# Research on the detection of abnormal traffic for multi-channel network

**Lixia Liu[1, 2], Hong Mei[1], Bing Xie[1]**

*[1] School of Electronics Engineering and Computer Science, Peking University, China*

*[2] Dept. of Information Engineering, Engineering University of CAPF, China*

*Corresponding author's e-mail: wjllx939@sohu.com*

*Received 25 October 2013, www.cmnt.lv*

## Abstract

With the rapid growth of the categories and numbers of network attacks and the increasing network bandwidth, network traffic anomaly detection systems confront with both higher false positive rate and false negative rate. A traffic anomaly detection system with high precision is presented in this paper. First, we use multi-level and multi-dimensional online OLAP method to analyze traffic data. In order to reduce the computational and space complexity in this analytical process, some optimization strategies are applied in building DetectCube, the minimal directed Steiner tree algorithm is adapted to optimize multiple query on the Cube, and the traffic data is summarized at appropriate level with the help of discovery-driven exploration method. Second, a concept of entropy to measure the distribution of traffic on some particular dimensions is given and the values of entropy in every window and every Group-By operation are collected to form multiple time series of entropy. Finally, we employ one-class support vector machine to classify this multi-dimensional time series of entropy to achieve the purpose of anomaly detection. The proposed traffic anomaly detection system is validated and evaluated by comparing it with existed systems derived from a lot of real network traffic data sets. Our system can detect attacks with high accuracy and efficiency.

*Keywords:* traffic anomaly detection; entropy; DetectCube; OLAP; one-class support vector machine.

## 1 Introduction

With the increasing of network scale and the appearance of various network applications, the Internet has become an indispensable living infrastructure to Humanity. However, network attacks and network failures are serious threats to network security. The traditional IDS tools based on feature matching (e.g., Snort and Bro) have been unable to meet the demand of high-speed network testing in performance, and what's more, they are incapable to detect new attacks effectively. As a kind of active safety technology, the network traffic anomaly detection can quickly detect various kinds of network anomalies in different network flow layers. Network traffic anomaly detection is to find anomalous behaviors that may occur in the network by analysing the normal network traffic. The key issue is to get the comprehensive traffic information and describe it with appropriate characteristics. In order to improve the accuracy of anomaly detection system and efficiency, the researchers have proposed many methods for traffic anomaly detection [1], e.g., the methods based on threshold, prediction, wavelet, data mining, machine learning, and statistics. However, network traffic has increased dramatically, and all kinds of new network applications user scale continue to expand, and hence increase complexity and heterogeneity of network. Besides, anomaly detection systems itself exist basic false probability. As the offensive and defensive game continued, all kinds of means of attack and attack methods evolve to be imperceptible. Owning to the single detection measure, most anomaly detection systems are only effective for some types of attack detection and are easy for the attacker to bypass. Therefore, it urgently needs to study a new detection method, which is satisfactory in the detection accuracy, efficiency and can detect new attacks.

Traffic anomaly detection method based on statistical model is the most widely used in the field of anomaly detection. This type of methods usually samples the network information according to certain interval, then calculates each sampling data and obtains a series of statistic information to describe network status, and finally detects abnormities by the change of the test statistic information. Different statistical model can produce different detection methods and each type of detection method has a corresponding measure. Researchers put forward a lot of testing measures for different types of exceptions, such as large-scale TCP connection exception, SYN packet with the SYN + ACK packet number do not matching, abnormal RST packet number, and abnormal FIN packet number. When DDOS attacks occur, the ratio of data packets in and out is abnormal. These measures can effectively solve the problem of some types of attack detection, but poor for other types of attack detection. Because the attack theory for each attack is different, network traffic anomaly is also different. Statistics data is usually gathered for a single level of a single dimension, so it can only describe a single view of network flow. The gathering for the single dimension or the multi-dimensional may loss useful information on other dimensions. For example, when gathering on the protocol type dimension, we find the ICMP packet data increase. However, it actually is network debug in progress with some hosts continuously for ping. These problems can be solved in multiple dimensions with statistics (e.g., we get the statistic data for each flow). However, this method needs to store the information for each flow. Suppose that each flow is compose of the IP address of source/destination, the port of source/destination and the sign for protocol type, then we must assign at least 104b space for storing the information of each flow, and the number of flow can be $2^{104}$. In order to solve this problem, OLAP is employed in this paper and we use the entropy to measure distribution differences for anomaly information.

