9-2007

Using Author Topic to Detect Insider Threats from Email Traffic

James S. Okolica
Air Force Institute of Technology

Gilbert L. Peterson
Air Force Institute of Technology

Robert F. Mills
Air Force Institute of Technology

Recommended Citation

Okolica, J. S., Peterson, G. L., & Mills, R. F. (2007). Using Author Topic to detect insider threats from email traffic. Digital Investigation, 4(3–4), 158–164. https://doi.org/10.1016/j.diin.2007.10.002

Using Author Topic to Detect Insider Threats from Email Traffic ⋆

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

Air Force Institute of Technology
AFIT/ENG
Bldg 641 RM 220
2950 Hobson Way
Wright Patterson AFB, OH 45433-7765

Abstract

One means of preventing insider theft is by stopping potential insiders from becoming actual thieves. This article discusses an approach to assist managers in identifying potential insider threats. By using the Author Topic (16) clustering algorithm, we discern employees interests from their daily emails. These interests then provide a means to create an implicit and an explicit social network graph. This approach locates potential insiders by finding individuals who either (1) feel alienated from the organization (a key warning sign of a possible disgruntled worker) or (2) have a hidden interest in a sensitive(e.g. proprietary or classified) topic. In both cases, this is revealed when someone demonstrates an interest in a topic but does not share that interest with anyone in the organization. By applying this technique to the Enron email corpus, we produce coherent, sensible topics and reveal Sherron Watkins, the famous Enron whistle-blower, as a potential insider threat from the viewpoint of the individuals behind the Enron scandal.

Key words: Author Topic (AT), Insider Threat, Datamining, Social Networks, Large Dataset

⋆ 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

Email address: james.okolica, gilbert.peterson, robert.mills @afit.edu (James S. Okolica and Gilbert L. Peterson and Robert F. Mills).

1 Introduction

The best time to address the insider threat is before it occurs. Mostly, individuals do not enter an organization with the intent to do harm (17). Instead, something changes while they work at the organization. An automated way is needed to detect when individuals begin feeling alienated from an organization. Managers can then use the results to allocate scarce resources to determine if active intervention is required.

One of the best indicators of a person's interests in today's organizations is their email traffic. Through datamining, topics of interest are extracted and people categorized by those topics they are most interested in. The topics of interest are then used to detect warning signs such as alienation from the organization as well as inappropriate interest in sensitive or classified topics. Especially likely are people who have shown an interest in a sensitive topic but 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.

In this paper, Author Topic (16) is used on the Enron email corpus to test its applicability on generating insider threat investigative leads. The results produce 48 clear topics and reveal Sherron Watkins, Enron's well-known whistleblower, as one of 3 individuals who has both a clandestine interest in the off-book partnerships' topic and a feeling of alienation from Enron.

This paper is organized as follows. In Section 2, the nature of the insider threat is discussed as well as a short overview of the Author Topic Model. This is followed by the methodology and results of the experiment in Sections 3 and 4. The article concludes in Section 5 with conclusions and suggestions for future investigation.

2 Background

Espionage is the practice of spying or using spies to gather information about a competitor. An insider is someone in an organization who has a legitimate right to (some of) the organization's information but uses it for an illegitimate purpose. In 2002, Herbig and Wiskoff published a report detailing all of the publicly known cases of espionage against the United States between 1947 and 2001 (9). They found that, although in most cases the individual responsible was an insider, they did not enter the organization intending to commit espionage. In many cases, an organizational change, such as restructuring, or a personal crisis, such as the end of a relationship or a severe financial problem,

contributed to the individual choosing to commit espionage. Managers can no longer spend sufficient time to detect these warning signs. Instead, they must rely on some tool to direct their their attention on a few individuals.

In today's Information Age, one of the best sources of personal information available at work is an individual's email activity. Probabilistic clustering takes a collection of email and lumps it into coherent "topics". In addition it discovers the principle topics an individual is interested in. Prior to betraying an organization, individuals generally feel alienated from it. Whether such an alienation is initiated by the individual or the organization is irrelevant. What is required is for this alienation to exist (17). To test for this, social networks can be used. If individuals have an interest in a topic but do not send or receive email about that topic with others within the organization, they can be considered to have a clandestine interest in that topic. Depending on the number and type of such topics, these clandestine interests can be considered an indicator that an individual feels alienated from the organization and has insider threat potential.

