one most common to all of them. It clearly gives the reader the sense of a senior management topic. It is interesting to observe that although in the first experiment only words found in a dictionary are included, at least one name seeped through because its stemmed base is the same as the stemmed base of a word in the dictionary (Kenneth Lay's first name, ken, is a word in the dictionary). Unlike the Senior Mgmt topic, the California Crisis topic emerges strictly by examining the most probable words. Despite this, the topic emerges clearly. The Research topic at first glance appears to show a mingling of two topics, one of research within Enron and the second involving universities (possibly recruiting). However, after examining relevant emails, it emerges that Vince Kaminski, head of the Research Group, had a close relationship with the faculty at Rice University (and is currently an adjunct professor there). He and several of his employees often spoke there and/or invited classes to Enron for research projects. As a result, the topic is clearly about Enron's Research Group. Finally, the Information Technology topic also emerges clearly with words like information system and server as well as the names of Enron's software packages (Unify, SAP, and Sitara).

Senior Mgmt		California Crisis		Research		Info Technology	
CATEGORY 45		CATEGORY 2		CATEGORY 47		CATEGORY 40	
Video	0.3%	Assembly	0.2%	Research	0.6%	Unify	1.1%
Boardroom	0.1%	AB	0.2%	Model	0.6%	Directory	1.1%
Sherri	0.1%	Crisis	0.2%	Resume	0.2%	Enterprise	1.0%
Task	0.1%	Deregulation	0.2%	Visit	0.2%	Hardware	0.7%
Safety	0.1%	Urgent	0.2%	University	0.2%	Script	0.5%
Peer	0.1%	Declare	0.2%	Finance	0.2%	Logistic	0.5%
Sera	0.1%	Freeze	0.2%	Rice	0.2%	Stage	0.4%
Palmer	0.1%	Legislature	0.1%	Dear	0.2%	Setup	0.3%
Medium	0.1%	Sold	0.1%	Student	0.1%	Solar	0.3%

Figure 4 PLSI-U Sample Categories with only Dictionary Words (from the 48 available).

Senior Mgmt		California Crisis		Research		Info Technology	
CATEGORY 11		CATEGORY 0		CATEGORY 44		CATEGORY 7	
PRC	0.3%	Governor	0.3%	Vinc	1.3%	Unify	0.7%
Video	0.2%	Calpin	0.3%	Kaminski	0.7%	SAP	0.4%
Weekly	0.2%	IEP	0.3%	Research	0.4%	Netco	0.3%
Ken	0.2%	Dasovich	0.3%	Model	0.3%	Sitara	0.3%
Dial	0.2%	Edison	0.3%	Shirley	0.2%	Script	0.3%
Kean	0.2%	Gov	0.2%	Rice	0.2%	Class	0.2%
Cindy	0.2%	IEPA	0.2%	Visit	0.2%	Setup	0.2%
VP	0.1%	Duke	0.2%	Crenshaw	0.2%	Path	0.2%
Passcode	0.1%	Mara	0.2%	University	0.2%	Regan	0.2%

Figure 5 PLSI-U Sample Categories with All Words (from the 48 available).

The next step is finding the topics that each individual is interested in. Recall that this is done by first calculating what the average interest is in a topic ($p(\text{topic}|\text{user})$) and then finding those individuals who have an interest in the topic greater than 95% of the population (i.e. $p(\text{topic} = T1|\text{user} = U1) \geq E(p(\text{topic} = T1|\text{user})) + 1.64(S^2(p(\text{topic} = T1|\text{user})))$). Those individuals with the highest interest in the selected topics are shown in Figures 6 and 7. From a cursory review of individuals' positions within Enron, the individuals with the highest interest in these topics appear appropriate. It is reasonable that the Senior Mgmt topic produces good results since it is created by looking at specific users. It is comforting to see Jeff McMahon who at different times held such positions as corporate treasurer, Chief Financial Officer and Chief Operating Officer. While it would have been desirable to have Jeff Skilling, Enron CEO, emerge in the top ten, the results are still promising. PLSI-U produces similar results for the California Crisis. Finding prominent public relations people (like Mark Palmer and Karen Denne) as well

as prominent government affairs people (like Jeff Dasovich and Richard Shapiro) is encouraging. The Research topic differs from the previous two by its limited nature. This topic is focused on a relatively small group within the Enron corporation. As a result, it produces excellent results. This is despite a mix of small and large email datasets for the top individuals. This suggests that when attempting to find individuals who all participate in a topic, if the topic is of limited interest, then the results are excellent. The only topic of concern is Information Technology. However, this may be due to the inability of the researcher in identifying most of the individuals and their positions. While the emails of many of these individuals seem to indicate a connection with I/T, their exact positions and responsibilities are unknown.

