Link-based Event Detection in Email Communication Networks

Xiaomeng Wan
Faculty of Computer Science
Dalhousie University
Halifax, Canada
xwan@cs.dal.ca

Evangelos Milios
Faculty of Computer Science
Dalhousie University
Halifax, Canada
eem@cs.dal.ca

Nauzer Kalyaniwalla
Faculty of Computer Science
Dalhousie University
Halifax, Canada
nauzerk@cs.dal.ca

Jeannette Janssen
Dept. of Mathematics & Statistics
Dalhousie University
Halifax, Canada
janssen@mathstat.dal.ca

ABSTRACT

People's email communications can be modeled as graphs with vertices representing email accounts and edges representing email communications. Email communication data usually comes in as continuous data stream. Event detection aims to identify abnormal email communications that serve as analogs of real-world events imposed upon the data stream. The goal is to understand the communications behaviors of the subjects. The contents of emails are often not available or protected by privacy, which makes linkage information the only resource we can rely on. We propose a link-based event detection method that clusters vertices with similar communication patterns together and then, considers deviations from each vertex's individual profile, as well as its cluster profile. Experiments show that this method performs well on both Enron and our own email datasets.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Sociology

Keywords

Event detection, communication networks

1. INTRODUCTION

Email has become an extremely popular communication tool in these days. Email communications among a group of people (e.g. friends, colleague) form a communication network. Studying this network can reveal people's communication behaviours. Event detection is to identify abnormal email communication patterns in this network caused by real world events. Since email communication data usually comes in as continuous data stream in vast volume, it is difficult to analyze their contents to identify events. Moreover, the contents of emails are often unavailable or protected by privacy regulations. Under such circumstances, linkage information becomes the only resource we can count on. Link-based event detection is to identify events using linkage information only and thus, is fast and requires less storage. We build a graph to represent the email communication network where vertices are email accounts and edges are email communications. Since email communications are temporal in nature, the edges of the graphs appear and disappear in time, resulting in a dynamic graph. The graph formed is a manifestation of the underlying social interactions. Consequently, mining the dynamic graph could identify abnormal behaviours caused by real-world events. Methods that apply to static graphs are not sufficient for dynamic graphs due to their constantly changing edges. A widely used simplification is to break the observation period into a number of time intervals and build a static graph summarizing the dynamic graph for each time interval [3]. Thus, the dynamic graph is transformed into a series of static graphs.

While event detection on email communications is rarely researched, techniques have been developed for other application domains, such as network intrusion detection and financial fraud detection. In general, there are two categories of event detection techniques on time series data - individual based techniques and cluster based techniques. They use different deviations served as evidence of abnormal behaviours: individual deviation and cluster deviation. Individual based methods build a profile for each entity based on its historic behaviour and then, measure the deviation of its current behavior from this profile. More specifically, such methods identify events by looking for variations using models like Hidden Markov model [5], change-point detection [4] and scan statistics [6]. Cluster based methods cluster entities with similar behaviors together, and measure cluster deviation for each entity as the distance of its current behaviour from the cluster center. Based on our experience of applying them on email datasets, both methods generate many false alarms.

SAC'09 March 8-12, 2009, Honolulu, Hawaii, U.S.A.

Event detection techniques on communication networks usually assume that, explicitly or implicitly, events cause variations on communication volume. While variations in communications volume may signal events, not every event worth detecting will result in a variation in the communications volume. For example, people may change their jobs or relocate to other cities. These kinds of changes may not be reflected on communication volume, but maybe reflected on some other features like the change of recipients. To capture the entire spectrum of events, we need to examine a set of features designed to capture a variety of deviations in communications.

In this paper, we propose a method that uses multiple link-based features and both individual and cluster deviations to infer the presence of events. The relationship between these two deviations is investigated and the method is evaluated on both Enron dataset and the email datasets from our department. The experimental results show that our method can identify various events and effectively reduce false alarms.

