A Link Prediction Approach to Anomalous Email Detection

Zan Huang and Daniel D. Zeng, *Member, IEEE*

Abstract—**In many security informatics applications, it is important to monitor traffic over various communication channels and efficiently identify those communications that are unusual for further investigation. This paper studies such anomaly detection problems using a graph-theoretic link prediction approach. Data from the publicly-available Enron email corpus were used to validate the proposed approach.**

I. Introduction

Detecting unusual communication patterns in various means and channels of communications represents an important class of application directly relevant to security informatics [1]. In this paper, we focus on anomalous email detection. Although a precise definition of what is normal or anomalous is typically elusive, the capability of detecting such communication anomalies, which can be heuristically explained as unusual relative to known common patterns, is critically needed for surveillance and monitoring purposes. This capability is particularly important for communications channels through which voluminous information is exchanged on the real-time basis.

One important aspect of anomalous email detection (or in general communication pattern anomaly detection) is that communicative acts take place in a networked structure. In many cases communication is between not only one sender and one recipient but also one sender and multiple recipients. This paper is intended to develop a technical approach that can model such one-to-many relations and efficiently detect anomalies.

Our proposed approach is based on the idea that email communications can be viewed as a network with time-stamped links representing communications between nodes (either senders or recipients) and that future communications can be predicted through link predictions [2, 7, 8]. In this framework, anomalous communications are those very unlikely links. We use the Enron email corpus to evaluate the usefulness of our approach.

The rest of the paper is structured as follows. In Section II, we present our link prediction-based model of communications and anomaly detection. In this section, related work is summarized. Section III reports an experimental study using the Enron email dataset. We conclude the paper in Section IV by summarizing the lessons learned and discussing future research.

II. An Link Prediction-Based Approach

In this paper we develop an anomalous email detection framework applying a link prediction-based formulation. We first discuss the assumptions we are making concerning data availability and the system operating environment when applying this link prediction-based approach. We then discuss in detail the anomalous email detection problem formulation and related computational approaches.

A. Assumptions and Problem Formulation

We assume that

- The set of email accounts (both senders and recipients) is known a priori and this set remains unchanged during the system operation.
- All email communications among people in this set of email accounts are being monitored and logged.
- All email messages are time-stamped.

The first two assumptions amount to a “closed-world” assumption. We now define the anomalous email detection problem more precisely by specifying related input and output fields.

The anomalous email detection approach has the following data as the input:

- The set of email accounts
- A list of tuples, each of them capturing a specific email message with its sender, recipients, and its delivery time.

As its output, at any given time t, the system predicts an *anomaly score* denoted by A(θ , t) for each distinct sender-recipient tuple θ (θ in the form of (sender, [recipient1, recipient2, …])).

B. Computational Approaches

We first describe how the anomalous email detection problem can be modeled as a link prediction problem [5]. After that, we summarize a set of link prediction computational methods used in our study.

As in the case of time series analysis, conceptually, for anomalous email detection, we use the data in time periods t-1, t-2, …, t-g as input to make predictions on possible email communications in time period t. In our current analysis, we treat all emails in time periods t-1, t-2, …, t-g equally. Certain weighting scheme (e.g., the more recent email exchanges

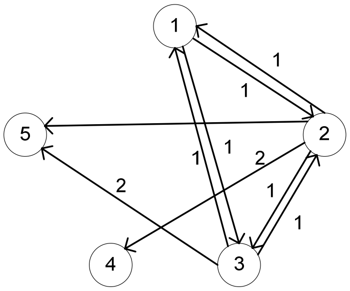
The reported research was supported in part by a grant for open research projects (ORP-0303) from the Key Lab of Complex Systems and Intelligence Science, the Chinese Academy of Sciences; and a grant (60573078) from the National Natural Science Foundation of China.

Z. Hang is with the Department of Supply Chain and Information Systems, the Smeal College of Business, Pennsylvania State University, University Park, PA 16802 USA (e-mail: zanhuang@psu.edu).