## 2 Related works

Collecting network traffic data by capture tool On the target network (such as NetFlow, catching works based on systems of LibPcap, etc.), select a set of dimensions and define hierarchy through statistics, online real-time statistics on the levels of these dimensions and definitions are the key point of detection system. This section elaborates on specific meaning of network traffic anomaly detection system and finally introduces the entropy measure that is used to detect in this article.

### 2.2 MULTI-DIMENSIONAL CHARACTERISTICS OF FLOW DATA

Network traffic data having a plurality of different attributes, KDD CUP 99 dataset gives 41 properties which is divided into three categories in accordance with the importance of the program can be used for anomaly detection, focusing more on active / destination IP, source / destination port, protocol type, etc., each attribute is defined intuitively as a dimension of detection. In order to reduce the complexity of statistical calculation it can be divided into different levels of statistics on these dimensions and may be gathered using the histogram method, can also be directly aggregation.

Network administrators usually need to select several dimensions in the process of traffic analysis and processing, and select a level of abstraction to deal with the flow of data classification on the various dimensions to facilitate comparison of horizontal and vertical. traffic analysis need to limit dimensions and levels of analysis of the problem, at the same time dimension also need to select a specific level (type and span of the time window) analysis window as the studied time data streams.

### 2.3 ENTROPY AND RELATIVE ENTROPY MEASURE

Entropy originated from information theory, it is a good measure for the change of the distribution system, functioning well in describing the long random process. The main idea of entropy-based anomaly detection system is: if there is abnormal flow occurs, the flow characteristics of the overall distribution should be changed, using the entropy to measure the distribution of characteristics, the anomaly can be detected by changing of the entropy. Entropy has been widely used in network traffic anomaly detection [7-9], comparison on the accuracy of the current classical anomaly detection systems are introduced in [10], anomaly detection system based on entropy achieved good results detection that is superior to other abnormalities detection method. Entropy used as a measure of anomaly detection in this paper, in order to dynamically adjust the aggregate level, using the relative entropy as a measure of abnormal situation may currently appear and using the relative entropy to drive the aggregation level and drill on the volume.

Set random variables X (such as the value of IP dimension), [n] is value space, set $m_i$ is the time of occurrence of data item, the overall data items in data stream is $m_i$ and $m=\sum_{i=1}^{n} m_i$ is the number of disparate data items data appeared, the entropy is defined as

$$H=-\sum_{i=1}^{n}\left(\frac{m_i}{m}\right)\mathrm{lb}\left(\frac{m_i}{m}\right). \tag{1}$$

Set up two distributions P and Q, they have the same value space, then define the relative entropy

$$I\left(PQ\right)=\sum p_i \mathrm{lb}\left(p_i/q_i\right). \tag{2}$$

In this paper, the distribution of network traffic data is calculated separately in different time window. Then use the relative entropy measure differences in the distribution within each time window.

## 3 System architecture and implementation

### 3.1 SYSTEM ARCHITECTURE

Anomaly detection system architecture shown in Figure 1, the entire system is composed of four parts: Part 1 is part of query optimization, network security experts to define MDX query pointed at the demand for network anomaly detection and the current state of the network, MDX grammar parser analyze queries forms a plurality of Group-By gathering operations, using multi-aggregate query optimization algorithm for the optimization scheduling of multi-query of Group-By gathering operations. Part 2 is the preprocessing of data. This section is based on packet header information and resolve internal start-byte data filtering. In the process of achieving it, blacklists and white lists are the comprehensively used.

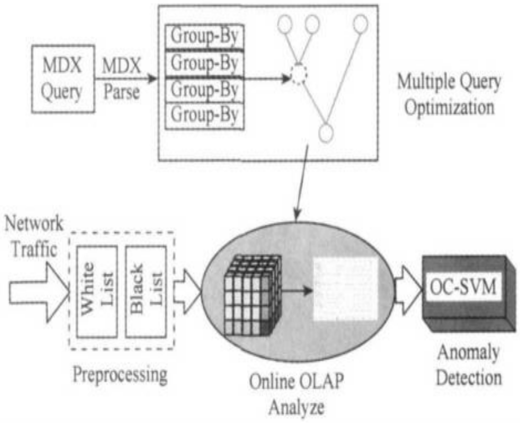


FIGURE 1 Architecture of the proposed system

### 3.2 DETECTCUBE DATA MODEL

Data cube is one of data analysis model in OLAP (On-Line Analytical Processing), supporting a common multi-dimensional multi-level. Through the data model characteristics of data sets can be studied from different perspectives inspection. Among Network traffic data, distinct traffic data property constitutes a different dimension, you can define he different levels of abstraction on different dimensions. For analysis of each dimension and level of the data can reflects t the current operating status of the network from different angles. In this paper, the data cube model can be extended to detect operational state of the network from many different angles. it alarms when running state deviates from the normal state far, the following definitions are given about DetectCube data model.