2. OUR METHOD

Methods based on cluster deviations identify those behaviours far away from the cluster profile as events. It faces problem when dealing with vertices whose normal behaviours are already far away from the cluster center (i.e. those vertices at the edge of the cluster). For those vertices, small variations from their typical behaviours can drive them out of the normal scope. Methods based on individual deviations identify events as communication patterns far away from their individual profiles, that is, behaviours never seen before. However, behaviours not seen before are not necessarily events. They can be normal behaviours for this kind of vertices that just never happened for this vertex. Without further discrimination, it leads to false alarms.

Based on these observations, we believe that under normal circumstances, a vertex's normal behaviour should be close to the typical behaviour of its cluster (cluster profile) and, at the same time, some variation around its own typical behaviour (individual profile) should be allowed. Based on this concept, we define "events" as behaviours with high scores on both individual and cluster deviations. In this paper, we use "events" and "anomalies" interchangeably.

The advantages of our approach include: first, it uses link-based features only and thus, can be used in situations where contents are not available; Second, it uses multiple features that capture different aspects of the network and, as a result, can detect various events.

2.1 Feature set

The feature set is selected with two goals in mind: discriminating between vertices and capturing events. The problem has two constrains. First is that the email data appears as a stream and the second is that the individual elements of the stream are large graphs containing hundreds to thousands of vertices. The first constraint allows only one pass of data, while the second constraint makes it impractical to use features that need the global structure of the whole network and, require intensive computation. Besides these, other criteria of choosing features are: first, they should characterize people's communications from different viewpoints; second, their variations should relate to events of interest. We have selected the following features based on the above criteria. They are divided into two categories: individual and neighbourhood features. Individual features focus on the center vertex's communications with its direct neighbours, while neighbourhood features focus on the impact that those communications bring to its neighbourhood. The feature set is shown in Table 1. A "new" link means that the sender never sent email to the recipient before.

Table 1: Overview of features

Features	Description
IN	number of people received emails from
OUT	number of people send emails to
NO	number of new recipients
IR	number of received and replied emails
OR	number of sending out emails get replied
AR	average emails to each recipient
RG	times of current recipients appear as recipients in previous 10 weeks
CC	number of emails between recipients
ICC	number of emails between people received emails from
NN	number of new emails between recipients

2.2 Dynamic graph

First, the time span of the dynamic graph G is split into disjoint time intervals (e.g. one hour, one day). Within each time interval t, a static graph $g(t)$ is built to summarize the dynamic graph. In other words, all the edges that ever appeared during this time interval are kept in the static graph. For each vertex v_i in $g(t)$, link-based features are extracted and a feature vector $V_i(t)$ is generated. Individual and cluster profiles are built based on these feature vectors in the training period.

2.3 Individual deviation

The profile of each vertex v_i is built based on its feature vectors in the training period (T, T'). It consists of individual center μ_i defined as the mean vector of its time series of feature vectors and covariance matrix Σ_i that captures the correlations of the features. After the profile is built, individual deviation $di(v_i, t)$ of v_i at time interval t is computed as the Mahalanobis distance from its current feature vector $V_i(t)$ to its individual center μ_i.

2.4 Cluster deviation

The clustering is based on the vertices' median feature values in the training period. We use medians because they are resistant to outliers. The vector of median feature values for a vertex v_i in the training period (T, T') is denoted as M_i. Vertices are first clustered using model-based clustering [2]. It assumes that the data are generated by a mixture of probability distributions in which each distribution forms a cluster. The profile for cluster c_j with n vertices consists of cluster center μ_{c_j} defined as the mean vector of its members' median vectors, and covariance matrix Σ_{c_j} as well. Cluster deviation $dc(v_i, t)$ of v_i at time interval t is defined as the Mahalanobis distance from its current feature vector $V_i(t)$ to its cluster center μ_{c_j}.

