Chapter 1

USING PLSI-U TO DETECT INSIDER THREATS FROM EMAIL TRAFFIC *

James S. Okolica, Gilbert L. Peterson, Robert F. Mills

Abstract Despite a technology bias that focuses on external electronic threats, insiders pose the greatest threat to commercial and government organizations. Once information on a specific topic has gone missing, being able to quickly determine who has shown an interest in that topic can allow investigators to focus their attentions. Even more promising is when people can be found who have an interest in that topic but have never communicated that interest within the organization. By datamining emails, an employee's interests can be discerned. These interests can then be used to construct social networks which can graphically expose investigative leads. This paper describes the use of Probabilistic Latent Semantic Indexing (PLSI)[5] extended to include users (PLSI-U) to determine topics of interest for employees from their email activity. It then applies PLSI-U to the Enron email corpus and finds a small number of employees (0.02%) who appear to have had clandestine interests.

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

1. Introduction

"Espionage is the practice of spying or using spies to obtain information about the plans and activities of a foreign government or a competing company"[2]. While professional spies can be inserted into an organization, today, the use of insiders is much more prevalent. Insiders are members of an organization who often have a legitimate right to the

*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

information that they are accessing. However, they abrogate the trust they have been given by using the information for illegitimate reasons.

The least costly time to address the Insider Threat is before it occurs, i.e. prevention. The next best time is as soon as it is detected and before the perpetrator manages to cover his tracks and escape. Given the scarcity of time and the large number of employees, investigators must quickly winnow the number of suspects to a number small enough to be manageable. As more and more information is stored electronically, it is now possible to datamine electronic information and extract likely investigative leads. One of the best indicators of a person's interests in organizations is email. Through datamining, topics of interest can be extracted from email and people can be categorized by the topics they are most interested in. To reactively identify investigative leads, only people who have shown an interest in specific topics of interest need to be studied further. Especially likely suspects are people who have shown an interest in the topic but have never communicated that interest with anyone within the organization.

In this paper, Probabilistic Latent Semantic Indexing (PLSI)[5] 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 103 employees with clandestine interests. Given that these 103 emerge from a collection of over 34,000 employees, the algorithm appears to produce a manageable collection of insider threats investigative leads.

2. Motivation: Datamining Email to detect insider threats

During a RAND workshop on the Insider Threat[4], the first priority for improving the detection of insider's misuse was "developing [user] profiling as a technique" [10]. One way to detect potential insiders is to consider whether a person's interests match with the people they are in contact with. A profile describing a person's interests is generated by analyzing the content of their email. If someone shows a high degree of interest in a specific topic but does not email anyone else within the organization that also has an interest in that topic, it may suggest a clandestine interest. If in addition, the category is one relevant to an insider threat investigation, the person may warrant additional attention.

Electronic mail is fast becoming the most common form of communication. In 2006, email traffic is expected to exceed over 60 billion message daily [7]. While using email as data is not new, it has only re-

cently begun to emerge as a tool for detecting deceptive communications [6]. Semantic analysis has been directly applied to countering insider threats by Symonenko, et al. [12]. 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 their 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.

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. The profiles are then used in generating an implicit social network between people for each interest. Individuals are considered connected if they share an interest. A second explicit social network for each interest is then constructed strictly based on the presence of email activity (containing that interest) between individuals. These two networks are then compared for discrepancies. People who fail to communicate via email for a specific interest (i.e. not connected to anyone according to the explicit social network) but who have shown an interest (i.e. connected according to an implicit social network) are then considered as possibly having a clandestine connection and worthy of additional investigation. Consider the example in Figure 1. 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* 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.