2.1 Enron as the data source

As part of their investigation into Enron, the Federal Energy Regulatory Commission (FERC) seized Enron's email and made a portion of it publically available (7). 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 to make probabilistic clustering effective. Prior to the public disclosure of Enron's questionable accounting, Ms. Watkins, a vice president in Enron's corporate accounting division, sent a letter to Ken Lay, Enron's chairman, detailing the dubious accounting practices and their likely impact on Enron's future. Her activities were considered insider theft by her boss, Andy Fastow, Enron's Chief Financial Officer, who demanded that she be fired immediately (12). This is an example of the importance of perspective on distinguishing whistleblowers from insider threats.

2.2 Current Social Network and Insider Threat Research

A social network is generally constructed based on perceived common interests. For example, ReferralWeb (10) uses the co-occurrence of names in close proximity in documents publicly available on the WWW (e.g. journal articles, newsgroups, chatrooms etc.) to denote a close relationship. Lada Adamic has done similar research (1) by using mailing lists and the homepages of students at Stanford and MIT. Since, when people create homepages, they link

to their friends' homepages (and ask their friends to link to theirs), she postulated that using homepages would result in an appropriate social network. She also used the text present on the web pages to further predict relationships (i.e. common interests) between people. While she was able to show that the text provided strong indications of friendships between people, it is unclear if this would generalize beyond the rather closed community of a university. Culotta, et al. (4) approached this problem differently but from a more general population. They began by extracting names from email messages. Then they used the WWW to find the person's "web presence" (generally his or her homepage) and used that to describe the person and to find friends of that person. After the network was created, they used graph partitioning algorithms to find highly connected components. While their dataset was small (53 email correspondents), their results were promising. However, one of the biggest drawbacks was a lack of web presence for many of the correspondents (31 of 53).

Since September 11, 2001 there has been increased research in uncovering potential threats through the use of social networks. However, despite several organization, such as Rand Corporation and Mitre, making proposals for using social networks to detect insider threats (15; 13), little public research has been done. Symonenko (18) has generated social networks of intelligence analysts and then used semantic analysis to detect when individuals are showing interests in areas outside of their group. While the results have been promising, the technique requires a large number of interviews with experts to provide the semantic analysis. In addition, this expert knowledge is then only applicable to the specific group and needs to be repeated each time the application is moved to another organization. Yee (19) has also performed some initial research into generating social networks from email headers for later analysis by social network analysts.

Other insider threat research has examined the problem from a formal security model and management perspective. By mapping employee performance metrics to insider threat traits a manager can determine an individuals potential for insider threat behavior (3). Additionally, Bayes network models of behavior have also been shown to be potentially helpful in detecting insider threats (2)

2.3 Author Topic

This work examines the application of Author Topic (16), a probabilistic clustering technique, on the task of revealing insiders from the Enron email corpus. Author Topic is a probabilistic clustering technique that takes the words and individuals that make up an email and organizes them into the most likely

probability distributions which represent the underlying topics. The remainder of this section describes the theoretical underpinnings of the Author Topic (AT) model and its application to an email corpus. AT is posited on a collection of M emails. Each email is made up of a (possibly different) number of words from a vocabulary of V words. Each email is further associated with a (possibly different) number of individuals (one sender and one or more recipients) drawn from a population of P people. Finally, the emails are concerned with one or more topics drawn from a collection of K topics.

For details about Author Topic, see (16) and Figure 1. In short, in order to create an email, N words are selected. To pick a word, an author, u, is chosen uniformly from the population P. Then a topic, z, is selected conditioned on the author chosen. Since the probability of picking a topic is also conditioned on θ (a prior distribution of multinomial probability distributions), the probability of picking a topic is $p(z|u, \theta)$. After the topic is selected, the probability of picking a word is conditional on the topic, z, and ϕ (a second prior distribution of multinomial probability distributions). Therefore, the probability of selecting a word is $p(w|z, \phi)$.