CATEGORY 45 SENIOR MGMT			CATEGORY 2 CALIFORNIA CRISIS		
Steven Kean	Chief of Staff, Government Relations Specialist	5.8%	Ken Lay	Chairman of Enron	7.6%
Stanley Horton	Chief Executive – Enron Transportation Group	4.0%	Karen Denne	Vice President of Public Relations	6.7%
Steven Kean	Chief of Staff – Government Relations Specialist	2.5%	Sandra McCubbin	Director of Government Affairs in California	4.9%
Maureen McVicker		2.4%	Paul Kaufman	Director of Government Affairs	3.9%
Rosalee Fleming	Secretary to Enron Chairman Kenneth Lay	2.1%	Jeff Dasovich	Government Affairs Executive	3.8%
Greg Whalley	President of Enron	1.8%	Harry Kingerski		3.6%
Mark Frevert	Vice-Chairman of Enron	1.6%	Steven Kean	Chief of Staff, Government Relations Specialist	3.3%
Kenneth Lay	Chairman of Enron	1.6%	Mark Palmer	Head of Corporate Communications	3.2%
Cindy Olson	Head of Human Resources	1.5%	Susan Mara	Director of Government Affairs in California	3.1%
Jeff McMahon	Chief Financial Officer of Enron	1.3%	James Steffes	Vice President of Government Affairs	2.8%

CATEGORY 47 RESEARCH			CATEGORY 40 INFO TECHNOLOGY		
Vince Kaminski	Managing Director and Head of Research	34.1%	Lisa Kinsey		1.0%
Jeffrey Shankman	Chief Operating Officer for Global Markets	6.2%	Robert Superty	Enron North America – Director Gas Procurement	1.0%
Shirley Crenshaw	Research Group Administrative Coordinator	5.0%	Patti Sullivan		1.0%
Stinson Gibner	Vice President in Quantitative Research Group	4.0%	Daren Farmer	Logistics Manager	0.8%
Vasant Shanbhogue	Vince Kaminski’s Second in Command	1.8%	Victor Lamadrid		0.8%
Tanya Tamarchenko	Director – Value at Risk	1.5%	Darla Saucier		0.8%
Zimin Lu	Director of Valuation and Trading Analytics Group	1.5%	Kirk Lenart		0.7%
Jennifer Burns		1.4%	Tammy Gilmore		0.7%
Grant Masson	Vice President – Research Group	1.2%	Cora Pendergrass		0.7%
Pinnamaneni Krishnarao	Vice President – Research Group	1.2%	Mark Schrab		0.6%

Figure 6 PLSI-U Sample Categories with only Dictionary Words and the Most Probable Individuals(from the 48 available).

CATEGORY 11 SENIOR MGMT			CATEGORY 0 CALIFORNIA CRISIS		
James Derrick	General Counsel	3.60%	Jeff Dasovich	Enron Government Affairs Executive	1.46%
Cindy Olson	Head of Human Resources	1.98%	James Wright		1.04%
Kay Chapman	Secretary of Management Committee	1.67%	Richard Sanders	VP and Asst General Counsel for Enron Wholesale	0.88%
Mark Koenig	Executive Vice President of Investor Relations	1.67%	Susan Mara	Director of Government Affairs in California	0.84%
Greg Whalley	President of Enron	1.66%	Scott Stoness		0.83%
Steven Kean	Chief of Staff, Government Relations Specialist	1.57%	Dennis benvides	Director of Green Power for Enron Energy in CA	0.80%
Mark Frevert	Vice-Chairman of Enron	1.54%	Sandra McCubbin	Director of Government Affairs in California	0.80%
Jeffrey McMahon	Chief Financial Officer of Enron	1.46%	Richard Shapiro	VP of Regulatory Affairs & principal DC lobbyist	0.80%
Kenneth Lay	Chairman of Enron	1.25%	James Steffes	Vice President of Government Affairs	0.76%
David Delainey	Enron Energy Services CEO	1.19%	Harry Kingerski		0.76%

