Using Graph to Detect Network Traffic Anomaly

Yingjie Zhou Guangmin Hu Weisong He

Abstract— **Comprehensive collection and accurate description of traffic information are core problems in network traffic anomaly detection. Aiming at the lack of traffic anomaly detection in analyzing multi-time series, we propose a network traffic anomaly detection method based on graph mining. Our method accurately and completely describes the relationships among multi-time series which are used in traffic anomaly detection by time-series graph; by means of the support count of the patterns, our method mines all the frequent patterns ,which is conducive to detecting many kinds of abnormal traffic effectively; through mining the relationships among all itemsets, our method introduces weight coefficients of the itemsets, which is able to solve relationship quantification issues of multi-time series in traffic anomaly detection. The simulation results show that the proposed method can effectively detect the network traffic anomaly and achieve a higher accuracy than the CWT-based (Continuous Wavelet Transform-based) method in term of DDos attacks detection.**

I. INTRODUCTION

Network traffic anomaly refers to the status that the traffic behaviors deviated from its normal behaviors. The feature of network traffic anomaly is that it erupts suddenly without any omen. It can bring great damage to networks and network equipments in a short time. People usually discover the abnormal behaviors that may occur in the network or system through the description and analysis of the network traffic, and send alerts to the administrator. This process is defined as network traffic anomaly detection. Comprehensive collection and accurate description of traffic information are core problems in network traffic anomaly detection. The abnormal traffic has brought great harm to the network, and there are more and more network traffic anomalies along with the rapid popularity of network applications. Therefore, to detect anomaly rapidly and accurately and make reasonable response has become the attractive and valuable subject in the present academic and industrial circles.

Existing traffic anomaly detection methods usually treat time-varying traffic information as a one-dimensional signal (or a one-dimensional time series), and detect traffic anomaly through a variety of signal analysis methods. Hussain et al. put forward the method that classified DDoS attacks through signal spectrum analysis [1]. Cheng et al. suggested identifying DDoS attacks by analyzing the cycle characteristics of TCP traffic based on the energy spectrum density of the traffic signals [2]. Alarcon-Aquino and Barria proposed a algorithm based on UDWT (Undecimated Discrete Wavelet Transform) and Bayesian analysis, using the wavelet coefficients at all levels to detect and position the weak change of a given time series in the variance and frequency [3]. P. Barford et al. indicated using wavelet filters to find subtle abnormalities flow, and detecting anomaly by scanning the rapid changes of the local filter data [4]. Jun Gao et al. proposed a scale-adaptive network traffic anomaly detection method based on wavelet packet, which had the same detective ability for high, medium and low-frequency anomaly traffic [5].

Since the complexity of abnormal traffic detection, the false negative rate and false positive rate of one-dimensional time series analysis methods are usually quite high. Lee et al. suggested detecting anomaly using information entropy, combined with information gain, information cost [6]. Lakhina et al. have proposed that using time series of the distributions of packet features (IP addresses and ports) to describe the network anomalies, and detecting and identifying anomalies by using entropy as a summarization tool [7]. Guan Xiaohong et al. divided network traffic into normal space and abnormal space, which are two independent components. By this way, it can detect traffic anomaly through the analysis of abnormal space [8].

Multi-time series analysis improves the accuracy of traffic anomaly detection, which can effectively reduce the false negative rate and false positive rate. Many methods, which are also used extensively in the detecting process, admit the interrelationships of multi-time series. However, the descriptions of multi-time series in the existing methods are independent. There isn't an effective approach to describe the relationships among multi-time series accurately and comprehensively, and apply it to the detection of abnormal traffic. Therefore, the information that traffic anomaly detection based on has limitations in the integrity, accuracy and etc.; and these limitations reduced the detection precision. In response to these issues, this paper suggests using a graph to describe the multi-time series and their relationships at each time; and graphs of different time constitute a time-series graph.

Time-series graph can describe the relationships among multi-time series which are used in traffic anomaly detection more accurately and completely. It is a new idea in traffic anomaly detection, which is able to use abundant information that existing multi-time series analysis methods are difficult to make full use of. To this end, this paper proposes a network traffic anomaly detection method based on graph mining. By means of the support count of the patterns, our

This work was supported by the National Science Foundation (60572092), the National High Technology Research and Development Program of China (2008AA011001) and the Program for New Century Excellent Talents in University (NCET-07-0148).