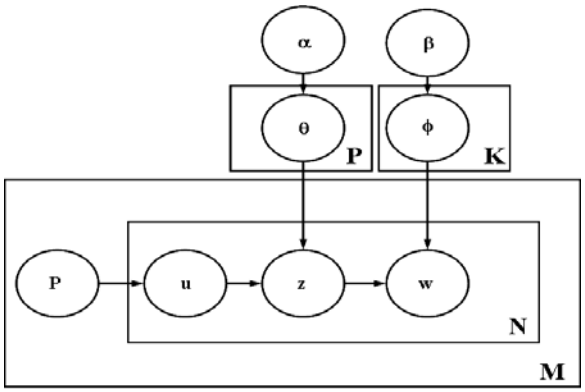

Fig. 1. Author Topic Model

The goal is to compute $p(z, \phi, \theta|d, \alpha, \beta)$. However, solving this equation is intractable. Instead Gibbs Sampling (8) is used to approximate the true probability distribution. It randomly assigns words to users and topics and calculates the resulting conditional probabilities. This process repeats until the conditional probabilities converge. Defining $n(u, z)$ as the number of times topic z is chosen from the topic distribution of the user, $n(u)$ is the number of times user u is assigned to *any* word. The definitions of $n(w, z)$ and $n(z)$ are then defined similarly, the conditional probabilities are:

$$p(z|u) = \frac{n(u, z) + \alpha}{\sum_{z'} n(u, z') + K\alpha} \tag{1}$$

$$p(w|z) = \frac{n(w, z) + \beta}{\sum_{w'} n(w', z) + V\beta} \tag{2}$$

$$p(u, z|w) = p(z|u)p(w|z) \tag{3}$$

Algorithmically:

(1) Assign random probabilities to all conditional probabilities, i.e. $p(z|w)$ and $p(u|w)$, such that they produce a probability distribution (i.e. the probabilities are all non-negative and sum to one).
(2) For every word in every document, "determine" what topic and user produced it. To do this, pick a random number between 0 and 1 and see which conditional probability it falls into.
(3) Based on the number of times each user and topic was assigned to a word, re-calculate the conditional probabilities.
(4) Repeat steps 2 and 3 until convergence.

3 Methodology

The remainder of this work focuses on applying Author Topic to the Enron email corpus (6). The Enron email corpus consists of all e-mail to and from the senior management of Enron with attachments removed. The Enron email corpus provides a good test case for the applicability of using Author-Topic to detect potential insider threats because it is data from real users at a real company, and their is enough historical information about the Enron company to identify a potential condition which should be detected. Specifically, in the Enron case, Sherron Watkins provided the authorities with information on the illegal behavior of the Enron management, Ken Lay, Andy Fastow, and Jeff Skilling. However, if we look at the question of insider threats from the point of view of the management, for example Ken Lay, then the whistleblower Sherron Watkins can be construed to be an insider threat.

By assuming this reversal of roles in this proof of concept, we demonstrate that Author Topic is a viable tool for detecting potential insider threats by identifying Sherron Watkins as an Enron insider threat. The first step is to take the Enron email and resolve it into a collection of word and individual frequency counts (i.e. the number of times each word and each individual occurs in each email). To reduce the size of the dataset and to improve clustering, prior to the generation of frequency counts, the words are stemmed and words with the same stem combined. These frequency counts are then fed into Author Topic. Two frequency distributions emerge: the number of times each word is associated with a topic and the number of times each topic is associated with an individual. Next, these frequencies are converted into four probability

distributions. The first is the probability of a word given a topic ($p(w|z)$). The second is the probability of an individual given a topic ($p(u|z)$). The third is the probability of a topic ($p(z)$). While the first two distributions follow naturally from the frequency count, the third is derived by comparing the number of times a particular topic is associated with any word to the total number of words in the corpus. The final probability distribution is the probability of a topic given a document ($p(z|d)$) which is derived by assuming that the probability of an email is proportional to the product of the probability of each of the words in the email (i.e. $p(d|z) \propto \prod_{i=1..N} p(w_i|z)$). This product is then changed to a sum of logarithms and adjusted to prevent the conditional probability of any one word being zero for a topic driving the conditional probability of the email given the topic to zero.

After the probability distributions are calculated, the next step is determining if an individual has an interest in a topic by taking the average probability of a topic and setting the threshold at a certain number of standard deviations above it. For instance, if, on average, individuals have a 0.5% probability of being interested in category 43 with a standard deviation of 0.7% and someone has a 1.7% probability of e-mailing on that topic, then (assuming a Gaussian distribution) that individual has over a 95% chance of being interested in that topic. For this experiment, 1.64 standard deviations above the mean (resulting in probabilities of at least 95%) is used to determine interest in a topic.