CATEGORY 44 RESEARCH			CATEGORY 7 INFO TECHNOLOGY		
Vince Kaminski	Managing Director and Head of Research	20.88%	Daren Farmer	Logistics Manager	2.37%
Vince Kaminski	Managing Director and Head of Research	5.30%	Robert Superty	Director – Gas Procurement Enron North America	1.82%
Shirley Crenshaw	Research Group Administrative Coordinator	3.49%	Patti Sullivan		1.54%
Vince Kaminski	Manager Director and Head of Research	2.86%	Victor Lamadrid		1.46%
Stinson Gibner	Vice President in Quantitative Research Group	2.40%	Lisa Kinsey		1.38%
Don Baughman	North America Power trader – East Desk	2.00%	Bryce Baxter		1.20%
Vasant Shanbhogue	Kaminski’s second in command	1.52%	Tammy Jaquet		1.04%
Zimin Lu	Director of Valuation and Trading Analytics Group	1.05%	Clarissa Garcia		0.97%
Eric Bass	trader	1.03%	Regan Smith	Network Administrator	0.89%
Vince Kaminski	Managing Director and Head of Research	1.00%	Kevin Heal		0.87%

Figure 7 PLSI-U Sample Categories with All Words and the Most Probable Individuals(from the 48 available).

Once the categories are resolved into words and individuals’ interests in those topics are determined, the next step is constructing social networks for each topic. The first network constructed for a topic connects pairs of individuals who share a common interest in that topic. One example of these implicit interest social networks is seen in Figure 8. The second network constructed for a topic connects pairs of individuals who have passed at least one email associated with that topic.

One example of these explicit email social networks is seen in Figure 9). In this small example, there is no one with an interest in the topic who has not passed at least one email related to that topic with another Enron employee.

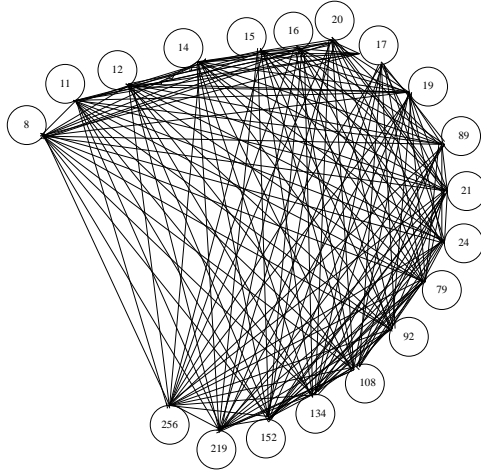


Figure 8 PLSI-U Enron Implicit Social Network for Database Topic.

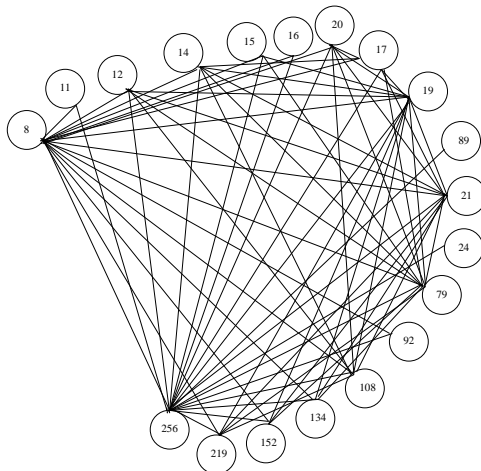


Figure 9 PLSI-U Enron Explicit Social Network for Database Topic.