After these two deviations are computed, "anomalies" are identified as vertices whose behaviours have high scores

on both individual and cluster deviations (i.e. exceeding thresholds α and β respectively).

$$\{v_i | di(v_i, t) > \alpha \text{ and } dc(v_i, t) > \beta\}, \quad (1)$$

3. EXPERIMENTS AND RESULTS

We test our method on two datasets, the Enron email dataset and the CS dataset that is derived from the log files on our own email server. Since no events are labeled for either data, we cannot ensure clean data for training purpose and have no ground truth for evaluation. We made two changes to our method to adapt to this situation. First, we use robust estimation to compute individual profiles. Robust estimation can estimate the center and variance matrix for data with outliers [7]. Second, we develop a novel evaluation metric based on the assumption that anomalies do not last longer than one time interval (i.e. one week here). It is called Back to Normal (BTN) ratio, which is defined as the percentage of detected anomalies in time interval t that return back to normal in the next time interval $t+1$. As long as false alarms at t still get caught as anomalies in next time interval $t+1$, BTN ratio is equal to $precision \times \lambda$, where λ is the percentage of real anomalies lasting less than one time interval. When the assumption is true (i.e. λ is 1), the BTN ratio is equal to the precision. If the assumption is only partially true, the BTN ratios are still comparable assuming that, for a specific dataset, λ is fixed. So we can use the BTN ratios to compare the detection results.

$$BTN(t) = \frac{|\{v_i | V_i(t) \text{is anomaly and } V_i(t+1) \text{is normal}\}|}{|\{v_i | V_i(t) \text{is anomaly}\}|} \quad (2)$$

3.1 Enron email dataset

We first applied our method to the Enron email dataset. It is derived from Enron employees' email folders. In this project, we use the processed version from [1]. We focus on the time period from 1999 to 2002 and break it into weeks. We use the first 120 weeks as the training period to build cluster and individual profiles. Among its 184 email accounts, 122 sent out at least one email during the training period. Our following study focuses on these 122 active accounts.

The relationship between individual and cluster deviations is shown in Fig. 1 for one time interval (i.e. week 124) . Each point represents one vertex. As we can see, most vertices have small scores on both individual and cluster deviations, indicating that they are close to both their individual profiles and cluster centers. Vertices with high individual deviations but small cluster deviations represent behaviours drifting away from individual profiles but actually moving closer to their cluster profiles. As we mentioned above, those behaviours are new for individuals but normal for their types, corresponding to the false alarms generated by methods based on individual deviation. Some other vertices have small individual deviations but high cluster deviations. They represent normal behaviours of those vertices whose typical behaviours are already far away from their clusters, corresponding to the false alarms generated by methods based on cluster deviation. Details of some false alarms that have high scores on one deviation but small scores on another are shown in Table 2.

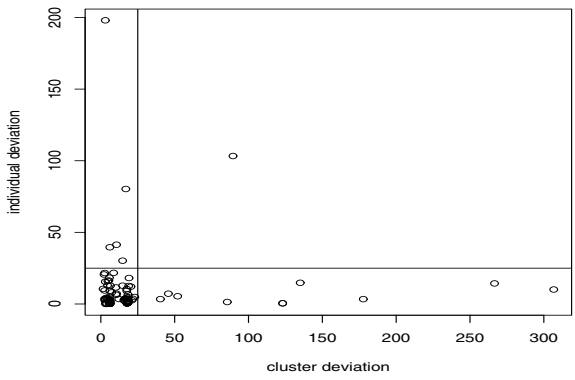

Figure 1: Relationship of cluster and individual deviations of week 124

Most of the false alarms are caused by people sending multiple emails to the same recipients (low outdegree and large number of emails per recipient). As we can see, the multiple emails were not necessarily on the same topic or related to each other. This kind of behaviour may get high score on either cluster or individual deviation depending on their individual profiles and the clusters that they belong to, but could not result in high scores on both.

