Real Network Traffic Anomaly Detection Based on Analytical Discrete Wavelet Transform

Marius Salagean
Department of Communications, Faculty of Etc, University "Politehnica" of Timisoara, Romania
marius.salagean@etc.upt.ro

***Abstract*-Signal processing techniques have attracted a lot of attention recently in the networking security technology, because of their capability of detecting novel intrusions or attacks. In this paper, we propose a new detection mechanism of network traffic anomaly based on Analytical Discrete Wavelet Transform (ADWT) and high-order statistical analysis. In order to describe the network traffic information, we use a set of features based on different metrics. We evaluate our technique with real traffic dataset, collected over several days period on a university public server. The test results show that the proposed approach accurately detects a wide range of anomalies.**

I. INTRODUCTION

Intrusion detection and network security has become an important problem in today's world. The Internet, the emergence of a variety of wireless networks, the mobility of network hosts along with the vulnerability of present-day software and protocols have a major impact regarding the evolution of network-based attacks. Network intrusions are carried out in various forms: virus, spamming, worms, malware, privilege escalation, unauthorized logins, access to sensitive information, attacks against vulnerable services, injecting unwanted packets into the target networks. As a consequence, this insecure environment has lead to the development of network intrusion detection systems (NIDS).

An intrusion detection system is a software tool that captures and analyzes host-based or network-based information, to identify the attacks that attempt to compromise the integrity, availability or confidentiality of the system/network. Primarily, NIDS are classified into signature/misuse detection and anomaly detection. The signature based detection relies on a database of a predefined set of attack signatures. By analyzing the collected packets and the observed trace left in the system, the attack can be identified. The great advantage of this detection system is that known attacks can be detected reliably with a low false positive rate. However, this requires frequent signatures updates with the latest known attacks. On the other hand, the misuse detection is faced with some difficulties: how to identify an attack that spans multiple packets over multiple discrete events, how to detect new attacks, to name a few.

As an alternative to the signature detection, the anomaly detection technique was introduced. First, the profiles of normal activity are created through various metrics, and afterwards the intrusion is detected when any system behaviours deviate from the baseline profiles. Some of the benefits of the anomaly detection systems are: ability to detect novel attacks, uncertainty regarding what activity the attacker can perform without triggering the alarm, capability to detect insider attacks. But, the anomaly detection has several drawbacks: establishing a normal traffic profile is challenging and can be time-consuming, an inappropriate normal traffic profile can cause poor performances, there are high percentage false alarms that make it very difficult to associate the alarms with the actual attacks that caused them, a malicious user can inject gradually malicious traffic, so that the anomaly detection system considers it as normal traffic.

In terms of traffic premises, there are four types of situations: intrusive but not anomalous (i.e. the detection system falsely reports the absence of intrusion), not intrusive but anomalous (i.e. the detection system falsely reports intrusion), not intrusive and not anomalous (i.e. is not reported as intrusive) and intrusive an anomalous (i.e. is reported as intrusive).

In recent years, as an alternative to the traditional network techniques in anomaly detection (e.g. machine learning based techniques, data-mining based methods), several signal processing techniques have found applications in NIDS: adaptive thresholding, cumulative sum [7], [8], principal component analysis [9], statistical analysis approach [10], [11], Hurst parameter analysis [12].

In this paper, we propose a new network anomaly detection method based on Analytical Discrete Wavelet Transform (ADWT), which consists of five components: feature analysis, wavelet transform, statistical analysis & thresholding, wavelet synthesis and anomaly detection. In the first step, the raw TCPDUMP packet data are converted into network flow logs, based on different basic metrics, used to measure the entire network's behaviour. Next, we employ the wavelet transform for statistical analysis by means of a sliding window. We then, selectively reconstruct the signal only from those wavelet coefficients that surpass the thresholds on each scale. Therefore, the reconstructed signal can be distinct from original signal to a greater degree. The final step is detection, in which attacks and anomalies are checked using thresholds. The thresholds are established through the research of historic traffics. The general architecture is illustrated in Fig. 1.

