Detecting Network-wide Traffic Anomalies based on Spatial HMM

Min Li, Shunzheng Yu, Li He
Dept. of Electronics and Communication Engineering,
Sun Yat-Sen University, Guangzhou, China, 510275
limin@mail2.sysu.edu.cn,syu@mail.sysu.edu.cn,heli_hefen@hotmail.com

Abstract

In contrast to many techniques exploiting temporal patterns of traffic from a single network element, network-wide traffic analysis mainly focuses on the spatial behavior across the whole network. This paper proposes a spatial hidden Markov model (SHMM) to learn the normal patterns of network-wide traffic. Combined with topology information, SHMM models traffic volumes on links as probabilistic outputs of underlying interactions between routers. Based on a trained SHMM, a nonparametric CUSUM algorithm is used to track the change of entropy of observation sequences in different sliding windows for anomaly detection. Background traffic collected from real network and synthetic anomalies are used for validation of the detection method. The results prove our method effective for network-wide traffic anomaly detection.

1. Introduction

In a large scale network of a large enterprise or an ISP, traffic anomalies such as DDoS, flash crowds, equipment misconfiguration and outage may occur quite frequently. These anomalies not only have a great impact on end users, but also may cause degradation of overall network performance. So it is very important for network administrators to keep a close eye on the variation of network traffic in real time in order to detect such anomalies as soon as possible.

There is a large literature on anomaly detection of traffic volume [1,2,3], but majority of them only focus on traffic from a single network element (link, router interface, or ingress/egress point) in isolation. These detecting schemes take a single traffic time series as input and exploit the temporal pattern of the input data for anomaly detection. And these techniques may work well at user end or at some critical point of a network. But at the view of network administrator, there may be tens even hundreds of routers and links in an ISP or enterprise network. Since we don't know in advance whenever and wherever an anomaly may occur, considering traffic of only one element is limiting. In order not to miss any anomaly in the whole network, deploying detection systems at every possible network element is straightforward. However it is very costly and may be not a good and effective solution. So there is a need for a novel scheme to combine data from multiple network elements.

In contrast to time series methods, modeling the spatial pattern across multiple sources is another reasonable means to explore traffic anomalies. Imagining an anomalous scene under DDoS attack, all the malicious traffic is aggregated at the victim end through different paths of the network. Since malicious traffic has to traverse multiple links on its way to the victim, the anomaly may not only be evident at the victim end, but also be possible to change the traffic volumes of the links involved. The unusual pattern of a set of links reflects the deviation of network behavior. Taking all the links into consideration may give more evidence to the detection and may be able to get an overall evaluation of network performance. So here comes the problem of network-wide analysis which is a simultaneous treatment of traffic from multiple or all network elements.

In this paper, we take traffic volumes on all links as observation variable. The change of network behavior can be characterized as a time series of such a multi-dimensional variable. Unlike usual multi-dimensional variable, traffic volumes on different links have special relationship based on the network topology. We consider the distribution of traffic volumes across all the links is controlled by interactions between all the routers. So we propose a spatial hidden Markov model (SHMM) which uses the network topology as its basic structure. Every router acts as a hidden state, and traffic from one router to another one is controlled by the transition probability of these two states. Using a trained SHMM, the change of network behavior can be

978-0-7695-3354-4/08 $25.00 © 2008 IEEE
DOI 10.1109/NPC.2008.89

detected by the variation of likelihood of the observation sequence.

The paper is organized as follows. In section 2, we describe the basic ideas of SHMM. In Section 3, we explain an anomaly detection method based on a non-parametric CUSUM algorithm. In Section 4, we describe the experiment setting and discuss the results. In Section 5, we summarize the recent related work. Finally in Section 6, we make a conclusion and discuss further extensions.

2. Spatial HMM for Normal Traffic

A large network is composed of routers that are connected by links. Based on network topology, traffic across network can be viewed as a weighted directed graph. Nodes in graph represent routers when arcs represent links. Traffic volumes on links act as weights of arcs which reveal interactions between routers.