The only vertex with high scores on both deviations in this week is Theresa Staab. The details of this event, as well as some events detected in other weeks using the threshold ($\alpha = 20, \beta = 80$), are shown in Table 3. All these events were caused by somebody sending emails to a group of recipients (large outdegree) and followed by discussions between the recipients (large number of links between neighbours). The first event was caused by Theresa Staab sending emails to Debra Perlingiere and Gerald Nemec talking about "May Confirmation Information", and Debra and Gerald sent emails to each other too talking about "Form Confirmation". Same thing happened in the third event when Mark Haedicke sent out an email to remind everybody the Magic of Myth exhibition at The Museum of Fine Arts, Houston, followed by discussions between the recipients about the catering and invitation. Most of the events that we detected in Enron dataset are of this type. Though we don't have the insider information to verify the linkage of some of those emails, it is clear that this is the typical model for events in such a corporation environment.

Fig. 2 show the BTN ratios for different threshold combinations (i.e. different α and β). When the threshold for cluster deviation (β) is zero, it means that we are using individual deviations only without considering the cluster deviation, and vice-versa. The BTN ratios show that our method performs much better than using either individual or cluster deviation alone. The best BTN ratio is achieved at $\alpha = 20$ and $\beta = 80$.

3.2 CS email data

Table 2: Details of false alarms on Enron data

center vertex	feature vector	$di(v_i, t)$	$dc(v_i, t)$	people involved	topic
Jeffery Skilling	3 1 0 0 0 7 2 0 2 0	10.1	306.6	Joannie Williamson	personal schedule
Dutch Quigley	1 1 0 0 0 9 3 0 0 0	14.3	266.5	Errol McLaughlin	gd to nymex, Bank of America
Greg Whalley	3 1 0 0 0 3 0 0 2 0	3.4	177.7	Rick Buy	Board of Mariner meeting
James Derrick	2 1 0 0 0 13 6 0 0 0	80.3	16.8	j.harris	May 16 clips, concur expense document,...
Stanley Horton	1 5 1 0 0 3.2 11 0 0 0	198	3.1	Rod Hayslett,...	clean fuels, staff meeting,...

Table 3: Details of identified events on Enron data

week	center vertex	feature vector	$di(v_i, t)$	$dc(v_i, t)$	people involved	topic
124	Theresa Staab	2 2 1 1 1 4 4 2 0 0	103.3	89.6	Debra Perlingiere, ...	May Confirmation Information
125	Joannie Williamson	3 15 4 0 1 4.13 1 27 1 3	714.7	156.2	Danny McCarty,...	Jeff Skilling at Houston Tech. Forum
127	Mark Haedicke	6 15 9 1 2 1.93 15 48 4 1	71.8	141.6	Dan Hyvl,...	Star Wars: The Magic of Myth
145	Mike Mcconnell	4 13 5 1 3 8.15 2 33 2 2	205.9	83.2	David Delainey, ...	Jeff Huff for VP

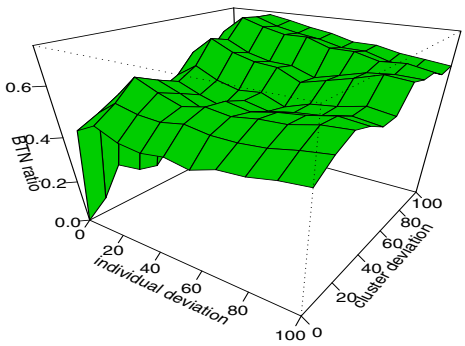

Figure 2: Precision from different α and β

The CS email dataset is derived from the log files on the email server of the Faculty of Computer Science of Dalhousie University, starting from May 2004. Overall, there are 26,425 email addresses involved. The log files record information such as the time, sender email addresses and recipient email addresses. The contents of the emails are not accessible and the email addresses were anonymized before being used in this project due to the privacy policy. Once again, we break the study period into weeks and use the first 50 weeks as the training period. We only examine active CS accounts (defined as accounts from which at least one message has been sent out in at least 10 weeks during the training period). This results in 633 unique e-mail accounts. We apply our method to detect events in the CS email data from week 51 to 55. Details of some detected events are shown in Table 4.