978-1-4244-7020-4/10/$26.00 ©2010 IEEE

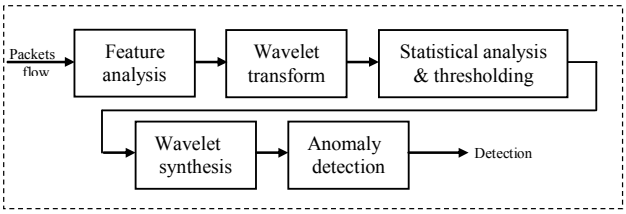

Fig. 1. The model of network traffic anomaly detection

The rest of the paper is organized as follow. Section II introduces related work on applying wavelet analysis techniques for intrusion detection. In section III we describe the ADWT and its improved Inverse Analytical Discrete Wavelet Transform (IADWT) [1]. Section IV presents our method for anomaly detection. In section V we illustrate the simulation results of our approach, and Section VI concludes the paper.

II. RELATED WORK

Techniques relying on the wavelet transform have been widely used recently in NIDS. The property of shift-invariance along with good time-frequency properties lead to improved detection results. Examples of typical works include literature [6-16].

In [6], Barford et al. use multi resolution analysis to evaluate traffic signal filtered only at certain scales. A deviation algorithm is presented to identify anomalies (e.g. flashcrowds, outages, DoS attacks etc.) by setting a threshold for the signal composed from the wavelet coefficients at different frequency levels. In [13], has been presented a framework for real time wavelet analysis of network traffic. The evaluation results show that Coiflet and Paul wavelets have better characteristics in detecting anomalies under the same environment. In [14], Daniotti et al. proposed a cascade architecture to detect volume-based anomalies caused by DoS attacks. The system is made of two different systems – the first one based on adaptive threshold and cumulative sum, and the second one based on Continuous Wavelet Transform.

To address the poor resolution in middle-high frequency of the multi resolution analysis, Chang et al. [15], proposed a new network anomaly detection method based on wavelet packet decomposition, which can adjust the decomposition process adaptively.

In [16], Wei Lu and Ali Ghorbani proposed a new network signal modelling technique for detecting anomalies. The network traffic behaviours are characterized by different features. Next, the normal daily traffic is represented by a set of wavelet approximation coefficients using an ARX model. The output for the normal daily traffic model is the residual that is used as an input signal to the outlier detection algorithm and finally a decision on the intrusion is made.

III. WAVELET ANALYSIS

The Analytical Discrete Wavelet Transform (ADWT) was first introduced in [2], and refers to the one-dimensional (1D) version of the Hyperanalytic Wavelet Transform (HWT).

A 1D wavelet transform (WT) is shift-sensitive if a small shift in the input signal can cause major variations in the distribution of energy between DWT coefficients at different scales. The shift-sensitivity of the DWT is due to the down-samplers used in its computation. In [2, 8] is devised the Undecimated Discrete Wavelet Transform (UDWT), which is a WT without down-samplers. Although UDWT is shift-insensitive, it has a redundancy of 2J, where J denotes the number of iterations of the WT and its implementation requires a large number of different filters.

Abry [3] first demonstrated that approximate shiftability is possible for the DWT with a small, fixed amount of transform redundancy. He designed a pair of real wavelets such that one is approximately the Hilbert transform of the other. This wavelet pair defines a complex wavelet transform. Kingsbury [4, 5] developed DTCWT which is a quadrature pair of DWT trees, similar to Abry's transform. Both transforms are quasi shift-invariant, but the filter design is quite complicated.

The ADWT is a complex wavelet transform, but instead of using complex mother wavelets, it uses regular mother wavelets (such those proposed by Daubechies), but the transform should be applied to the analytical signal x_a associated to the input signal x, computed as:

$$x_a = x + j\mathcal{H}\{x\} \tag{1}$$

where $\mathcal{H}\{x\}$ is the Hilbert transform of the signal.