We use a traffic matrix $f(t)=\{f_{ij}(t)\}$ to represent traffic across network at time *t*. $f_{ij}(t)$ is volume of traffic from router *i* to router *j*. If there is no direct link between router *i* and *j*, $f_{ij}(t)=0$. And $f_{ii}(t)$ is traffic dropped at router *i*. Since data packet passing a network is delivered hop by hop (or router by router), we assume a_{ij} as the probability that traffic at node *i* would be forwarded to node *j*,

$$a_{ij}=\overline{f_{ij}}/\sum_{m}\overline{f_{im}} \qquad (1)$$

where $\overline{f_{ij}}$ is the mean value of $f_{ij}(t)$. When a packet arrives at node *i* , it will be forwarded to node *j* based on a_{ij} or be dropped at node *i* based on a_{ii}. So the delivery process of such a packet is controlled by a Markov chain $A=\{a_{ij}\}$. Since traffic is composed of packets, we think that all the routers construct a Markov chain to control the distribution of traffic across network, as shown in figure 1.

Due to fluctuation of traffic, *A* is not a constant matrix. But we assume that during the relative stationary periods of traffic (busy period or free period on weekdays), the Markov model in figure 1 is homogeneous. And the variation of *f(t)* can be viewed as a random process controlled by *A*. So we can build a spatial HMM to learn the profile of normal traffic.

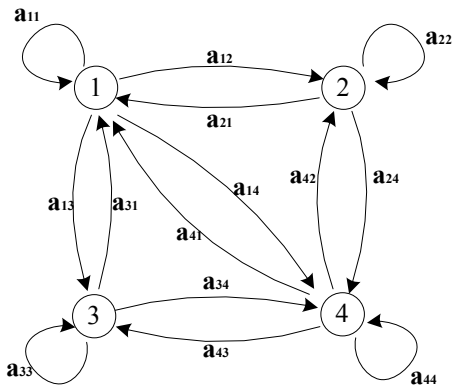

Figure 1. **Markov model of routers**

2.1 Basic parameters of Spatial HMM

An HMM[4] is a doubly stochastic process with an underling stochastic process that is not observable, but can only be observed through another set of stochastic processes that produce the sequence of observed symbols. An HMM is characterized by five-tuple $\lambda=\{S,V,\pi,A,B\}$ where $S=\{S_1,S_2,\ldots,S_N\}$ and $V=\{v_1,v_2,\ldots,v_K\}$ are the set of hidden states and the output symbols. $\pi=\{\pi_i\}$ is the initial state distribution and $A=\{a_{ij}\}$ is the state transition probability distribution. And *B* is the observation symbol probability distribution.

For our SHMM, We take all the nodes as the hidden states and add a node to represent the outside network. In contrast to an ergodic(fully connected) HMM, the connections between states are the same as the network topology. It means that if there is no direct link connecting node *i* and node *j* in the topology, $a_{ij}=a_{ji}=0$ will be set in A.

We choose byte counts on links as raw data. For traffic matrix *f(t)*, we use equation (2) to change *f(t)* to observation value $O(t)=\{O_{ij}(t)\}$ as input of SHMM,

$$O_{ij}(t)=\frac{f_{ij}(t)-\mu_{ij}(t)}{\sigma_{ij}(t)} \qquad (2)$$

where $\mu_{ij}(t)$ and $\sigma_{ij}(t)$ are the mean and variance of previous 30 measurements of f_{ij} before time *t*.

We consider SHMM as an arc-emission HMM where the symbol emitted at time *t* depends on both the state at time *t* and at time *t+1*. So we can denote observation symbol probability distribution $B=\{b_{ij}(v_k)\}$ as

$$b_{ij}(v_k)=P[O_{ij}(t)=v_k \mid q_t=S_i, q_{t+1}=S_j] \qquad (3)$$

2.2. Training of SHMM

Estimation of model parameters is done by forward-backward algorithm [4]. Given a fixed length observation sequence $O_1^T = O_1 O_2 ... O_T$, we define a forward variable

$$\alpha_t(i) = \Pr[O_1^t, q_t = S_i \mid \lambda] \quad (4)$$

and a backward variable

$$\beta_t(i) = \Pr[O_{t+1}^T \mid q_t = S_i, \lambda] \quad (5)$$

For simplicity, we use $O_{ij}(t)$ to represent $O(t)$. The forward-backward algorithm is as below:

Forward procedure:

$$\alpha_1(i) = \pi_i \qquad i \in S \quad (6)$$

$$\alpha_{t+1}(j) = \sum_{i \in S} \alpha_t(i) a_{ij} b_{ij}(O_{ij}(t)) \qquad 1 \le t \le T, j \in S \quad (7)$$