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

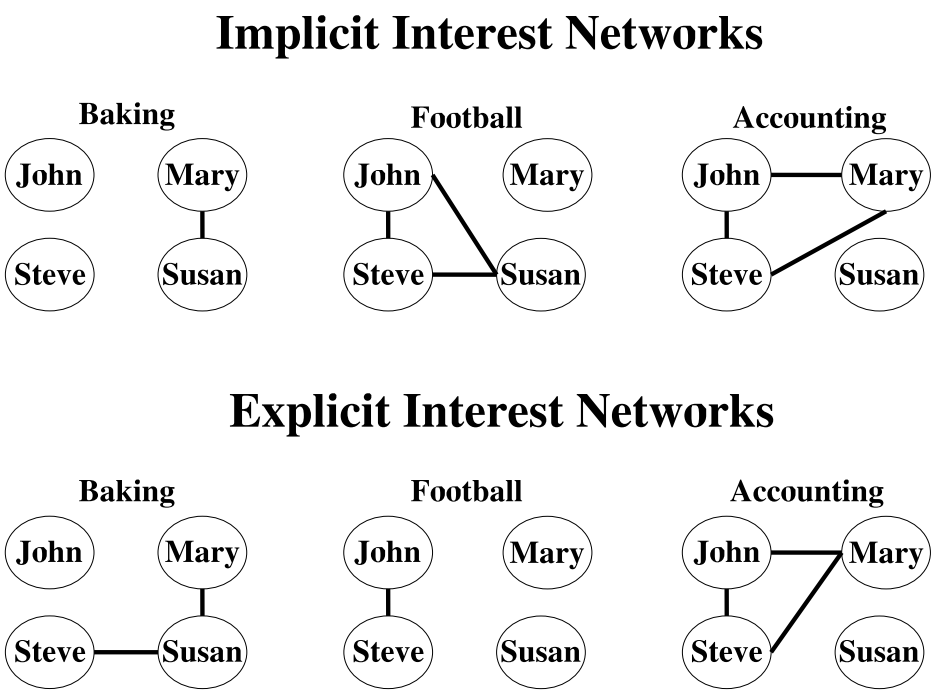

Figure 1. An Example of Clandestine Interests (implicit network = external; internal email explicit network = internal email only).

in a topic that exceeds a threshold, specifically 0.5%, a link is created between those two people. Mathematically, if $p(z = Z_1|u = U_1) > 0.005$ and $p(z = Z_1|u = U_2) > 0.005$ where Z_1 is a topic and U_1 and U_2 are people, 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 $p(z = Z_1|d = D_1) > 0.005$ where D_1 is an email, 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. Each implicit network is compared to the explicit network is turn. 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, an exception is generated.

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 [8]) 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 cate 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 requires incorporation, 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 $p(z|d, w, u)$.

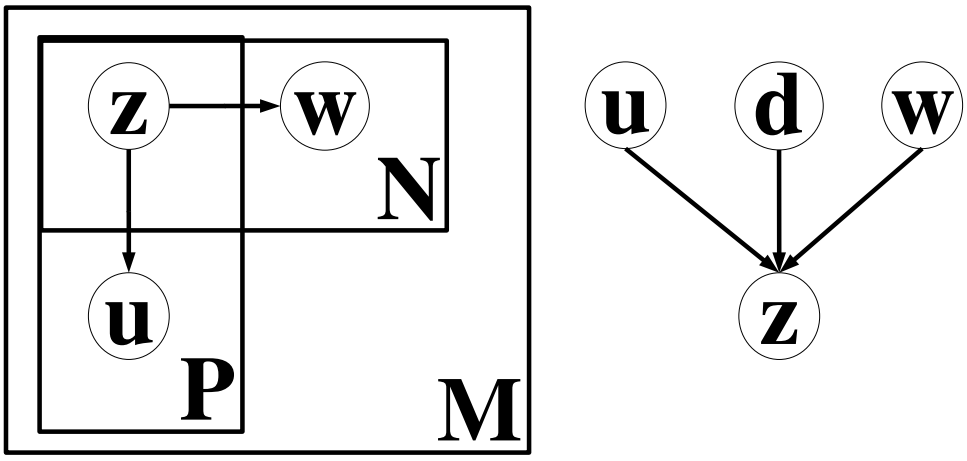

Figure 2. PLSI-User Mixture Model

To begin, using Bayes rule:

$$p(z|u, d, w)p(u, d, w) = p(u, d, w|z)p(z). \tag{1}$$

Now, consider the model in Figure 2 and observe that u, d, and w are all conditionally independent given z. Therefore,

$$p(z|u, d, w) = \frac{p(u|z)p(d|z)p(w|z)p(z)}{p(u, d, w)} \tag{2}$$

But $p(u, d, w)$ is simply $p(u, d, w|z)$ marginalized across all possible z's. So finally,

$$p(z|u, d, w) = \frac{p(u|z)p(d|z)p(w|z)p(z)}{\sum_{z' \in Z} p(u|z')p(d|z')p(w|z')p(z')} \tag{3}$$