Yingjie Zhou, Guangmin Hu, Weisong He are with Key Laboratory of Broadband Optical Fiber Transmission and Communication Networks, UESTC, Chengdu, 610054, China yjzhou@uestc.edu.cn hgm@uestc.edu.cn weisonghe@uestc.edu.cn

978-1-4244-4888-3/09/$25.00 ©2009 IEEE

method mines all the frequent patterns, which is conducive to detecting many kinds of abnormal traffic effectively; through mining the relationships among all itemsets, our method introduces weight coefficients of the itemsets, which is able to solve relationship quantification issues of multi-time series in traffic anomaly detection. The simulation results show that the proposed method can effectively detect the network traffic anomaly and achieve a higher accuracy than the CWT-based method in term of DDos attacks detection.

II. The Construction of Time-Series Graph

Since each NetFlow message contains tens of thousands of or even hundreds of thousands of lines of flow information, it is difficult to deal with such flow information directly. Regarding each attribute of NetFlow data as a sequence of random events, the concept of comentropy can be effectively used to measure the concentration and dispersion situation of data, which corresponds to each attribute in order to obtain the rough-granularity expression of mass data. We select four entropy sequences, which are source IP address, destination IP address, source port and destination port from the data of NeFlow protocol. We also represent their values and the relationships between them through constructing time-series graph.

A. NetFlow Data and Entropy

NetFlow technology is a prevalent IP / MPLS traffic analysis and measurement standard in the Internet. NetFlow message is constituted of two parts: message header and a number of flow information. The flow information contains source IP address, destination IP address, source port, destination port, the number of packets, and some other attributes.

Entropy measures the uncertainty of random events. We can use entropy to analyze [9] if we regard NetFlow data as a discrete information source and each of the attributes as a sequence of random events.

The entropy of NetFlow data's attribute S:

$$H(S) = -\sum_{i=1}^{n} P_i \ln P_i \tag{1}$$

where P_i is the attribute's frequency of a certain value, n is the total number of instances, and $\sum_{i=1}^{n} P_i = 1$. The entropy can effectively represent the concentration and dispersion situation of corresponding data on the same attribute. Where the data is more concentrated, the smaller the entropy is; where the data is more decentralized, the greater the entropy is, especially in the large-scale network traffic flow.

In this paper, we focus on four attributes: source IP address, destination IP address, source port, destination port. They can effectively represent the abnormalities of large-scale network traffic flow.

B. Graph Representation of Multi-Time Entropy Series

As shown in Figure 1, four nodes represent the the four attributes at the time sampling point respectively and edges denote the relationships between nodes.

Details are as follows:

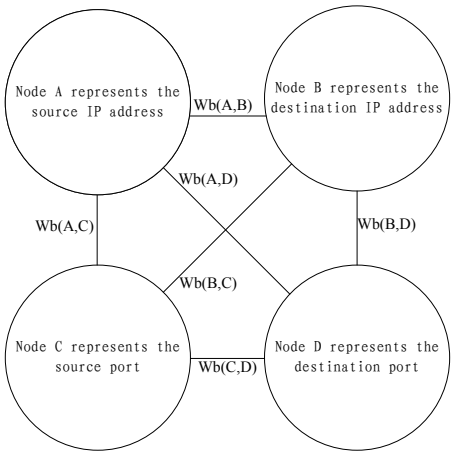


Fig. 1. Graph representation of multi-time entropy series, where $Wb(p,q)$ ($p,q = A,B,C,D$ and $p \neq q$) denote respectively the edge weights of corresponding edges

1) Node representation

We normalized the entropies by mapping their values to $[0,1]$, and then divided into four values levels according to the mapped values. They were divided by the following rules: level 1 for $[0, 0.25]$, level 2 for $[0.25, 0.5]$, level 3 for $[0.5, 0.75]$, level 4 for $[0.75, 1]$. A, B, C, D denote the four attributes, while A_j, B_j, C_j, D_j (j denotes the value level of the corresponding node's entropy, $j = 1, 2, 3, 4$)are the entropies modes of them at the time sampling point.