The final step is to focus on the topics of interest. For instance, for Enron, the topic might concern the off-book partnerships that resulted in Enron's downfall. These off-book partnerships were named LJM1 and LJM2 after the Chief Financial Officer's family (his wife, Lea, and his sons, Jeffrey and Matthew). Unfortunately since LJM1 and LJM2 are acronyms, they are not found in the experiment where only words in the dictionary are used. Instead, the most troubling transactions performed by these partnerships, named the Raptors, were used (since raptor is a word found in the dictionary). When the 48 topics are examined to find which contain a non-zero conditional probability for the word raptor (i.e. $p(\text{word} = \text{raptor} | \text{topic})$), only 1 topic emerges for each experiment. For the experiment where words are restricted to the dictionary, five individuals emerge as having a hidden interest in this topic while when the restriction is removed, only one individual (a different one) emerges as having a hidden interest in the topic.

The second topic of interest is one on socializing. In this case, there is no single word that can be used. Instead the words *dinner*, *drink*, *fun*, *tonight*, *love*, *weekend*, *family* and *game* are used. Using these words, three topics are found for each experiment with non-zero probabilities for all of these words. In total, 293 individuals emerge as having a clandestine interest in at least one of these three topics for the experiment where only words in a dictionary are used. When the restriction is removed, the number of individuals drops to 89. The last step is to find those individuals with clandestine interests in both topics. Unfortunately, in neither experiment do any individuals emerge as having a clandestine interest in both topics.

These results are very promising, especially when one considers that this is out of a population of over 34,000 individuals. Even without using the cross-referencing technique, only 0.9% of the individuals in the population emerge as likely investigative leads for potential insider threats for the experiment restricted to words in the dictionary. When the restriction is removed, this number decreases even further to only 0.3% of the population.

In addition to finding clandestine interests, the social networks generated are also useful. If investigators needed to track down information on the database topic (Figure 9), a good place to start would be user 256 since he is connected to everyone. If, on the other hand, they needed to start looking at possible suspects, perhaps users 89 or 24 would be better since they have only a weak connection to other people interested in this topic. In this case, it might be suspicious that user 89, who has sent or received 1985 emails in total and has a 31% interest in this topic, has only emailed one other person about it.

5.1 Social Network Analysis Comparison

While probabilistic clustering is one method for finding the individuals most interested in a topic, a second method is social network analysis (SNA). Instead of using conditional probabilities, SNA uses several other measurements to determine an individual's importance or centrality Wasserman (1994). Degree is the simplest measurement of centrality and assumes that the most central actors are linked with the greatest number of other actors. By counting the number of links an individual has, his importance can be easily calculated. A second measurement of centrality, closeness, measures the distance of an actor from the other actors in the network (the weight of each edge is the same). This measurement assumes that those individuals "in the middle" are the most important. Unfortunately, one of the drawbacks to these measurements is that if the network is not connected, this is no agreed upon way to measure closeness. A third SNA measurement, betweenness, overcomes this by counting the number of shortest paths an actor resides on.

By applying these SNA measurements to the social networks generated by the Detection Algorithm, it is possible to validate whether the individuals with greatest probability for a topic are also the most central for that topic. Further, observe that while SNA can extract the most central individuals from these social networks, it is unable to generate the topics themselves. It is only useful once probabilistic clustering has provided the groundwork. However, once the groundwork has been laid, SNA does provide additional validation.

There appears to be little difference between centrality rankings. Degree, Closeness, and Betweenness in general show the same individuals as most central. In fact, only five individuals do not appear in at least two of the rankings. This phenomenon repeats for all four topics across both experiments. Therefore, for brevity, only the top ten most central individuals based on betweenness for the sample topics are shown for the two experiments (Figures 10 and 11). Each of these shows individuals very similar to those produced by probabilistic clustering supporting the effectiveness of probabilistic clustering in associating individuals with appropriate topics and providing an alternative means of discovering these individuals.

While these SNA techniques are revealing when used on the topic sub-graphs generated from the probabilistic clustering, when used on the social network as a whole (i.e. a network where two individuals are linked if they communicate via email), they reveal nothing. None of the individuals in the company are disconnected from this total email graph. All of the individuals exchanged at least one email with another Enron employee. Therefore, it is only when these SNA techniques are combined with the results of probabilistic clustering that they are revealing.