Backward procedure:

$$\beta_{T+1}(i) = 1 \qquad i \in S \quad (8)$$

$$\beta_t(i) = \sum_{j \in S} a_{ij} b_{ij}(O_{ij}(t)) \beta_j(t+1) \qquad 1 \le t \le T, i \in S \quad (9)$$

Using an iterative procedure called Baum-Welch method (an example of Expectation-Maximization algorithm) based on forward and backward variables, the training procedure may lead to a local maxima. Since just one training sequence may be insufficient for the estimation of all the parameters, we also use the training algorithm based on multiple observation sequences in [4] to get a more precise model.

3. Anomaly Detection Based on SHMM

Given an trained SHMM for normal patterns, we would like to compute P(O| λ) – the probability of $O=\{O_1 O_2 ... O_T\}$ to examine how well the model matches the observation sequence.

$$P(O \mid \lambda) = \sum_i \alpha_{T+1}(i) \quad (10)$$

And we use the entropy or average log likelihood to measure the normality of O.

$$ent = \ln(P(O \mid \lambda) / T \quad (11)$$

Since monitoring of network traffic is a continuous job, the observation sequence would have an unlimited length which contains measurements from many time slots. However, the network administrator may have major interest in the network situation of the recent period rather than from the beginning. Furthermore a too long observation sequence may not be sensitive and quick enough to reflect the current change. So we divide the long observation sequence into short sequences in a sliding way.

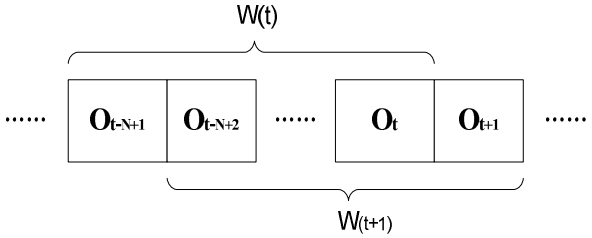

Figure 2.**Sliding windows for observation sequences**

Figure 2 shows two successive sliding windows of fixed-size N both containing N observation symbols. $W(t)$ is the sliding window at time t. The sliding window moves forward one symbol at a time. It means that $W(t)$ contains $O(t)=\{O_{t-N+1} O_{t-N+2} ... O_t\}$ while $W(t+1)$ contains $O(t+1)=\{O_{t-N+2} O_{t-N+3} ... O_{t+1}\}$.

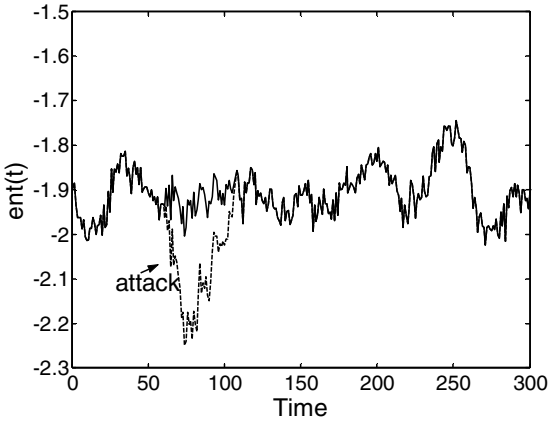

Figure 3. **Entropy varying with time**

Let $ent(t)$ be the entropy of observation sequence in $W(t)$. The variation of $ent(t)$ may reflect the change of network traffic. Figure 3 shows an unusual change of $ent(t)$ under DDoS attack comparing to the same background traffic without anomalies. While network traffic was under normal condition, $ent(t)$ would fluctuate in a quite small range. Then $ent(t)$ began to drop right after the beginning of the attack. Even after the attack stopped, $ent(t)$ still kept at a lower level because the sliding window had part of the measurements from the attack period. So we would like to use a nonparametric CUSUM to track the change of $ent(t)$ and to signal alarms.