$pq(p, q = A, B, C, D$ and $p \neq q)$ represent a 2-itemsets, and $p_i q_j (p, q = A, B, C, D$ and $p \neq q$ and $i, j = 1, 2, 3, 4)$ denote a 2-itemsets model. $p, q, r (p, q, r = A, B, C, D$ and $p \neq q$ and $p \neq r$ and $q \neq r)$ represent a 3-itemsets, and $p_i q_j r_k (p, q, r = A, B, C, D$ and $p \neq q$ and $p \neq r$ and $q \neq r$ and $i, j, r = 1, 2, 3, 4)$ denote a 3-itemsets model.

Anomalies in network traffic usually have a strange pattern of itemsets in the graph.

2) Edge representation

Let $H^k(A), H^k(B), H^k(C), H^k(D)$ denote the entropies of source IP address, destination IP address, source port and destination port at the time sampling point k.

Denote edge weight which is decided by the two connected endpoints as:

$$Wb^k(p,q) = \frac{H^k(p) - H^{k-1}(p)}{H^k(q) - H^{k-1}(q)}, \tag{2}$$

where$p, q = A, B, C, D$ and $p \neq q$.

Edge weight reflects the similarity degree about the entropy value change that the two endpoints of a certain edge have at the time sampling point. The two endpoints of the connected edge belong to two different time sequences. Therefore, edge weight changes in the entire time-series reflect the extent of the relationship between two time sequences: the smaller the change is, the closer the relationship is; the greater the change is, the looser the relationship is.

In this way, we formed an undirected weighted graph at each time sampling point. From the entire time-series, we obtained a time-series graph.

III. Mining Time-Series Graph and Traffic Anomaly Detection

Graph-based anomaly detection is an emerging area in anomaly detection. Noble and Cook introduced two techniques used in graph-based data to find unexpected patterns [10]. But both of them analyzed without using the relationships among graph elements. Analyzing with the relationships among graph elements can bring extra information, which is conducive to improve the accuracy of traffic anomaly detection. Therefore, this paper first analyzes on graph elements and the relationships among them by graph mining, which can get the supports of patterns [11] and weight coefficients of itemsets, and then determines the abnormal rule with the obtained results.

A. Mining Time-Series Graph

Data mining in time-series graph includes: the support count of patterns and the relationship mining among all itemsets.

1) The Support Count of Patterns: Support reflects the frequent degree of a pattern and shows the frequency of the pattern. The smaller the support is, the greater the likelihood that network traffic anomaly may occur. For each graph at a certain time sampling point, the frequent degree in which the patterns of its six 2-itemsets and four 3-itemsets are can reflect the possibility of abnormal flow at the time sampling point. The more the patterns which are frequent patterns at the time sampling point are, the less the possibility that abnormal flow may appear.

The support of the pattern is defined as the ratio that the number of the pattern divided by the total number of instances in its itemsets. It provides a basis for judging the network traffic abnormal. Using support count method to mine 2-itemsets model and 3-itemsets model in the time-series graph, we could obtain the support of each 2-itemsets model $Sup_t(p,q)$ and the support of each 3-itemsets model $Sup_t(p,q,r)$ (p,q,r express different nodes in the graph respectively, t denotes the time sampling point).

2) Mining Edge Weight: In order to solve relationship quantification issues of multi-time series in traffic anomaly detection, we mined the edge weights, which introduced weight coefficients of the itemsets (2-itemsets or 3-itemsets in this article) to describe.

Weight coefficient of the pattern set represents the relation degree of elements in the pattern set. It not only defines the inner connection degree of the pattern set, but also identifies the likely connection intensity of the patterns in the pattern set. The more intense the relation degree of elements in the pattern set is, the stronger the impact on the support of the pattern is, and the greater the weight coefficient of the pattern set is.

We could get weight coefficients of the itemsets by mining the edge weights. The second-order central moment of the edge weight $Wb^k(p,q)$ ($p,q = A,B,C,D$ and $p \neq q$) in the entire time-series is defined as $S(p,q) = \frac{1}{n}\sum_{k=1}^{n}(Wb^k(p,q) - \mathrm{E}(Wb^k(p,q)))^2$. The smaller the second-order central moment of a certain edge is, the steadier the relation between the entropy values of the two endpoints connected by the edge is, which stands for the relation is stronger and have a greater impact on the support of the 2-itemsets.

Along with the increasing of the second-order central moments, the impact on weight coefficients of the itemsets due to its continual increasing will decrease. Thus, we can approximately use a monotone increasing exponential function to describe the relationship between the second-order central moments and weight coefficients of the itemsets.