CATEGORY 45		SENIOR MGMT	CATEGORY 2		CALIFORNIA CRISIS
Tracey Kozadinos		0.30	Alan Connes	Director of Government Affairs in California	0.28
Jeff Skilling	Chief Executive of Enron	0.22	Kenneth Lay	Chairman of Enron	0.25
Constance Charles	Human Resources – Associate/ Analyst Program	0.17	Simone La		0.13
Steven Kean	Chief of Staff – Government Relations Specialist	0.15	Clayton Seigle		0.12
Rosalee Fleming	Secretary for Chairman Ken Lay	0.15	Jeff Dasovich	Government Affairs Executive	0.08
Rhonda Denton		0.04	Steven Kean	Chief of Staff – Government Relations Specialist	0.08
Bill Donovan		0.04	Karen Demne	Vice President of Public Relations	0.07
Brian Ripley		0.04	Ginger Demeuhl	Admin Assistant – Global Government Affairs	0.07
Janet Butler		0.04	Richard Shapiro	VP of Regulatory Affairs, Chief DC Lobbyist	0.07
Rhonda Denton		0.04	Leonardo Pacheco		0.06

CATEGORY 47		RESEARCH	CATEGORY 40		INFO TECHNOLOGY
Vince Kaminski	Managing Director and Head of Research	0.34	Cheryl Johnson		0.47
Outlook Team		0.15	Outlook Team		0.26
Jewel Meeks		0.11	Emma Welsch		0.15
Kristin Gandy	Associate Recruiter for Enron	0.09	Jim Schwieger	Vice President in Gas Trading Division	0.12
Shirley Crenshaw	Research Group Administrative Coordinator	0.09	Julie Meyers		0.10
Jeff Dasovich	Government Affairs Executive	0.08	Darren Vanek	Credit Analyst – Credit Risk Management	0.09
Nicki Daw		0.08	Carolyn Gilley	Enron Networks – Information & Records Mgmt	0.08
Richard Shapiro	VP of Regulatory Affairs, Chief DC Lobbyist	0.07	Geoff Storey		0.08
Ashley Baxter	Recruiter – Global Technology Track	0.07	Kevin Dumas		0.06
Althea Gordon	Recruiter – Associates/ Analyst Program	0.07	Daren Farmer	Logistics Manager	0.05

Figure 10 PLSI-U Sample Categories with only Dictionary Words and Individuals with Highest Betweenness Measurements (from the 48 available).

CATEGORY 11		SENIOR MGMT	CATEGORY 0		CALIFORNIA CRISIS
Joannie Williamson	Secretary to CEO Jeff Skilling	0.21	Susan Mara	Director of Government Affairs in California	0.28
Bobbie Power		0.09	Jeff Dasovich	Government Affairs Executive	0.24
Tracy Ralston		0.07	Alan Connes	Director of Government Affairs	0.12
Billy Lemmons		0.07	Joseph Alamo		0.09
David Delainey	CEO of Enron Energy Services	0.06	Sandra McCubbin	Director of Government Affairs in California	0.09
Jeff Skilling	CEO of Enron	0.06	Dan Leff		0.08
Cindy Olson	Head of Human Resources	0.06	Tamara Johnson		0.06
Rosalee Fleming	Secretary to Chairman Ken Lay	0.06	Michael Tribollet	VP of Underwriting and Investment Valuation	0.06
Paula Rieker	Deputy Director of Investor Relations	0.05	Leticia Botello		0.04
Liz Taylor		0.04	Thomas Bennett		0.02

CATEGORY 44		RESEARCH	CATEGORY 7		INFO TECHNOLOGY
Vince Kaminski	Managing Director and Head of Research	0.44	Cynthia Morrow		0.55
Vince Kaminski	Managing Director and Head of Research	0.18	Regan Smith	Network Administrator	0.17
Shirley Crenshaw	Research Group Administrative Coordinator	0.16	Georgia Ward	QA in Development Support	0.13
Ravi Thurasingham	Director of Global Bandwidth Risk Management	0.07	Brandee Jackson		0.09
Anjam Ahmad		0.05	Bryce Baxter		0.08
Anita Dupont		0.05	Kenneth Harmon		0.07
Vince Kaminski	Managing Director and Head of Research	0.05	Rita Wynne	Manager for Volume Management Group	0.06
Vasant Shanhogue	Vince Kaminski's Second in Command	0.04	Brian Ripley		0.05
Zimin Lu	Director of Valuation and Training Analytic Group	0.04	Tony Dugger		0.05
Steven Leppard		0.04	Anwar Melethil		0.04

Figure 11 PLSI-U Sample Categories with All Words and Individuals with Highest Betweenness Measurements (from the 48 available).