As we can see, the vertices are picked up for different reasons. Some are caused by people sending out a large number of emails to many recipients. The recipients were sometimes old friends (small number of new links, i.e. event 4 and 7) or completely new (large number of new links, i.e. event 1 and 2). The former case indicates that they were raising interesting topics within his old friends and which in turn, triggered

Table 4: Details of identified events on CS data

ID	week	vertex	feature vector
1	51	1183	10 264 246 2 4 1 14 341 21 87
2	51	689	4 25 25 0 0 1 0 2 2 0
3	52	332	34 3 0 2 3 1 10 6 238 0
4	52	2721	6 18 1 2 4 1.06 21 113 6 9
5	54	1140	6 15 12 1 1 1.13 2 97 5 37
6	55	836	0 42 23 0 0 3.19 27 21 0 0
7	55	4	28 207 5 16 16 1.54 304 343 119 38

discussion between them. It also means that some kinds of communities were formed especially when the communications between the recipients were dense (large number of links between neighbours). The latter case with large number of new links indicates that the person was sending emails to somebody to whom he never sent email, which should raise an alert that something unusual happened. The third event was caused by a person who received large number of emails and as we can see, the communications between those senders were quite dense. In general, using multiple features does give us the advantage of identifying various types of events with different causalities and effects. Based on the BTN ratios shown in Fig. 3, our method performs better than the methods using either deviation alone. It achieves a 100 percentage BTN ratio with high α and β combinations. Without ground truth, we couldn't verify whether it is sacrificing the recall ratio to achieve that, which could be avoided in practice by manually checking the identified events.

4. CONCLUSIONS AND FUTURE WORK

A link-based event detection method is proposed in this paper, which is based on two different aspects: deviation from the individual pattern, and deviation from the cluster pattern. Experiments on email communication datasets show that this method can captures a variety of events and is robust to false alarms. The results also show that the two parameters α and β are data-dependent and, thus, need to be tuned for each dataset. A training process could be introduced for this purpose. In the future, we want to experiment more on feature selection and parameter tuning.

5. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the MITACS NCE and NSERC.

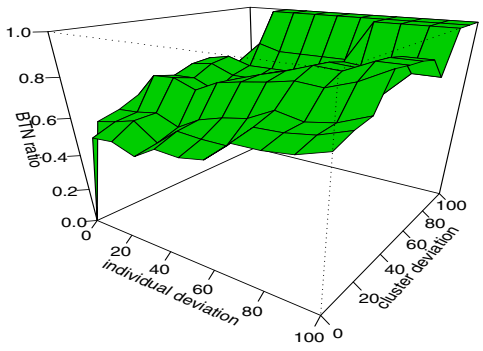

Figure 3: BTN ratios from different α and β

6. REFERENCES

[1] Scan statistics on enron graphs. http://cis.jhu.edu/ parky/Enron/enron.html.

[2] J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803–821, Sep. 1993.

[3] C. Cortes, D. Pregibon, and C. Volinsky. Computational Methods for Dynamic Graphs. *Journal of Computational and Graphical Statistics*, 12(4):950–970, 2003.

[4] V. Guralnik and J. Srivastava. Event detection from time series data. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–42, New York, NY, USA, 1999. ACM Press.

[5] A. Ihler, J. Hutchins, and P. Smyth. Adaptive event detection with time-varying poisson processes. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 207–216, 2006.

[6] C. Priebe, J. Conroy, D. Marchette, and Y. Park. Scan Statistics on Enron Graphs. *Computational and Mathematical Organization Theory*, 11(3):229–247, 2005.

[7] P. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.