In order to evaluate the conditional probabilities in the above equations, consider:

$$p(w|z) = \frac{p(z|w)p(w)}{p(z)} \tag{4}$$

By marginalizing across u and d, we get:

$$p(w|z) = \frac{\sum_{u \in U} \sum_{d \in D} p(z|u, d, w)p(w|u, d)}{\sum_{u \in U} \sum_{d \in D} \sum_{w' \in W} p(z|u, d, w')} \tag{5}$$

Finally, consider what $p(w|u,d)$ means. This is the probability of a given word occurring for a given email and person. Since the email and person are already specified the probability space is the one email. Therefore the probability is the number of times the word appears in the email divided by the number of words in the email. Therefore:

$$p(w|z) = \frac{\sum_{u\in U}\sum_{d\in D} p(z|u,d,w)n(d,w)}{\sum_{u\in U}\sum_{d\in D}\sum_{w'\in W} p(z|u,d,w')n(d,w)} \tag{6}$$

where $n(d,w)$ is the number of times a word occurs in an email. Observe that since an email is the same regardless of which "author" is considered, it is sufficient to specify $n(d,w)$ so long as it is summed across all people. Furthermore, since the denominator sums across all words, the net effect is the quotient described previously. This equation extends naturally to emails and users:

$$p(d|z) = \frac{\sum_{u\in U}\sum_{w\in D} p(z|u,d,w)n(d,w)}{\sum_{u\in U}\sum_{d'\in D}\sum_{w\in W} p(z|u,d',w)n(d,w)} \tag{7}$$

$$p(u|z) = \frac{\sum_{d\in D}\sum_{w\in W} p(z|u,d,w)n(d,w)}{\sum_{u'\in U}\sum_{d\in D}\sum_{w\in W} p(z|u',d,w)n(d,w)} \tag{8}$$

$$p(z) = \sum_{u\in U}\sum_{d\in D}\sum_{w\in W} p(z|u,d,w) \tag{9}$$

These equations can now form the expectation (eq. 3) and maximization (eq. 6, eq. 7, eq. 8, eq. 9) 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. During the investigation into the Enron scandal, the Federal Energy Regulatory Commission (FERC) made email from Enron publicly available. As a part of this process, it placed the email of 150, primarily senior, employees

of Enron accessible electronically on the World Wide Web. 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. Observe that while it would be applicable to use Enron as a case study, this paper is 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. The entire Enron corpus was used which consisted of 245,904 emails made up of 54,147 stemmed words and 87,395 users of which 34,885 were Enron employees. In addition, it was decided a priori that the corpus was made up of 48 categories based on some previous research by McCallum, et al.[8]. While the theoretical joint distribution of $p(z|d, w, u)$ could consist of approximately 5.5×10^{16} probabilities, the actual distribution for the corpus consisted of 3.4×10^9 probabilities. This size still forced the the implementation of a parallel algorithm to reduce the processing time per iteration to two hours. 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.

Some of the words from the resulting categories are shown in Figure 3. 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 distinguished categories proved easy to identify. In order to produce a list that excluded common, non-distinguishing words, only words that appeared in at most 5 categories were used to define a category. While in general, this made the categories much easier to identify, the removal of some words, for instance 'program' from category 16, made some more difficult to understand. By manually examining the documents that had the highest conditional probabilities given the topic (i.e. $p(d|z)$), it became evident that in general the words of highest probability did describe the categories well. For instance, in category 1 many of the most likely documents concerned accessing scheduling databases. In category 14, many promising documents described the Associate/ Analyst Program, a mentoring program for new hires. Category 16 seemed to be about improving Enron's online presence through web traffic and email. And category 40 describes a migration to new computer systems, backing up of directories, and a weekend outage.