The next step is to construct the social networks needed to determine if individuals have a hidden interest and/or potentially feel alienated from the organization. In the same way as an individual, an email is considered to have an interest in a topic if its conditional probability is 1.64 standard deviations above the mean. It is therefore possible to create two networks for each topic. The first is an implicit interest network that links two individuals if they share an interest in the topic, i.e. the network is a fully connected graph of all individuals with an interest in the topic, shown in Figure 2. The second network is an explicit email network that links two individuals only if they have passed an email related to the topic between them, shown in Figure 3. If an individual has links in the implicit interest network but not the explicit interest network, they can be considered to have a clandestine interest in that topic. For this topic, the only individual of concern would be user 89 who has only a single link in the Explicit Social Network for the Database Topic.

After creating the social networks, it is possible for an investigator to use the social networks and the word to topic associations to determine a number of things. Specifically, an investigator would be interested in topics of a corporate security nature. For these topics, the investigator would use the social networks to identify individual who should potentially not be e-mailing about a topic, and also the possible information leak. Additionally, the process can assist managers who would be interested in individuals who do not seem to take an

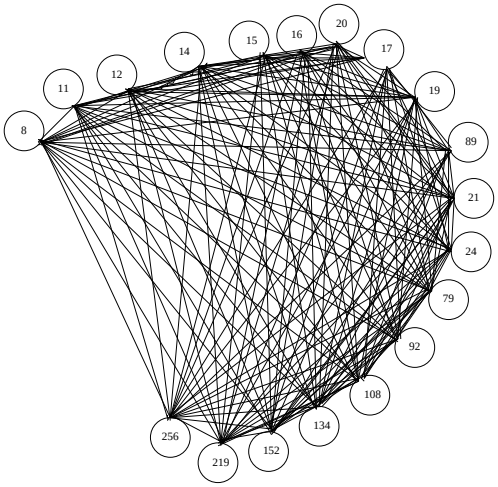

Fig. 2. PLSI-U Enron Implicit Social Network for Database Topic.

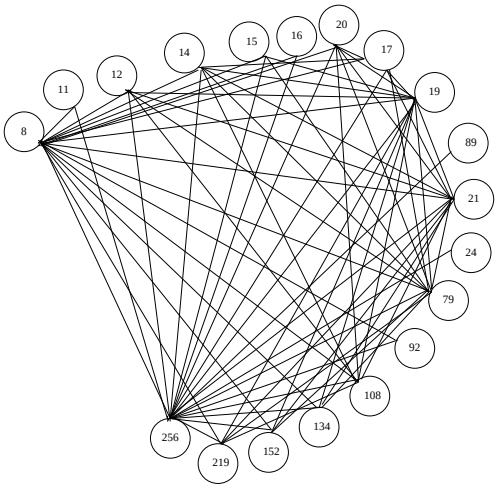

Fig. 3. PLSI-U Enron Explicit Social Network for Database Topic.

interest in the corporate environment, i.e. socializing with their co-workers. In these cases, the manager may want to intervene as this may not just be a potential insider threat but a person who is having life issues that if caught early can be more easily dealt with.

In the following section, using the Enron email corpus and corporate history, we demonstrate the validity of Author Topic as an appropriate mechanism for for identifying possible insider threats.

4 Results

The inputted parameters for Author Topic include the number of topics and the number of iterations. The number of iterations was set to 2000. The num-

ber of topics was set to 48 based on similar work done by McCallum, et al. (11).

First, the topics generated by Author Topic are examined for useability. For brevity, rather than discussing all 48 topics, only four representative topics given knowledge of the Enron corporation are discussed. While there was initial concern that stemming words might make the topics more difficult to understand, the results show stemming caused no problems while presumably improving clustering. However, rather than showing the stems, the most probable full words are extrapolated and appear in Figure 4. The first topic, Senior Mgmt, was generated by observing the preferred topics of Ken Lay, Jeff Skilling (Enron's CEO) and Andy Fastow 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 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 names, ken and lay, are words 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.