D. Zeng is with the Department of Management Information Systems, the Eller College of Management, University of Arizona, Tucson, AZ 85721 USA (e-mail: zeng@eller.arizona.edu). Zeng is also affiliated with the Institute of Automation, the Chinese Academy of Sciences, Beijing, China.

hold more predictive power) might improve the performance of the system and is planned for our future research.

In our approach, we first construct an email graph using email data in time periods t-1, t-2, …, t-g. This graph is a weighted and directed graph. An email with the sender-recipient tuple (1, [2, 3]) contributes two links 1 → 2 and 1 → 3 to the graph. The multiple occurrences of an identical link are captured in the weight of the link. For each of exposition, we denote this graph using matrix $M = (m_{ij})$, i, j = 1, …, N. The following example illustrates an email graph constructed from the following emails: (1, [2, 3]), (2, [3, 4, 5]), (2, [1, 4]), (3, [5]), (3, [1, 2, 5]).



Based on this graph-theoretic formulation, link prediction algorithms, which predict possible links based on existing observed links, can be applied to derive anomaly scores for each of the possible links or emails *i* → *j*: *P*(*i*, *j*). Note that, however, most existing link prediction algorithms only work on undirected and un-weighted graphs. Adaptation is needed for these algorithms to meet the need of anomalous email analysis. Below we summarize some of the main link prediction algorithms and the necessary adaptation needed.

a) Preferential Attachment:

This is a basic link prediction algorithm which corresponds to the preferential attachment phenomena in many real world graphs. In our domain, this algorithm basically computes the *sender degree* DS_i ($\sum_j m_{ij}$) and recipient degree DR_i ($\sum_j m_{ji}$) for each node *i*. The potential score for a link *i* → *j* is given by $P(i,j) = DS_i{*}DR_j$

b) Spreading Activation:

The main advantage of the spreading activation algorithm is its ability to capture transitive associations among the email senders and recipients. By exploring the transitive associations (paths in the email graph), we assume the following: frequent links from 1 to 2 and 2 to 3 imply that links from 1 to 3 might be likely.

The spreading activation algorithms developed in associative information retrieval can be adopted to explore transitive association efficiently. In this study we customize an algorithm with competitive performance in recommendation applications, the *Hopfield net* algorithm [4]. In this approach, we compute for each node *i* an activation level μ_j for all other nodes, $j = 1, \dots, N, j \neq i$. We first initialize the activation level to 1 ($\mu_j = 1$). Activation levels of all other nodes are set to 0. After initialization the algorithm repeatedly performs the following activation procedure: $\mu_j(t+1) = f_s\left[\sum_{i=1}^{N} \beta^{1/m_{ij}} \mu_i(t)\right]$, where f_s is the continuous *SIGMOID* transformation function or other normalization functions; β governs the decay of activation over long paths, which takes a value between 0 and 1. The algorithm stops when activation levels of all nodes converge. The final activation levels μ_j*s* give the potential scores *P*(*i*, *j*). In essence this algorithm achieves efficient exploration of the *connectedness* of a sender-recipient pair within the email graph. The connectedness concept corresponds to the total effect of the weighted paths between the pair.

c) Generative Model:

Under the generative approach, latent class variables are introduced to explain the patterns of email communications [3, 9]. Typically one can use one latent class variable to represent the unknown cause that governs the email communications. The email matrix *M* is considered to be generated from the following probabilistic process: (1) select an account as sender with probability *P*(*s*); (2) choose a latent class with probability *P*(*z*|*s*); and (3) generate an email link from *s* to *r* (i.e., add 1 to m_{rs}) with probability *P*(*r*|*z*). Thus the probability of observing an email link from *s* to *r* between *c* and *p* is given by $P(s,r) = \sum_z P(s)P(z \mid s)P(r \mid z)$. Based on the interaction matrix *A* as the observed data, the relevant probabilities and conditional probabilities are estimated using a maximum likelihood procedure called *Expectation Maximization* (*EM*). Based on the estimated probabilities, *P*(*s*, *r*) gives the potential scores.

We now describe the computational steps needed to leverage the link potential scores to detect anomalous emails.

(a) Based on the potential scores for each sender-recipient pair, we derive an estimated likelihood distribution of all possible sender-recipient tuples. The space of all possible sender-recipient tuples (any account as sender to any *n* other accounts as recipients (*n* = 1, …, *N*-1)) is large.