Weight coefficient of a 2-itemsets: $W^2_{p,q} = 10^{-S(p,q)}$ (p,q express different nodes in the graph respectively). We use exponential function to make sure that weight coefficients of the itemsets value between $0-1$. When the second-order central moment of a certain edge is 0, the entropy values of the two endpoints connected by the edge are positive correlated, and weight coefficient of the 2-itemsets is 1; When the second-order central moment of a certain edge is $+\infty$, the entropy values of the two endpoints connected by the edge are unrelated, and weight coefficient of the 2-itemsets is 0.

The relationships among three nodes could be expressed by the addition of the relationship of two nodes. Weight coefficient of a 3-itemsets: $W^3_{p,q,r} = 10^{-(S(p,q)+S(p,r)+S(q,r))}$ (p,q,r express different nodes in the graph respectively).

B. Rule for Abnormity Determining

The frequent degree or the support of the pattern at a certain time sampling point can reflect the possibility that abnormal flow may appear at the time sampling point. Weight coefficients of the itemsets have solved the relationship quantification issues of multi-time series in traffic anomaly detection. It quantified the contribution that the support of patterns in where the itemsets are could make to the abnormality at the time sampling point.

From the above ideas and mining results of time-series graphs, we define the abnormity coefficient W_t to measure the degree of network traffic anomaly at a certain time sampling point on a single router. It is important to realize that the frequent 2-itemsets model is more valuable than the frequent 3-itemsets model in network traffic anomaly detecting, as it reflects the most direct and basic relationship of multi-time series. Therefore, we multiplied the abnormity coefficient of 3-itemsets by 0.6.

$$W_t = -\min_{0<l<N}\{W_t\} - \log_{10}\Big(\sum_{1\le p,q\le 4, p\neq q} W^2_{p,q} \bullet Sup_t(p,q) + 0.6 \times \sum_{1\le p,q,r\le 4, p\neq q, p\neq r, q\neq r} W^3_{p,q,r} \bullet Sup_t(p,q)\Big), \tag{3}$$

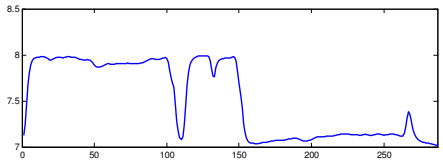
where p,q,r express different nodes in the graph respectively, t denotes the time sampling point, and N is the total number of the time sampling points.

The greater the abnormity coefficient is, the more possibility that abnormal flow may appear. If the the abnormity

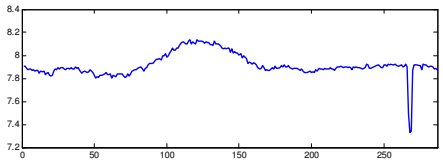
coefficient is more than twice the abnormity coefficient of nearby, there is an abnormity at the time sampling point.

C. Simulation and Analysis

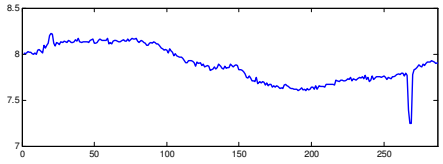
In this paper, the simulation uses the sampling data from Abilene [12]. We have collected a whole day's flow data on a single node, which were from Abilene's IP-level sampling flow data (packets sampling 1/100, cycle sampling at intervals of five minutes) in December 13, 2006. Every five minutes' flow exports constituted the data of a time sampling point, and a day contained 288 time sampling points.



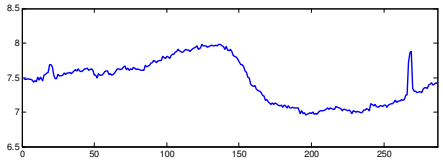
(a) Source IP address entropy sequence



(b) Destination IP address entropy sequence



(c) Source port entropy sequence



(d) Destination port entropy sequence

Fig. 2. Sequence diagrams of entropy sequences before inserting attack

Through conversion and calculation with data from Abilene backbone network, we could obtain four information entropy sequences as shown in Figure 2.

Figure 3 drew the sequence diagram of the abnormity coefficient on a single node on that day. There are four peaks in all, corresponding to 4 detected anomalies. Through manual analysis of original IP packet data, we can find that flows in the four time sampling points are unusual; where peak 1, 2, 3 are point-to-multipoint anomalies, and peak 4 is DDoS attack.