**Definition 1** Give a data set S, a set of dimensions $A=\{A_1,A_2,\ldots A_n|n\rangle 0\}$, also give the level set of each dimension $H=\{H_1,H_2,\ldots H_n|n\rangle 0\}$.

Taking a hierarchical level of each component of $\{H_1,H_2,\ldots H_n\}$, make combinations. A group of elements constituting a cubic data cube is created through performing packet aggregation operations on data sets S on All such combinations like $\{H_1^{j1},H_2^{j2},\ldots,H_n^{jn}\}$.

**Definition 2** a cubic unit is a two-dimensional array $(A_H,M)$ where $A_H=\{H_1^{j1},H_2^{j2},\ldots,H_n^{jn}\}$ and $H_i^{ji}\in H_i$ is a specific level of abstraction selected in each dimension. When $H_i^{ji}$ can be *, representing the highest level of abstraction of the dimension, when $H_i^{ji}$ = * the dimension cube does not work on choosing unit. M is the numerical attributes value obtained after polymerization operation on the selected data sets, in this paper M is the number of IP packets.

### 3.1 MULTIDIMENSIONAL SEQUENCE ENTROPY ANOMALY DETECTION

The DetectCube each Group-By operation corresponds to a distribution, representing the distribution of traffic data gathered on these dimensions, degree of uniformity of these distributions can be measured by entropy, according to the characteristics of entropy when the distribution of the aggregated individual deviations in normal circumstances far, the network can be considered that there is an exception occurs. In order to detect anomalies in the entropy of the sequence, this paper arranged the entropy of each Analysis of view in detection vector using OCSVM detection vector classification.

Constructing hyper plane in H or a linear judgmental function f (x):

$$f(x)=w^T.\phi(x)-\rho\,. \tag{3}$$

Hyper plane should be able to separate the vector of training samples and vector origin, where w is the hyper plane normalization vector, ρ is the hyper plane offset. To solve w and ρ, we need to address the following quadratic programming.

$$\min\phi(w,\xi,\rho)=\frac{1}{2}w^Tw+\frac{1}{v\cdot l}\sum_{i=1}^{l}\xi_i-\rho$$

$$\text{s.t.}w^T\cdot\phi(x)\geq-\rho-\xi_i,\xi_i\geq 0\,. \tag{4}$$

$\xi_i$ is a slack variable, $v\in(0,1)$ its value needed to be compromised. For the Quadratic programming problems above, you can use Lagrange multipliers to solve:

$$w=\sum_{i=1}^{l}\alpha_i\phi(x_i),0\leq\alpha_i\leq\frac{1}{v\cdot l}\,. \tag{5}$$

Then determining the function f(x) becomes a nonlinear function

$$f(x)=\sum_{i=1}^{l}\alpha_iK(x_i,x)-\rho\,. \tag{6}$$

$K(x_i,x)$ is the kernel function.

## 4 Experimental results and analysis

### 4.1 EXPERIMENTAL DATA

During the experiment mainly uses two types of data: one is synthetic data; one is the real backbone traffic data. Synthetic data injected all kinds of malicious network traffic data injected into the background traffic according to certain rules, where the background traffic using the trace data capture d by WIDE project group when they across t the backbone network traffic in the Pacific Ocean. What is used in the experiment is flow from 2:00 to 4:00 on 2009-03-30 (this period is relatively stable flow), malicious traffic data using CAIDA organization released data Witty worm, DDOS 2007 data as well as data worm.

20% data of Witty is injected into background data traffic, 2% of DDOS 2007 data was poured into background traffic, 50% of Conficker worm data traffic is injected into the background as the synthetic experimental data set, the specific flow rate curve shown in Figure 3, which were plotted normal traffic, malicious traffic and synthetic traffic in each graph where the abscissa is the time, according to statistics of time window in 1 minutes, the ordinate is the number of IP packets.