A key decision to make at this step is how to derive the likelihood of sender-recipient tuples from the potential scores. We adopt a method in previous studies to use the average potential score of all involved links in a sender-recipient tuple. Example: Likelihood score ((1, [2, 3])) = avg(*P*(1, 2), *P*(1, 3)). Another important modeling aspect at this step is to adjust the likelihood scores based on the number of recipients. It is often unreasonable to assume that the likelihood for an email (1, [2, 3, 4, 5]) is the same as (1, [2, 3]) when *P*(1,*j*) takes the same value, *j* = 2, 3, 4, 5. We use the distribution of the number of recipients q_k from the periods *t*-*g* to *t*-1 as the basis for such adjustment. We adjust q_k by setting all zero values to 1/10 of the smallest non-zero value and renormalize them to obtain q'_k to allow for possibility of previously unobserved recipient numbers. To account for the possibility of previously unobserved sender-recipient tuples, we similarly adjust the potential scores $P(i, j)$ obtained above (required to be non-negative) to $P'(i, j)$ by setting the zero values to 1/10 of the smallest nonzero value

(b) With the adjusted recipient number distribution q'_k and adjusted potential scores P', the probability of a particular sender-recipient tuple $\theta = (i, [j_1, j_2, \ldots, j_k])$ of k recipients is given by

$$p(\theta) = q'_k \frac{L(\theta)}{\sum_{\theta' \text{has } k \text{ recipients}} L(\theta')}.$$

It is prohibitively expensive to compute $\sum_{\theta' \text{has } k \text{ recipients}} L(\theta')$. However, due to selection of the average function to derive $L(\theta)$, we can eliminate the need to enumerate all possible sender-recipient tuples to compute the likelihood score for each possible θ. It can be shown that $\sum_{\theta' \text{has } k \text{ recipients}} L(\theta')$ is the sum of individual adjusted potential scores $P'(i, j)$s and each potential score is summed by exactly $\frac{1}{k} C_{N-2}^{k-1}$ times. Thus

$$\sum_{\theta' \text{has } k \text{ recipients}} L(\theta') = \frac{1}{k} C_{N-2}^{k-1} \sum_{i} \sum_{j} P'(i, j).$$

(c) Based on the probability of all possible sender-recipient tuples $p(\theta)$, we can compute the likelihood score for all emails in time period t: $L(t) = \prod_{\theta \text{ appeared in time } t} p(\theta)^{c(\theta)}$, where $c(\theta)$ is the number of occurrences in time period t. This likelihood score can be used to compare the effectiveness of different link prediction algorithms for email modeling. Average log likelihood score log(L(t)) / (# messages in t) can indicate the predictability of emails in time period t and enable comparison across different time periods. More importantly, an email anomaly score can be given by c(θ)/p(θ) to identify the anomalous sender-recipient tuples. A high anomalous score for θ indicates that one or more emails corresponding to the sender-recipient tuple θ is very unlikely to occur and thus may warrant investigation.

In the next section, we report an experimental study using a real-world email dataset to demonstrate the potential usefulness of the proposed link prediction-based anomalous email detection.

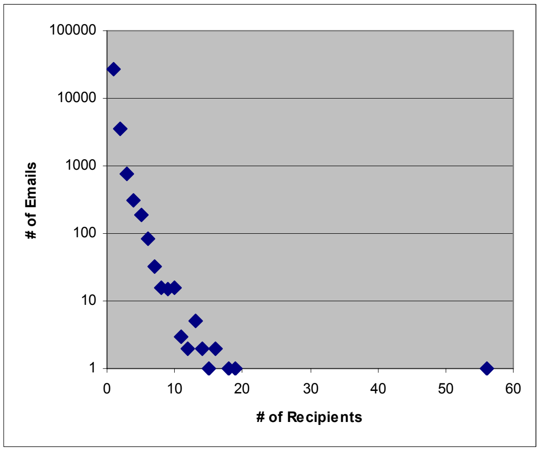
III. An Experimental Study using the Enron Dataset