CUSUM algorithm is a popular change-point algorithm which accumulates deviations of incoming measurement $\{X_t\}$ relative to a specific value. Since we don't know the distribution of $ent(t)$, we would like to use a nonparametric CUSUM as in [9] :

$$Z_t = \delta - ent(t) - d \quad (12)$$

where δ is the mean of *ent(t)* which can be updated by EWMA. And *d* is the upper bound of $\delta - ent(t)$.

$$S_t = \sum_{i=0}^{t} Z_i, S_0 = 0 \tag{13}$$

$$Y_t = S_t - \min_{1 \le k \le t} S_t = (Y_{t-1} + Z_t)^+, Y_0 = 0 \tag{14}$$

where

$$X^+ = \begin{cases} x, x > 0 \\ 0, x \le 0 \end{cases} \tag{15}$$

Setting a threshold h for the cumulative sum Y_t, an alarm will be signaled when $Y_t > h$ occurs

4. Experiment and Results

4.1 Background traffic

We propose our method operating on link traffic volumes which can be easily collected by SNMP. But there is no data containing volumes of all links in a network open to public. The only data set we get from real network are 8 weeks traffic matrices of origin destination(OD) flows collected from Abilene network which is an educational network in USA. An OD flow represents traffic exchanges between two nodes in communication network.

Using a routing matrix A where A_{ij}=1 if flow *i* traverses link *j*, and is zero otherwise, the vector of traffic volumes on links *y* are the superposition of the vector of volumes on OD flows *x*.

$$y = Ax \tag{16}$$

Traffic of OD flows in the data set is aggregated in 5 minute bins. Since 5 minute bins would be a bit large for quick detection, we would like to extend the aggregated interval to a smaller scale. Based on the volume in a 5 minute interval, we could get the mean utilization λ. Then we assume the traffic in this 5 minute is a Poisson flow at intensity λ.By doing so, we can get estimations of volumes in varying length interval and collect traffic matrices in 1 minute bins.

According to the research of [5], traffic on links may have diurnal patterns and is non-stationary. It is not suitable to use only one model to characterize traffic all day long. So we just focus on the traffic from the busy hours at 16:00-22:00 on weekdays (not including Saturday and Sunday). 4 week-long data is used as the training set. And the other 4 week-long data is used as the testing set to generate background traffic under normal condition.

4.2 Synthetic Anomalies

Since we have no more detailed data to label the actual anomalies that may exist in our data set, we would like to use synthetic anomalies to test the detection rate(DR) of our method. The first step is to select a set of OD flows to be involved in the anomaly. Then anomalies are added on top of the original volumes of these OD flows. After that, the volumes on links carrying anomalies can be generated based on Equation (16).

The synthetic anomalies could be characterized by three parameters: the set of OD flows involved, volume and duration. In order to mimic the common DDoS attack, we would like to set the number of OD flows involved from 2 to 5. Total volume of anomalies would be percentage of the volume on the egress link connecting the victim. The percentage is set in the range of 20% to 100%. Then the volume of anomalies is divided evenly to the OD flows involved. The duration of DDoS attacks could last from minutes, to hours and even to days. According to [6], the majority of DDoS events in Abilene network lasted between 10 and 30 minutes, and a few outliers exceeded 2 hours. So we set the duration from 15 to 30 minutes.

For each of the 20 sequences of background traffic in the testing set, we generate 50 different anomalous scenes by varying the three parameters of anomalies. So there are totally 1000 different anomalous scenes.

4.3 Results

The common measure for performance of anomaly detection method is the Receiver Operating Characteristic (ROC) curve. ROC curve is a plot of DR against false positive rate (FPR) on different thresholds. DR is the percentage of the detection of anomalies. And FPR is the percentage of how many sliding windows under normal condition are classified as anomalous.

Figure 4 shows the overall ROC curve of our anomaly detection algorithm based on SHMM. For a FPR of about 5%, the DR is 70%. When the FPR reaches 8%, about 90% of the anomalies can be caught.

The reason why FNR is a little bit high is partly because we ignore the actual anomalies due to the limitation of the data we have. Actual anomalies in training set would affect the training of SHMM. Furthermore, background traffic carrying actual anomalies would trigger a false alarm. So we believe that our method would have a finer performance if we could separate the actual anomalies from the background traffic.