6 Conclusions and Future Work

The Potential Insider Threat Detection Algorithm emerges from this research as a promising tool. The topics generated by PLSI-U are easily identifiable both based on the most probable words as well as the most probable individuals. In addition it generates a small, manageable number of individuals as investigative leads. However, much work remains. While a small number of investigative leads emerge, none of the principle perpetrators in Enron's fall (such as Ken Lay, Jeff Skilling, and Andy Fastow) are among them. This may be because any revealing emails would have been to other people *inside* the organization, thus thwarting the algorithm described in this paper. To overcome this, work needs to be done to extract insider threat collusion networks including a small number of individuals as well as extracting individual insider threats.

Secondly, while many of the categories were easy to identify by the most probable words, some were not. A different model for extracting topics might produce better results. Latent Dirichlet Allocation (LDA) has been shown to be a more general case of PLSI Girolami (2003). By not assuming that the mixture of topics in the corpus is the only possible mixture of topics, LDA has a better chance of describing previously unseen emails. Rosen-Zvi, et al developed the Author-Topic model Rosen-Zvi (2004) that expands on LDA by including clustering on individuals.

A final area for improvement is expanding PLSI-U from email to internet activity. This work has already been done with PLSI by Cohn Cohn (2000). By introducing internet activity, the implicit interest profiles would not be generated from the same data used to generate the explicit email networks. As a result, better topics should emerge as well as more clandestine interests. While internet activity was not available for Enron, it is generally available from the same sources that supply email history logs.

Acknowledgements

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government

References and Notes

- 1 D. Cohn, and H. Chang (2000) 'Learning to Probabilistically Identify Authoritative Documents', *Proc. 17th International Conf. on Machine Learning*, 167–174. Morgan Kaufmann, San Francisco, CA.
- 2 *Merriam-Webster Collegiate Dictionary, Espionage*, URL: <http://www.m-w.com/cgi-bin/dictionary>.
- 3 M. Girolami and A. Kaban, 'On an equivalence between PLSI and LDA', URL: citeseer.ist.psu.edu/girolami03equivalence.html.
- 4 K.L. Herbig and M. F. Wiskoff (2002) 'Espionage Against the United States by American Citizens 1947 - 2001', *Technical Report, Defense Personnel Security Research Center (PERSEREC)*.

- 5 T. Hoffman (1999) ‘Probabilistic Latent Semantic Indexing’, *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*.
- 6 S. Martin, A. Sewani, B. Nelson, K. Chen, and A. Joseph (2005), ‘Analying Behavioral Features for Email Classification’, *Second Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA.
- 7 A. McCallum, A. Corrada-Emmanuel, and X. Wang (2004), ‘Topic and Role Discovery in Social Networks’, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, San Jose, CA.
- 8 J. Okolica, G. Peterson, and R. Mills (2006), ‘Using PLSI-U to Detect Insider Threats from Email Traffic’, *Springer-Verlag 2006*.
- 9 RAND (1999), ‘Research and Development Initiatives Focused on Preventing, Detecting, and Responding to Insider Misuse of Critical Defense Information Systems’, (<http://www.rand.org/publications/CF/CF151/CF151.pdf>).
- 10 M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth (2004), ‘The Author-Topic Model for Authors and Documents’, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 487-494.
- 11 Salvatore Stolfo and Shlomo Hershkop and Ke Wang and Olivier Nimeskern and Chia-Wei Hu (2003), ‘A Behavior-based approach to Securing Email Systems’, *International Workshop on Mathematical Methods, Models and Architectures for Computer Networks Security (ACNS-2003)*.
- 12 S. Symonenko, E.D. Libby, O. Yilmazel, R. Del Zoppo, E. Brown, and M. Downey (2004), ‘Semantic Analysis for Monitoring Insider Threats’, *Second Symposium on Intelligence and Security Informatics (ISI 2004)*.
- 13 Wasserman, Stanley and Katherine Faust (1994) *Social Network Analysis: Methods and Applications*, Cambridge University Press, New York, NY.