A. Enron Email Data

The Enron email corpus is well suited for empirically evaluating our proposed email monitoring and anomaly detection framework. It is a large-scale email collection from a real organization with potential anomalous email communications over the course of more than 3.5 years. We used a pre-processed version of the dataset provided by Jitesh Shetty and Jafar Adibi [8] (data available at ftp://ftp.isi.edu/sims/philpot/data/ enron-mysqldump.sql.gz). This dataset contains 252,759 emails from 151 Enron employees, mainly its senior managers. According to the "closed-world" assumption of our analysis framework, in our study we have focused on emails sent from *and* to these 151 people. Thus all email communications among these 151 Enron employees are assumed to be covered by the available dataset.

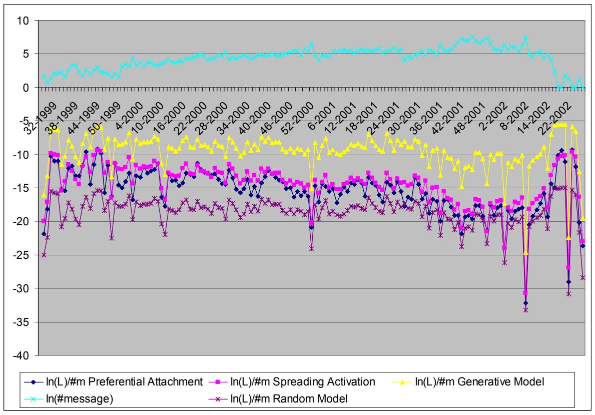
After preprocessing, the Enron email collection analyzed in our study contains 40,489 emails from May 11th of 1999 to June 21st of 2002. We performed the analysis on the weekly basis based on the volume of the email traffics. Altogether there were 159 weeks in which at least one email occurred among the 151 Enron employees. An important feature of email communication as modeled in our framework is the one-to-many communications. The distribution of the numbers of recipients in this email collection is shown in Fig. 1.

Fig. 1. Distribution of the numbers of email recipients



B. Analyzing Weekly Likelihood of Emails

Fig. 2. Weekly average log(likelihood scores) based on three link prediction models and the baseline random model.



We have applied the proposed anomaly detection framework on the Enron email dataset. The three link prediction approaches described in Section III were implemented: the Preferential Attachment, Spreading Activation, and the Generative Model. In these experiments we specified the number of lags to be 8 (emails of the 8 weeks before the current week are used to derive the link potential scores). Similar results were obtained by varying the number of lags in the neighborhood. For the Spreading Activation model we set the key parameter, the activation decay parameter β, to be 0.5 in our experiment (using other values of β did not change the results significantly). For the Generative Model we set the possible number of latent class variable values according to the number of links in each email graph for a particular time period (ranging from 6 to 12 possible values).

In addition to the three link prediction models, we also implemented a baseline random model, in which a naïve prediction is given for link potential scores: all links are equally possible. This purely random model provides a baseline model where only the previously observed distribution of the numbers of email recipients is used for anomaly detection.

The weekly average log likelihood scores obtained from the three link prediction models and the baseline random model are presented in Fig. 2. The logarithm of the number of email messages within each week is also presented. The figure shows that generally all three link prediction models fit better with the actual email data than the baseline random model. The Preferential Attachment had the worst fit among the three link prediction models. The Spreading Activation model slightly outperformed the Preferential Attachment model in general, except for a couple of weeks. The Generative Model provided a significantly better fit with the actual email data than the other two methods. This finding reflects the characteristic of the Enron email communication process: (1) emails can be thought to belong to different types with each individuals having varying sending and receiving probabilities on different types of emails; (2) individuals sending a large number of emails do not always send to people that always receive large numbers of emails; (3) the transitivity property (A is more likely to send an email to C if A often sends emails to B and B often sends emails to C) is not prominent in email communication.