In order to further validate the effectiveness of the method, we inserted attack flow into flow data of IP-level sampling. This would make it easier to know precisely when "attack" occurred. We simulated the process that 200 agents attacked

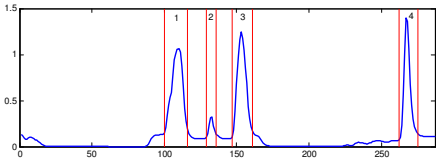
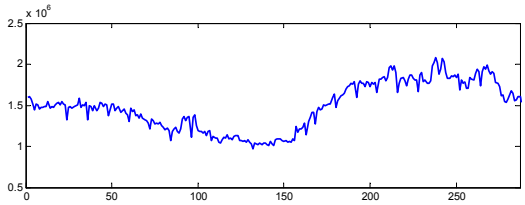
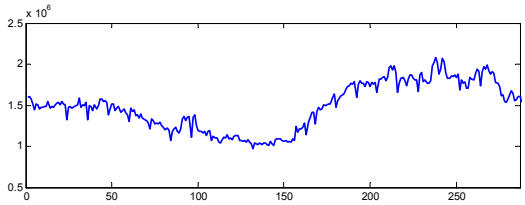


Fig. 3. Sequence diagram of abnormity coefficient before inserting attack



(a) Flow before inserting attack



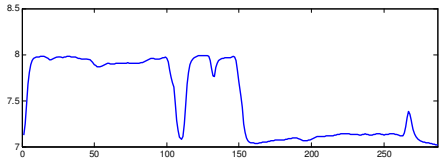
(b) Flow after inserting attack

Fig. 4. Flow diagrams of packet number before and after inserting attack

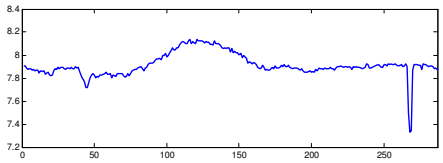
a host, and collected packet data on the backbone router. 10 time sampling points' attack data formed an attack flow. We injected the attack flow into 10 time sampling points from 40 to 49. Figure 4 drew flow diagrams of packet number before and after inserting attack. Figure 5 drew sequence diagrams of entropy sequences after inserting attack. Figure 6 showed sequence diagram of the abnormity coefficient on a single node after inserting attack.

If we watched directly from the traffic signal after inserting attack, we would see the addition intensity was not prominent, especially compared with the traffic at nearby time sampling points, although the addition attack data somewhat strengthened the intensity of the traffic between the time sampling point 40 and 49. If we watched from the four entropy sequences after inserting attack, we would notice that only the value of the destination IP address entropy sequence had decreased in a small range at the time sampling points between 40 and 49.Through the analysis of our method, we finally got the sequence diagram of the abnormity coefficient as shown in Figure 6, which could easily determine the anomaly and make sure its occurrence time (the position that label a showed in Figure 6). From Figure 6 we found that there was an additional exception at the time sampling points between 40 and 49, which was actually the abnormity that artificially injected into the traffic.

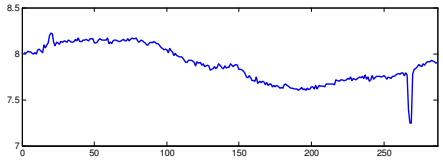
In order to compare our method with Alberto Dainotti's wavelet-based method, we experimentally simulated the same data by using the wavelet-based method (original flow and injection attack are the same).The results which were



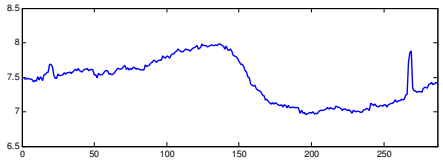
(a) Source IP address entropy sequence



(b) Destination IP address entropy sequence



(c) Source port entropy sequence



(d) Destination port entropy sequence

Fig. 5. Sequence diagrams of entropy sequences after inserting attack

counted by the IP packet data of Netflow traffic data are as shown in Figure 7. The program, parameter settings and etc. completely followed the wavelet-based traffic anomaly detection method introduced in literature [13]. In order to detect as many abnormalities as possible and compare detecting ability with our method, we set all output alarm of the detecting module on, and directly did the precise detecting based on the packet number. The simulation results showed that artificial injected anomalies (between 40 and 49) caused a few changes of the packet number, but the changes were not significant; moreover, from the entire time sampling points, we could not determine the existence of abnormalities at those points.