Part 2 of the experimental data using the backbone network traffic data collected by WIDE project groups in 2010-04-15 2008-03-19,2009-03-30, the data is collected in A Day under the direction of project called In The Life Of The Internet organized by CAIDA initiated, with a typical representative.

First, located centralized flow anomalies in data , the flow as shown in Figure 4, shown by the vertical line identifies the traffic anomaly section 4, where there are four abnormal segment 2008-03-19, 2009-03-30 respectively 3 abnormal segments, 2010-04-15 no obvious abnormalities.

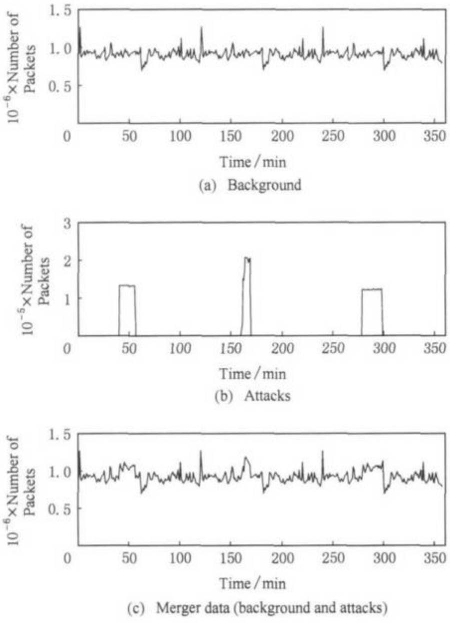
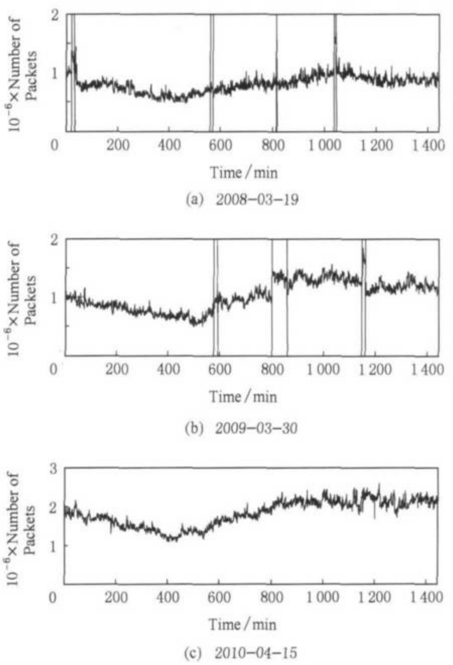


FIGURE 2 Time series of synthetic traffic traces



(a) 2008–03–19

(b) 2009–03–30

(c) 2010–04–15

FIGURE 3 Time series of synthetic traffic traces

### 4.2 THE DESIGNING FOR THE EXPERIMENTS

In order to verify the exactness of the algorithm, the basic flow of EWMA test method for the prediction, which also has higher detection accuracy, is adopted to do the comparison experiment with detection method based on the Sketch and the detection method based on relative entropy value. In the first part of the experiments, the above three methods and the method proposed in this research are respectively used to do experiments on the two data sets. And the detection time window is set to 1 min. In order to verify the influence of various parameters in the experiments, the second part experiments are designed to mainly test the influence to the time window and the data itself characteristics of the proposed method detection accuracy. When the accuracy caused by the time window is considered, the data selected by the Witty worms attack is set as the experimental data. And the detecting window is respectively set to 1 min, 2 min and 5 min. In order to study the influence of the proportion of attack traffic on the detection accuracy, the DDOS attack traffic of 0.5%, 1% and 2% are respectively injected into the background traffic, which can constitute three traffic data collection. The method proposed by this research is based on the three data sets.

### 4.3 THE ANALYSIS OF THE EXPERIMENT RESULTS

The characteristics of network traffic data are fully achieved by adopting multi-dimensional multi-level detection technique proposed by the research. By combining more detection measures than the other methods, the abnormity detecting can be more effective than the others. And the detection error is made to the minimization based on a kind of support vector machine (SVM) on multidimensional sequence entropy. So the method proposed in this research can achieve the higher detection accuracy.