Senior Mgmt	
CATEGORY 9	
Business	1.4%
Skilling	1.4%
Lay	1.4%
Company	1.3%
Year	1.2%
Ken	1.1%
Opportunity	1.1%
President	1.0%
Chairman	1.0%

California Crisis	
CATEGORY 43	
Electricy	1.8%
Commission	1.7%
State	1.6%
Utility	1.5%
Energy	1.2%
Public	1.1%
Legislature	1.1%
Regulatory	1.0%
Senate	0.9%

Research	
CATEGORY 30	
Research	2.3%
dear	1.9%
University	1.7%
School	1.0%
Model	0.9%
Program	0.9%
Resume	0.9%
Interest	0.9%
Finance	0.9%

Info Technology	
CATEGORY 28	
Access	2.3%
Service	2.1%
User	2.1%
Information	1.6%
System	1.6%
Contact	1.6%
Manage	1.4%
Server	1.4%
Thi	1.4%

Fig. 4. Sample Categories Described by the Most Probable Words

Similar results appear when considering the individuals with the highest probability of being associated with the topics (Figure 5). However, prior to examining these individuals, their positions with Enron need to be established. Luckily, there is a copious amount of information (7; 12; 5) on the rise and fall of Enron. As a result, despite a lack of job information in the actual corpus, it is still possible to determine many individuals' positions.

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, or Ken Lay, Enron's president, emerge in the top ten, the results are still promising. Author Topic produces similar results for the California Crisis. Finding prominent public relations people like Mark Palmer and Robert Frank is encouraging. However, the large number of unidentified individuals again causes some concern. By looking at their emails, they do appear to have been involved in conducting business in California. However, what their exact positions were is unknown. 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. Finally, while Information Technology has no positions identified for individuals, the presence of emails like "SAP Security" (SAP was the corporate enterprise software package) and "Integrated Solutions Center - I/T Help Desk" shows excellent results for this topic as well.

CATEGORY 9		SENIOR MGMT
Michael Horning		0.05%
Jeff McMahon	Chief Operating Officer	0.05%
Anthony Duenner	Senior Vice Pres Global Assets & Services	0.05%
ethink		0.05%
Mitch Meyer		0.05%
All Enron Worldwide		0.05%
Matthew Scrimshaw		0.04%
Nate Ellis	Director Enron Energy Services	0.04%
Mariano Gomez		0.04%
Margaret Carson	Director of Corporate Strategy	0.04%

CATEGORY 43		CALIFORNIA CRISIS
'Kevin Fulton		0.1%
Eric Letke	Enron Energy Services	0.1%
snovose		0.1%
Robert Frank	State Government Affairs	0.1%
Hap Boyd	Enron Wind Corporation	0.1%
.sue		0.1%
Tamara Johnson		0.1%
tamara Johnson		0.1%
Mark Palmer	Head of Corporate Communications	0.1%
Becky Merola		0.1%

CATEGORY 30		RESEARCH
grant Masson	Vice President – Research Group	0.04%
Kenneth Deng	Manager of Quantitatve Research	0.04%
Mary Bailey		0.04%
Vince Kaminski	Managing Director and Head of Research	0.04%
Network Security		0.04%
Althea Gordon	Recruiter – Associates/ Analysts Program	0.03%
Jason Sokolov	Risk Management Group employee	0.03%
Lenos Trigeorgis	Risk Management Group employee	0.03%
Rehman Sharif		0.03%
Nedre Strambler		0.03%

CATEGORY 28		INFO TECHNOLOGY
houston.report		0.1%
Eric Saibi	Enron Capital & Trading – East Desk	0.1%
SAP Security		0.1%
EES Power Settlements		0.1%
subscribers@mailman		0.1%
weatherward@mailman		0.1%
Integrated Solutions Center – I/T Help Desk		0.1%
Jeffrey Jackson		0.1%
ISC Systems Notification		0.1%
Enron Users		0.1%

Fig. 5. Sample Categories and the Most Probable Individuals

In demonstrating that these results do indeed identify a potential insider, we treat the domain as an investigator might. Consider Figure 6. The first step an investigator must take is to identify topics of interest. As mentioned in Section 3, this would be a topic of associated with corporate information security. Specifically tied to the Enron case is the topic concerning the off-book partnerships. While the initial one was called Rhythms, later off-book partnerships were called Raptor I, Raptor II, Raptor III, and Raptor IV. Excluding a general "email" topic, the four topics that the word "raptor" had a non-zero conditional probability for are topics 12 ($p(w = raptor|z = 12) = 0.0011$), 25 ($p(w = raptor|z = 25) = 0.0004$), and 30 ($p(w = raptor|z = 12) = 0.0002$). Observe that topic 30 is the Research category discussed above.

This is very appropriate considering that the Research division was the first to examine and then reject the feasibility of the Raptor projects. Despite "raptor" not appearing as one of the most probable words, with most probable words like trade, agreement, credit, swap, and financial, this does appear to be a topic related to the Raptors.