Category 1		Category 14		Category 16		Category 27		Category 40		Category 45	
Database	6.0%	Associate	1.6%	Outlook	5.2%	Image	0.4%	Migrate	2.3%	Ken	0.3%
Alias	4.0%	Analyst	1.6%	Migrate	4.5%	Expect	0.3%	Application	1.8%	Video	0.3%
Error	3.9%	Pilot	1.4%	Calendar	1.4%	Monitor	0.2%	Unify	1.1%	Lay	0.2%
Unknown	3.6%	Accept	0.4%	Mailbox	1.3%	Fool	0.2%	Directory	1.1%	Effort	0.2%
Variance	3.0%	Important	0.4%	Button	1.1%	Senate	0.2%	Enterprise	1.0%	Return	0.2%
Detect	2.4%	Opportunity	0.4%	Client	0.8%	Free	0.2%	Outage	0.9%	Corporation	0.2%
Log	2.2%	Feedback	0.3%	Journal	0.8%	Source	0.2%	Begin	0.8%	Board	0.2%
Parse	2.1%	Condition	0.3%	Web	0.7%	Security	0.2%	Log	0.7%	Rick	0.2%

Figure 3. PLSI-U Sample Categories (from the 48 available).

Once the categories are resolved into words, the next step is constructing the social networks. For the implicit social network constructed from interests, this consists of connecting every pair of people that had the same interest (Figure 4). For the explicit social network, consisting of the relevant emails, every pair of people are examined to see if they passed at least one email for the relevant topic between them; if so, a link is made between them (Figure 5). In the case of Category 1, there is no one with an interest in category 1 that does not have at least one email with at least one Enron employee who also has an interest in category 1.

The final step is to consider which are the principle categories each user fell into. For instance user 14920 was principally interested in category 9 (62%) and secondarily interested in category 38 (38%). Each Enron employee who sent emails was reviewed to see if there was at least one email sent or received for every category for which they had at least a 10% interest in. The results were promising; across all 48 categories, there were only 103 employees that showed up as having clandestine interests. This is even more remarkable when one considers that employees could show up as having clandestine interests in multiple categories (i.e. $34,885$ employees $\times 48$ categories $= 1,674,480$ possible clandestine interests). Even restricting the number to $34,885$, this means that less than 0.4% of the people appeared to have clandestine interests. Of these people, only 22 had a total of at least 10 emails sent or received (the overall corpus average was 46). Therefore, it is very likely that the remaining 81 people are false positives having rarely or never used email. In the remaining 22 cases, the data does bear out that although PLSI-U shows a person with a strong interest in a specific topic, no email sent or received within Enron is composed of that topic. For instance, one employee (person 14920) received 20 emails from a single non-Enron employee all pertaining to category 9. Of the other 4 emails received and 15 emails sent within Enron, none of them had anything to do with category 9 (Figure 6).