By comparing average log likelihood scores of different weeks we could perform the high-level anomaly detection: to determine whether the email communication patterns in a particular time period *t* deviates significantly from previous times. Based on the Generative Model results, we observe that the overall weekly email communication patterns were relatively stable over the periods with sufficient number of emails. However, there are several weeks with significant deviation in email communication patterns (the sudden decrease of average likelihood score, for example, the 8th and 50th weeks of 2000, the 31st, 34th, 40th, 47th, and 52nd weeks of 2001, and the 6th week of 2002). Further investigation on these anomalous weeks revealed that many of these weeks involved many unusual emails with a large number of recipients.

C. Individual Anomalous Email Detection

We have also investigated the anomalous scores of individual emails to gain further insights into the effectiveness of our approach. Because emails with a large number of recipients were rare in our dataset and were easy to detect, we focused on emails with one or two recipients. In Table 1, we have shown the most unlikely and most likely emails according to the anomaly score derived from the Generative Model for the 21st week of 2001. This week was chosen because the most anomalous emails had substantially larger anomaly scores than the following emails.

Note that our detection framework only relies on the sender/recipient information of the emails (not the content of the email). The anomaly scores characterize the likelihood of the emails within the context of the email graph constructed with emails of 8 prior weeks. For example, the email sent by louise.kitchen regarding Wharton was sent to greg.whalley and andy.zipper. In the 8 prior weeks, louise.kitchen sent one email to greg.whalley and did not send any email to andy.zipper. In Fig. 3 we present the email graph based on emails during these 8 prior weeks. We observe significant network structural differences between the group and another group that relates to one of the most likely emails shown in Table 1 (the email sent by stephanie.panus to sara.shackleton and carol.clair regarding the Montana Power Company).

Table 1. Example anomalous and normal emails for the 21st week of 2001.

Number of Recipients	Most Unlikely Emails	Most Likely Emails
2	**Sender**: darrell.schoolcraft **Subject**: Outage update **Body**: --Forwarded by Darrell Schoolcraft …. Outage updateStation 1:Unit 101 was completed on the 24th…	**Sender**: stephanie.panus **Subject**: The Montana Power Company **Body**: All,Tanya thought that there were issues raised by Carol and/or Sara regar ISDA for Montana Power Company …
	Sender: louise.kitchen **Subject**: Re: Wharton **Body**: We all attended (parts of) a day long event which I believe Vince is hosting on Wharton and a project they are doing for us, ...	**Sender**: gerald.nemec **Subject**: Revised Palo Alto Docs. **Body**: Attached are the revised docs. ba discussion. Please let me know prior to to Palo Alto.
1	**Sender**: k..allen **Subject**: **Body**: http://www.caiso.com/SystemStatus.html	**Sender**: jeff.dasovich **Subject**: Re: Transwestern s comments CPUC hearing **Body**: Nice job. Only comment is that th California now find market power under e and in every pipeline. ...
	Sender: michelle.cash **Subject**: Re: Enron Law Conference **Body**: Hi, Jim.We had a great group of people working on the CLE committee this year. The committee was a combination of people identified by the various general counsels, plus …	**Sender**: carol.clair **Subject**: Status **Body**: Randy:As I mentioned in my voice you, we have accepted the CSP languag October 2002 date and with my other mo which Susan sent to you yesterday….

Although our anomaly detection is entirely derived from sender and recipient information, we do observe from Table 1 that the content of these emails was to a certain extent consistent with our detection results. For example, the email about Wharton was related to a joint project with Wharton e-Business Initiative. There are altogether three emails related to this project in our entire dataset. In contrast, the email about Montana Power Company seemed to be the continuation of a long discussion on the subject.

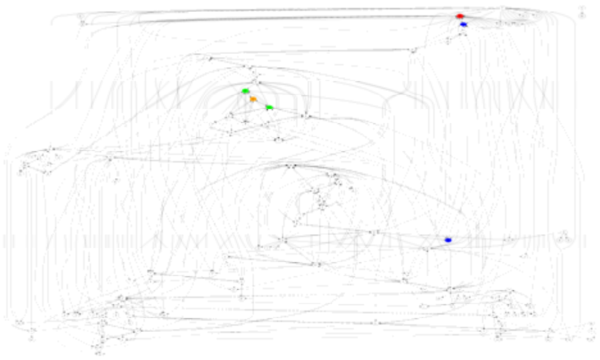


Fig. 3. Sample email graph based on emails during 8 weeks prior to the 21st week of 2001 (Network visualization by Graphviz (http://www.research.att.com/sw/tools/graphviz/)).