STEP 1: DETERMING THE TOPIC TO INVESTIGATE

Non−zero Raptor probabilities − p(z|w)	
Topic 12 Financial Trade Agreements	0.11%
Topic 25 Financial Risk Management	0.04%
Topic 30: Resarch	0.02%

Financial Trade Agreements	
CATEGORY 12	
Trade	2.0$
Copy	1.6%
Agreement	1.3%
receive	1.2%
Executive	1.2%
click	1.2%
Credit	1.1%
Swap	1.1%
Financial	1.0%

STEP 2: FIND INDIVIDUALS WITH CLANDESTINE INTERESTS IN THE TOPIC

Clandestine Interests − Topic 12 Financial Trade Agreements			
Stacey Ramsey	Angela Liknes	K. Longoria	John Disturnal
Corbin Barnes	Ilan Caplan	Kimberly Hardy	Dave Kistler
Peter Berger	Andrea Reed	Sherron Watkins	Edosa Obayagbona
Trevor Randolph	Frank Lobdell	Mac Mclelland	Junellen Pearson
Kelly Lovvorn	Joshua Koenig	Mika Watanabe	John Bottomley
Mark Haedicke	Tori Hayden	Michelle Schultz	Esther Gerratt
Jayanta Sengupta	Nikole Vander	Mchael Nanny	Bryan Garrett
Adam Pollock	Cecil John	Carmella Jones	Victoria McDaniel
Habiba Bayi	Felicia Solis	Anita Grandos	Kimberly Nelson
James Puntumapanitch	Adriana Wynn	Jim Roth	Michael Rump
Melissa Allen	Olivier Herbelot	Nelly Carpenter	Michele Beffer
Katherine Chisley	Laura Johnson	Clay Spears	Patrick Conner
Jeffrey Austin	John Boomer	Tom Halpin	Mary Hubbard
Darla Steffes	Omar Aboudaher	Lena Kasbekar	Peter Traung
James Foster	Gardiner Corby	Robert Pickel	Duncan Croasdale
Peter Maheu	Warren Schick	Joe Hoang	Barbara Hankins
Christi Nicolay	Jay Johnson	Brenda Funk	Fabian Valle
Llewelyn Hughes	Linda Noske	Jesse Alvorado	

Fig. 6. Investigating if Sherron Watkins is an Insider

The next step is checking which individuals have clandestine interests in these topics. Although this investigation needs to be performed for all three topics, only topic 12 is shown here. 71 individuals emerge as having a clandestine interest in this topic 6.

Because there are multiple social networks over multiple topics, we can also check these 71 individuals to determine if they felt alienated at work. This occurs from calculating the intersection of the set of individuals who have a clandestine interest in the off-book partnerships and the set of individuals with a clandestine interest in socializing. For this analysis, there is no clear word that defines socializing, and so several are used. Appropriate words include *dinner, drink, fun, tonight, love, weekend, family* and *game*. In each case, only one or two topics emerge as having a non-zero probability for each word.

By performing the intersection between the social topics and the off-book partnerships, three individuals emerge as having a clandestine interest in one of the two socializing topics and the off-book partnerships topic. They are Dave Kistler, Llewlyn Hughes and Sherron Watkins. Therefore, for this experiment, Sherron Watkins emerges as possibly feeling alienated and a potential insider threat. If this had been a real world case and the CFO had combined results

between these two topics, he could have quickly zeroed in on Watkins as a possible insider threat.

5 Conclusions and Future Work

Author Topic emerges from this research as an effective tool at revealing potential insiders by datamining email. The topics generated by Author Topic are easily identifiable both based on the most probable words as well as the most probable individuals. In addition, Author Topic effectively reveals Sherron Watkins as a potential insider by flagging both her clandestine interest in the Raptor topic as well as an indication that she feels alienated from Enron. While certainly viewed negatively by her boss Andy Fastow, had her letter to Ken Lay been addressed quicker, her insider threat actions might have saved the company.

However, it is one thing to show that the signs existed that an individual was an insider after the fact. What is needed is to show that Author Topic would have revealed her before the fact. The analysis shows this as well since Sherron Watkins is one of only three individuals (out of a possible 34,000) with a clandestine interest in both the Raptor topic and a topic on socializing. If there was cause for concern that someone might leak information about the Raptors, this analysis would have generated, prior to the leak, a short list of people to pay closer attention to.