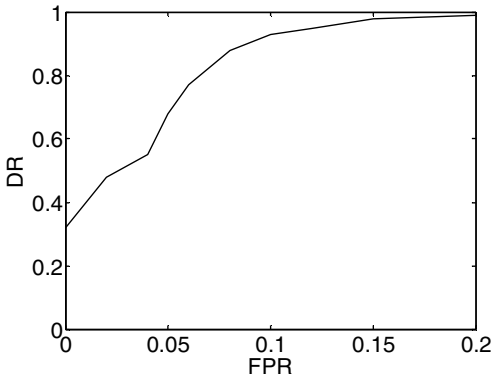

Figure 4. **The ROC curve**

We also test the sensitivity of volume which tells the impact of DR and FNR as the volume of anomalies get smaller and smaller. First we fix the detection threshold. Then we change the volume of anomalies from 20% to 100% to check the detection performance under different situations. Table 1 lists different FPRs and FNRs(1-DR) relative to different levels of anomaly volume. We can see that for anomaly volume under 20%, we can hardly catch the anomalies even at a very high FPR. The FNR and FPR begin to drop rapidly in the range of 40-60%. When the volume exceeds 80%, the majority of anomalies can be caught at the cost of low FPR.

Table 1. **The sensitivity of anomaly volume**

Volume level	FNR	FPR
20%	85%	62%
40%	38%	24%
60%	21%	18%
80%	8%	10%
100%	4%	5%

5.Related work

The first and the most important work to combine the network-wide analysis and traffic anomaly detection is [7]. The author uses Principal Component Analysis (PCA) as a dimensionality-reduction technique to project the original dataset to two disjoint subspaces. Based on the separation of traffic measurements into normal and anomalous subspaces, a G-statistic test on the anomalous subspace tells whether an anomaly occurs.

Another network-wide analysis work in [8] uses Kalman Filter to filter out the "normal" part of the original traffic. Then the residual is examined for anomalies by several statistical hypothesis testing.

These previous work both take traffic volumes on links as a multi-dimensional vector and use traditional multivariate tools to make a separation of the "normal" part and consider the residual responsible for anomalies. The spatial behavior is described as common patterns across links.

In this paper, our SHMM models the spatial behavior as interactions between routers. From the view of routers, we can combine the topology information to reveal the inside relationship of traffic volumes on different links. Rather than a traditional multivariate method, it is a strong compliment to recent work on network-wide anomaly detection.

6. Conclusion

In this paper we propose a novel model for network-wide traffic analysis which takes routers as the major elements responsible for the spatial behavior. Setting the structure of SHMM based on the topology, volumes on different links can be modeled as probabilistic outputs of interactions between different routers. Then the entropy, output of SHMM is tracked by a nonparametric CUSUM to keep eyes on the change of network condition. The experiment results show quite a good performance of our method. Further improvements will be made in order to eliminate impact of actual anomalies in the data set. Also there is a need to develop an online updating algorithm for the model parameters in order to catch the non-stationary properties of network traffic.

Acknowledgement

This work was supported in part by the Key Program of NSFC-Guangdong Joint Funds under Grant No.U0735002, the National High Technology Research and Development Program of China under Grant No.2007AA01Z449.

References

[1] P. Barford, J. Kline, D. Plonka, and A. Ron. A signal analysis of network traffic anomalies. In ACM Internet Measurement Workshop, Nov. 2002.

[2] J. D. Brutag. Aberrant behavior detection and control in time series for network monitoring. In 14th Systems Administration Conference (LISA 2000), Dec. 2000.

[3] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. Sketch based change detection: Methods, evaluation, and applications. In ACM Internet Measurement Conference, Oct. 2003.

[4] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, vol. 77, pp. 257-286, 1989.

[5] A. Lakhina, K. Papagiannaki, C. D. Mark Crovella, E. D. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. In ACM SIGMETRICS, 2004.

[6] A. Lakhina, M. Crovella, and C. Diot. Characterization of network-wide anomalies in traffic flows. In ACM Internet Measurement Conference, Oct. 2004.

[7] A. Lakhina, M. Crovella, and C. Diot. Diagnosing networkwide traffic anomalies. In ACM SIGCOMM, Aug. 2004.

[8] A. Soule, K. Salamatian, and N. Taft. Combining filtering and statistical methods for anomaly detection. In ACM Internet Measurement Conference, Oct. 2005.

[9] P Qiu and D Hawkins. A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. Journal of the Royal Statistical Society: Series D (The Statistician) vol.52 (4) , pp.706–712, 2003.