It is extremely difficult to rigorously evaluate the effectiveness of anomaly detection systems, mainly due to the lack of a gold set of true anomalous emails [6]. The example results only provide anecdotal evidence on the anomalous nature of the detected emails. As another attempt to demonstrate the detection quality, we generated five fictitious emails by randomly selecting senders and recipients (number of recipients ranging from 1 to 4) and insert them into different time periods. These faked emails can be generally treated as anomalous emails assuming that the true email communication process is not a random process. For all five fake emails, all three detection models generated gave the highest anomaly score among all emails with the same number of recipients, although with different numbers of real emails with the same ranking. For example, a faked email with one recipient in the first week of 2002 was ranked as one of the two most unlikely emails by the Generative Model (one of three by the Spreading Activation Model and one of ten by the Preferential Attachment Model). The relative fake email detection quality of the different link prediction models were consistent with the overall fit with the actual email data shown in Figure 1.

Year	Week	Total Number of Emails	Sender Id	Recipient Ids	Number of Recipients (k)	Number of k-recipient Emails	Email Likelihood Score Ranking within All k-recipient Emails (# of Real Emails with the Same Ranking)		
							Preferential Attachment	Spreading Activation	Generative Model
2000	13	28	62	32, 87	2	9	1st (2)	1st (1)	1st (1)
2001	1	47	72	3, 33, 53, 143	4	3	1st (0)	1st (0)	1st (0)
2001	13	105	87	1, 11, 111	3	6	1st (1)	1st (1)	1st (1)
2002	1	96	91	49	1	66	1st (10)	1st (3)	1st (1)
2002	13	21	42	92	1	17	1st (2)	1st (0)	1st (0)

IV. CONCLUSION

In this paper, we develop an email communication anomaly detection framework using link prediction methods. Obviously this approach can be used to analyze other types of communication channels (telephone, Webpage linkage, etc.) as well. One of the key intended technical contributions of our research is concerned with modeling the one-to-many relationship between a sender and potentially multiple recipients. We are currently further evaluating our proposed approach by conducting additional computational experiments. We are also developing formal graph-theoretic models to explicit capture hyper-arcs that link multiple nodes together for monitoring and anomaly detection purposes.

REFERENCES

[1] Badia, Antonio and Mehmed M. Kantardzic, "Link analysis tools for intelligence and counterterrorism.," *Proceedings of the IEEE International Conference on Intelligence and Security Informatics*, Atlanta, GA, 2005, 49-59.

[2] Diesner, Jana and Kathleen M. Carley, "Exploration of communication networks from the Enron email corpus," *Proceedings of the Workshop on Link Analysis, Counterterrorism and Security at the SIAM International Conference on Data Mining*, Newport Beach, CA, 2005.

[3] Hofmann, Thomas, "Latent semantic models for collaborative filtering," *ACM Transactions on Information Systems*, 22, 1, (2004), 89-115.

[4] Huang, Z., H. Chen and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," *ACM Transactions on Information Systems (TOIS)*, 22, 1, (2004), 116-142.

[5] Lin, Shou-de and Hans Chalupsky, "Unsupervised link discovery in multi-relational data via rarity analysis," *Proceedings of the 3rd IEEE International Conference on Data Mining*, Melbourne, FL, 2003, 171-178.

[6] Priebe, C.E., J.M. Conroy, D.J. Marchetteand Y. Park, "Scan statistics on Enron graphs," *Computational and Mathematical Organization Theory*, 11, 3, (2005), 229-247.

[7] Rattigan, M. and D. Jensen, "The case for anomalous link detection," Proceedings of the 4th Multi-Relational Data Mining Workshop, 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, 2005,

[8] Shetty, J. and J. Adibi, "The Enron dataset database schema and brief statistical repront," (2005).

[9] Ungar, L. H. and D. P. Foster, "A formal statistical approach to collaborative filtering," *Proceedings of the CONALD'98*, 1998,