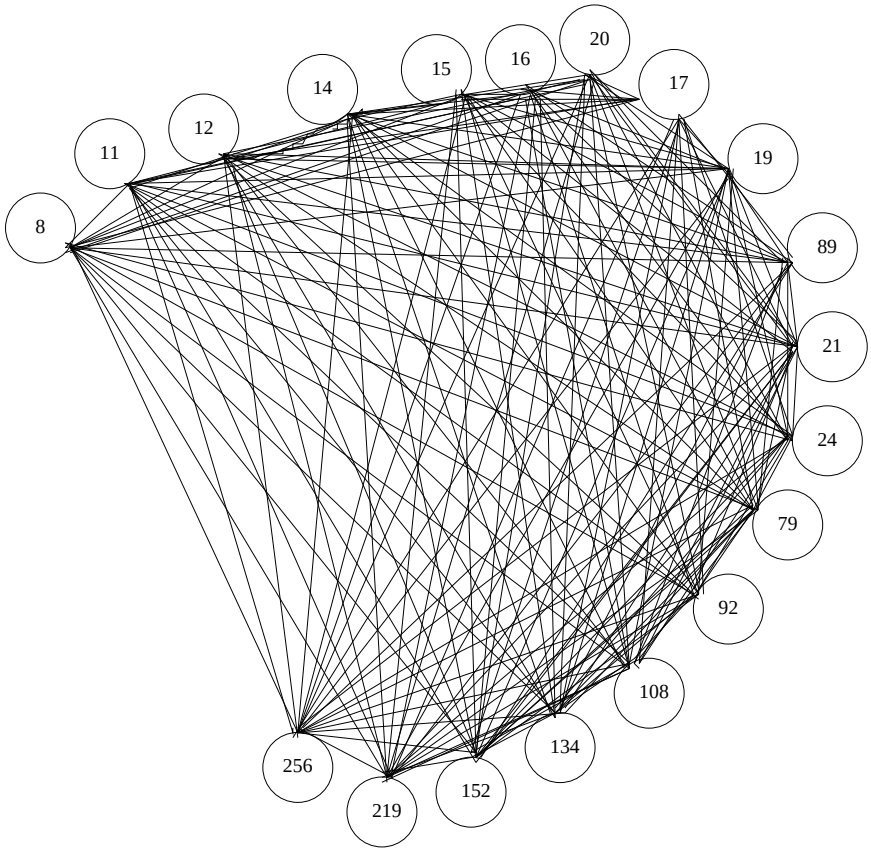

Figure 4. PLSI-U Enron Implicit Social Network for Category 1.

In addition to finding clandestine interests, the social networks generated are also useful. If investigators needed to track down information on category 1 (Figure 5), 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.

6. Conclusions and Future Work

The results of the experiment show that the theory is sound. PLSI-U works well extracting topics from the email corpus and the simple mechanism for finding clandestine interests produces very few false positives.

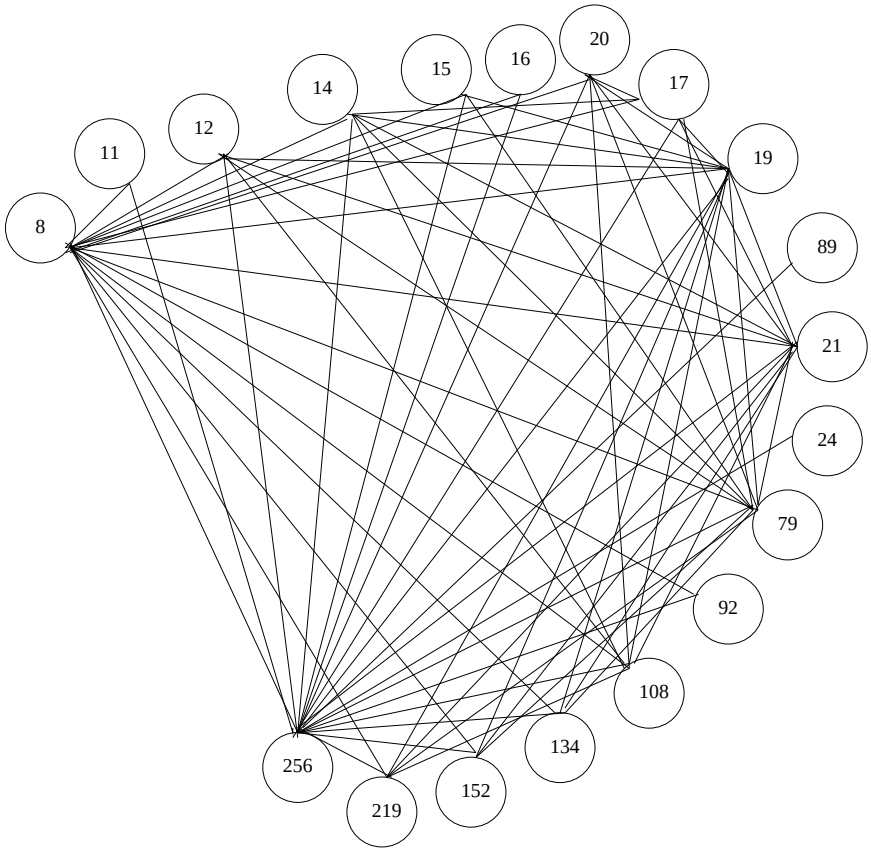

Figure 5. PLSI-U Enron Explicit Social Network for Category 1.

User 14920	User 17594	User 46814	User 22446
15 Messages Sent	15 Messages Sent	2 Messages Sent	22 Messages Sent
24 Messages Recv'd	84 Messages Recv'd	2102 Messages Recv'd	57 Messages Recv'd
TOPIC 9 62%	TOPIC 10 12%	Topic 2 42%	TOPIC 23 18%
Topic 38 38%	Topic 34 88%	TOPIC 12 10%	Topic 41 41%
		Topic 30 48%	Topic 43 2%
INTERNAL	INTERNAL	INTERNAL	INTERNAL
15 Messages Sent	15 Messages Sent	2 Messages Sent	22 Messages Sent
4 Messages Recv'd	77 Messages Recv'd	335 Messages Recv'd	39 Messages Recv'd
Principal Email Topics:	Principal Email Topics:	Principal Email Topics:	Principal Email Topics:
38 and 20	34, 9 and 11	2, 30 and 13	41, 43 and 15

Figure 6. PLSI-U Sample Users with Clandestine Interests (from the 22 extracted.)