## 5 Conclusion

The irrelevant traffic is removed by adopting the black and white list technology in the method on the front end. Since the experiment data which has been processed anonymously without the black and white list technology, the need for dealing with the real-time traffic can be greatly reduced. And the detection precision of the system is improved to a certain extent. But malicious traffic on the front end will be filtered out with omission records. This problem can be solved by carefully choosing the black and white list of filters. And the method proposed is mainly based on the analysis of the view on the choice, but do not give specific guidance. It greatly relies on the user's expert knowledge. The selection of the analysis view is the important work in the next step research. And, at the same time, the system cannot provide details about the exception type information. When an exception occurs, traffic alarms are only given out in the network. The linkage and other network security equipment will be required, for more detailed location and blocking abnormal.

The time and space complexity of our proposed method is very higher, especially the OLAP analysis in the data flow process. The construction process of DetectCube will increase with the network traffic data. Accordingly the storage overhead and computing time will increase with the traffic data, too. The characteristic of DetectCube data processing will be further excavated in the next step in order to reduce the computational overhead. In the part of the abnormal drive set, when the level of the gathered drill down to a finer level, if the time relative entropy change trend is the same as that before the drill, the possibility of abnormal can be recognized increasing. This may be the content of further study in this part.

## References

[1] Patcha A, Park J 2007 An overview of anomaly detection techniques: Existing solutions and latest technological trends *Computer Networks* **51**(12) 3448-70

[42] Axelsson S 2000 The bass rate fallacy and the difficulty of intrusion detection *ACM Trans on Information and System Security* **3**(3) 186-205

[43] Cheng Jiren, Yin Jianping, Liu Yun, et al 2009 Detecting distributed denial of service attack based on address correlation value *Computer Research and Development* **46**(9) 1334-40

[44] Sarawagi S, Agrawal R, Gupta A 2006 *On computing the data cube, RJ10026* San Jose, CA: IBM Almaden Research Center 1-18

[45] Charikar M, Chekuri C, Cheung T 1999 Approximation algorithms for directed steiner problems *Journal of Algorisms* **33**(1) 73-91

[46] Scholkopf B, Platt J C, Shawe-Taylor J 2001 Estimating the support of a high-dimensional distribution *Neural Computation* **13**(7) 1443-71

[47] Gong Jian, Peng Yanbing, Yang Wang, et al 2006 Reconstructing the parameter for massive abnormal TCP connections with bloom filter *Journal of Software* **17**(3) 434-44

[48] Gu Yu, Mccallum A, Towsley D 2005 Detection anomalies in network traffic using maximum entropy estimation *ACM SIGCOMM Conf on Int Measurement (IMC). New York: ACM* 345-50

[49] Krishnamurthy B, Sen S, Zhang Yin, et al 2003 Sketch-based change detection: Methods, evaluation, and applications *ACM SIGCOMM Conf on Int Measurement (IMC). New York: ACM* 234-47

[50] Measurement and Analysis on the WIDE Internet (MAWI) Working Group Traffic Archive. http://mawi.wide.ad.jp/mawi/

## Authors

| | |
|---|---|
| | **Lixia LIU, born in 1975, Shanxi, China**<br>**Current position, grades:** Ph.D., associate professor, master advisor of the Engineering University of CAPF<br>**University studies:** Bachelor in Computer Science from Engineering University of CAPF in 1998, Master's degree in Computer Science from Northwestern Polytechnical University in 2003, Doctorate degree in Electronics Science and Technology from Xidian University in 2011<br>**Scientific interest:** PSO,SVM, System Software<br>**Publications:** 59 |
| | **Hong MEI, born in 1963, Guizhou, China**<br>**Current position, grades:** Ph.D., professor, vice president of SJTU<br>**University studies:** Bachelor and Master's degrees in Computer Science from Nanjing University of Aeronautics & Astronautics (NUAA) in 1984 and 1987 respectively, Doctorate degree in Computer Science from Shanghai Jiao Tong University in 1992<br>**Scientific interest:** Software Engineering,System Software<br>**Publications:** 186 |
| | **Bing XIE, China**<br>**Current position, grades:** Ph.D., professor, vice dean of School of Electronics Engineering and Computer Science at Peking University<br>**University studies:** Bachelor degree in Computer Science from PLA Information Engineering University in 1984 and 1987 respectively, Master's and Doctorate degrees in Computer Science from National University of Defense Technology in 1995 and 1998 respectively<br>**Scientific interest:** Software Engineering<br>**Publications:** 150 |