Although this technique appears promising, much work remains. When it came time to check which topics were most likely to be focused on the off-book partnerships, a problem occurred. The name of the partnerships, LJM, is not a word one finds in a dictionary. Therefore, only the most problematic Raptor transactions, not the partnership itself, could be checked. Not restricting vocabulary to words found in the dictionary might produce better results.

References

[1] Adamic, L., and Adar, E.. it Friends and Neighbors on the Web. Retrieved on 30 June, 2005 from http://www.parc.xerox.com/istl/groups/iea/ papers/web10/.

[2] AlGhamdi, G.A., Laskey, K.B., Wright, E.J., Barbara, D., and Chang, K.C. " Modeing Insider User Behavior Using Multi-Entity Bayesian Network". *International CCommand and COntrol Research and Technology Symposium*. McLean, VA, June 13-16, 2005.

[3] Butts, J.W., Mills, R.F., and Peterson, G.L. "A Multidiscipline Approach

to Mitigating the Insider Threat". *International Conference on Information Warfare and Security (ICIW)*. Princess Anne, MD, March 2006.

[4] Culotta, Aron, Bekkerman, Ron, and McCallum, A.. "Extracting social networks and contact information from email and the Web".*First Conference on Email and Anti-Spam (CEAS)*. Mountain View, CA. 2004.

[5] Eichenwald, Kurt. *Conspiracy of Fools: A True Story*. Broadway Books, New York, NY, 2005.

[6] Carnegie Melon University.*Enron Email Dataset*. March 2, 2004. Retrieved on 30 June, 2005 from http://www-2.cs.cmu.edu/ enron.

[7] FindLaw Legal News and Commentary. "Enron Investigation News". Retrieved 15-Nov-2005 from http://news.findlaw.com/legalnews/lit/enron/.

[8] Griffiths, T.L., and Steyvers, M. "Finding Scientific Topics".*Proceedings of the National Academy of Sciences*, 101. 5228-5235, 2004.

[9] Herbig, Katherine L. and Martin F. Wiskoff. *Espionage Against the United States by American Citizens 1947 - 2001*. Technical Report, Defense Personnel Security Research Center (PERSEREC), 2002. Retrieved on 15 June, 2005 from http://www.fas.org/sgp/library/spies.pdf.

[10] Kautz, Henry, Selman, Bart, and Shah, Mehul. "Referral Web: Combining Social Networks and Collaborative Filtering". *Communications of the ACM*. 40(3):63-65. 1997.

[11] McCallum, Andrew and Andres Corrada-Emmanuel and Xuerui Wang. "Topic and Role Discovery in Social Networks". *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*.

[12] McLean, Bethany and Peter Elkind. *The Smartest Guys in the Room*. Penguin Group (USA), New York, NY, 2003.

[13] Mitre. *Detecting Insider Threat Behavior*. 2004. Retrieved 25-Jan-2006 from www.mitre-corporation.com/ news/events/tech04/briefings/1344.pdf.

[14] Okolica, James, Gilbert Peterson and Robert Mills. *Using PLSI-U to Detect Insider Threats from Email Traffic*. chapter Advances in Digital Forensics. Springer-Verlag 2006.

[15] RAND. *Understanding the Insider Threat: Proceedings of a March 2004 Workshop*. 2004. Retrieved 25-Jan-2006 from www.rand.org/pubs/conf_proceedings/2005/RAND_CF196.pdf.

[16] Rosen-Zvi, Michal and Thomas Griffiths and Mark Steyvers and Padhraic Smyth. "The Author-Topic Model for Authors and Documents". *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. 487-494, 2004.

[17] Shaw, Eric and Jerrold Post and Keven Ruby. "Inside the Mind of the Insider". *Security Management*, 43:34–41.

[18] Symonenko, Svetlana, Libby, Elizabeth D., Yilmazel, Ozgur, Del Zoppo, Robert, Brown, Eric, and Downey, Matt. "Semantic Analysis for Monitoring Insider Threats". *Second Symposium on Intelligence and Security Informatics (ISI 2004)*. 2004.

[19] Yee, J., Mills, R.F., Peterson, G.L., and Bartczak, S. "Automatic Genera-

tion of Social Network Data from Electronic-Mail Communications". *Tenth International Command and Control Research and Technology Symposium.* McLean, Virginia, June 2005.