While it would have been desirable to find Kenneth Lay, Jeffrey Skilling, and Andrew Fastow emerging as having clandestine interests, this did

not occur. This may be because any questionable emails would have been to other people *inside* the organization, thus thwarting the algorithm described in this paper. While they did not emerge as having clandestine interests, it is informative that all three had only 1 or 2 topics of interest. All three of them had a significant interest ($p(z|u) > 0.10$) in category 45 (Figure 3) while Jeffrey Skilling and Kenneth Lay also had a significant interest in category 27 (Figure 3). Therefore, while they do not appear to have clandestine interests, by examining the other people who also had a significant interest, additional people involved in the questional business practices may emerge. Even Sherron Watkins, the one Enron person considered to be a whistle-blower, did not emerge as having a clandestine interest. By her own admission, she saw nothing gained by going outside the company[9]. Her only attempt at whistle-blowing was to try to talk to Kenneth Lay herself.

Although this technique appears promising, much work remains. 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 show to be a more general case of PLSI[3]. 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[11] that expands on LDA by including clustering on individuals. This model may produce better topics results. Another change that might produce better results is not restricting vocabulary to words found in the dictionary. Acronyms and words like "socal" (for southern California) figure prominently into many Enron emails but are excluded because they do not appear in a dictionary. By allowing these words to appear in the topics, the topics may become more identifiable.

The final cause for concern is the overloading of email. In this application email is used to both define the topics and to determine who is not revealing their interest in a topic. On the surface this should result in no clandestine interests since the only way someone is considered to have an interest in a topic is if they send or receive an email about it. The only reason this is not the case is because during the definition of topics internal and external emails are considered while during the search for clandestine interests only internal emails are used. By using a different data source for generating topics of interest, more clandestine interests may emerge. One logical data source is internet activity. Internet history is kept on servers in the same way that email history is. In addition, PLSI can easily morph from documents made up of words to web pages made up of hyperlinks[1]. While internet activity was not

available for Enron, it is generally available from the same sources that supply email history logs.

References

[1] D. Cohn, and H. Chang, Learning to Probabilistically Identify Authoritative Documents, *"Proc. 17th International Conf. on Machine Learning"*, 167–174. Morgan Kaufmann, San Francisco, CA, 2000.

[2] Merriam-Webster Collegiate Dictionary, Espionage, (http://www.m-w.com/cgi-bin/dictionary).

[3] M. Girolami and A. Kaban, On an equivalence between PLSI and LDA, (citeseer.ist.psu.edu/girolami03equivalence.html).

[4] K.L. Herbig and M. F. Wiskoff, Espionage Against the United States by American Citizens 1947 - 2001, *Technical Report, Defense Personnel Security Research Center (PERSEREC)*, 2002.

[5] T. Hoffman, Probabilistic Latent Semantic Indexing, *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, 1999.

[6] P. Keila and D. Skillicorn. Detecting Unusual and Deceptive Communication in Email, *Technical Report, Queen's University*, Kingston, Ontario, Canada, June 2005.

[7] S. Martin, A. Sewani, B. Nelson, K. Chen, and A. Joseph, Analying Behaviorial Features for Email Classification, *Second Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA, 2005.

[8] A. McCallum, A. Corrada-Emmanuel, and X. Wang, Topic and Role Discovery in Social Networks, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, San Jose, CA, 2004.

[9] B. McLean and P. Elkind, *The Smartest Guys in the Room*. Penguin Group (USA), New York, NY, 2003.

[10] RAND, 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).

[11] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, The Author-Topic Model for Authors and Documents, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 487-494, 2004.

[12] S. Symonenko, E.D. Libby, O. Yilmazel, R. Del Zoppo, E. Brown, and M. Downey, Semantic Analysis for Monitoring Insider Threats,

Second Symposium on Intelligence and Security Informatics (ISI 2004), 2004.