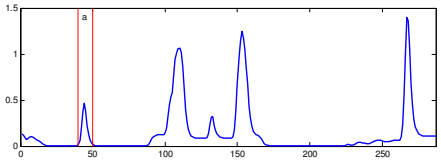
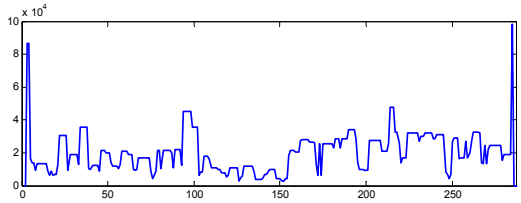
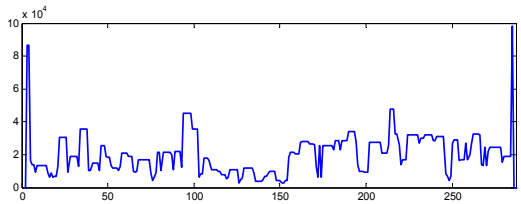


Fig. 6. Sequence diagram of the abnormity coefficient after inserting attack

The results from Figure 6 and Figure 7 show the proposed method has achieved a higher accuracy than the CWT-based method in term of DDos attacks detection.



(a) Detecting results before inserting attack



(b) Detecting results after inserting attack

Fig. 7. Detecting results of the wavelet-based method

IV. Conclusion

This paper proposes a network traffic anomaly detection method based on time-series graph mining. It accurately and completely describes the relationships among multi-time series which are used in traffic anomaly detection by time-series graph, and can effectively detect the network traffic anomaly, especially DDos attacks. In the next step, we will combine traffic characteristic parameters and state characteristic parameters to study more effective time-series graph constructing and mining methods. Further more we will assess network security situation on this basis.

References

[1] A. Hussain, J. Heidemann and C. Papadopoulos, A Framework for Classifying Denial of Service Attacks, *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, Karlsruhe, Germany, 2003

[2] C.M. Cheng, H.T. Kung and K.S. Tan, Use of Spectral Analysis in Defense Against Dos Attacks, *Proceedings of IEEE GLOBECOM*, 2002

[3] V. Alarcon and J.A. Barria, Anomaly Detection in Communication Networks Using Wavelets, *IEEE Proc-Commun*, Vol.148.No.6, 2001

[4] P. Barford, J. Kline, D. Plonka and A. Ron, A Signal Analysis of Network Traffic Anomalies, *In:Proc of ACM SIGCOMM Internet Measurement Workshop*, Marseilles, France, November 2002, 412-423

[5] Jun Gao, Guangmin Hu and Xingmiao Yao, Anomaly Detection of Network Traffic Based on Wavelet Packet, *APCC'06. Asia-Pacific Conference on Communications*, 2006

[6] W. Lee and D. Xiang, Information-Theoretic Measures for Anomaly Detection, *In: Proc of IEEE Symposium on Security and Privacy, Oakland*, CA, May 2001,130-143

[7] A. Lakhina, M. Crovella and C. Diot, Mining Anomalies Using Traffic Feature Distributions. *In: Proc of ACM SIGCOMM 2005*, Philadelphia, Pennsylvania, USA, August 2005, 9-20

[8] http://www.apng.org/9thcamp/matbdfs.ppt

[9] Yuexiang Yang, Hailong Wang and Xicheng Lu, Entropy-Based Classification of Large-Scale Network Traffic Anomalies, *Computer Engineering & Science*, Vol.29 ,No.2 ,2007, 40-43

[10] C.C. Noble and D.J. Cook, Graph-Based Anomaly Detection, *SIGKDD '03*, August 24-27, 2003, Washington, DC, USA

[11] Jiawei Han and M. Kamber, *Data Mining-Concepts and Techniques*, Morgan Kaufmann Publishers, 2000

[12] [EB/OL], http://www.internet2.edu/network/

[13] A. Dainotti, A. Pescape and G. Ventre, Wavelet-based Detection of DoS Attacks, *Proceedings of IEEE GLOBECOM*, 2006