The implementations of the ADWT and of its inverse (IADWT) we employ in this paper are those proposed in [1]. The decision to use the improved version of IADWT in the reconstruction process was dictated by a few advantages: better shift-invariance properties, simplicity, flexible structure (e.g. can use any orthogonal or biorthogonal mother wavelets) and reduced computation time required.

IV. ANOMALY DETECTION OF NETWORK TRAFFIC METHOD

1. *Feature Analysis*

Because most of the current NIDS use network flow data, the feature analysis block extract network flow features so that network traffic volumes can be characterized and discriminated. Specifically, a network flow should include a source (source IP, source port), a destination (destination IP, destination port), IP protocol, number of packets, number of bytes. In order to measure the network traffic, we use different basic metrics: flow count over a time period, average number of packets in a flow over a time period, average number of bytes in a flow over a time period, average

packet size in a flow over a time period, and ratio of flow count to average packet size [15].

The evaluation of the proposed detection method is done with the real traffic dataset collected on public server of our university, which consists of several days' sniffed traffic (tcpdump files). In this paper we have included the results over one day traffic only. Converting the raw tcpdump files into flow logs takes two steps: the editcap tool is used to split the original tcpdump file into different tcpdump files over one minute time interval, and then the resulted tcpdump traffic data is converted into flow logs based on TCP, UDP and ICMP protocols using the tshark tool.

2. *Wavelet Transform*

The ADWT with 4 level of decomposition is computed using the Daubechies filter. The approximation and detail coefficients are: [cA4, cD4, cD3, cD2, cD1]. With the increase of the levels of the wavelet decomposition, the number of the ADWT coefficients halves. If the length of the detection series is N, the length of the wavelet coefficients series from level j is N/2j.

After weighing the effect of localization (both in time and frequency), vanishing moment, linearity and symmetry, we select Daubechies(db6) filter.

Generally, the common method of deciding upon a wavelet for a certain time series signal is to choose a wavelet that most matches the variations in the data itself. This is an adequate technique when one is dealing with (semi)stationary signals, where the signal frequencies are constant throughout the signal. Because this is valid only for the (semi) stationary signals, where the signal frequencies are constant throughout the signal, in the case of non-stationary signals (like network traffic), the problem is more difficult. Hence, no single wavelet can be easily matched to all the types of traffic and/or the anomalies.

3. *Statistical Analysis and Thresholding*

The statistical detection algorithm is based on the high-order statistics, specifically on the fourth-order cumulant. This high-order statistic hold a very interesting property: has a high value for signals with rapid and impulsive transitions, therefore using it in our context is quite appropriate.

For p and g, we define with M_p^q the complex moment of the signal s in (2) and $\tilde{M}_p^q$ his estimate, in (3).

$$M_p^q[s] = E\left(s^p s^{*q}\right) \tag{2}$$

where $E()$ denotes the expectation value.

$$\tilde{M}_p^q[s] = \frac{1}{N}\sum_{k=1}^{N-1} s^p(k)\, s^{*q}(k) \tag{3}$$

The fourth-order cumulant $\tilde{C}_2^2$ is then computed by:

$$\tilde{C}_2^2[s] = \tilde{M}_2^2[s] - 2\left(\tilde{M}_1^1[s]\right)^2 - \tilde{M}_2[s]\tilde{M}^2[s] \tag{4}$$

For the approximation and every details coefficients obtained by applying the ADWT transform, we calculate the forth-order cumulant using a detection window. The size of the detection windows must be carefully chosen. A smaller window is well suited in high-frequency detection, and a little bigger under low-frequency detection. If the traffic is anomalous in the detection window, there must be an increase in the magnitude of forth-order cumulant.

One important note here is that high-order statistics analysis is able to catch attacks early in the attack launch, far ahead of congestion build-up due to the attacks. In order to determine the thresholds, an initiation procedure and adaptive adjustment are needed. Through the research of historic traffic, we establish the threshold value in each level of decomposition. By setting a high threshold at each level, anomalies can be detected with high confidence. Varying the window size may possibly lead to changes to these threshold values. At the confirmation of anomaly of time t on each level, we preserve the corresponding wavelet coefficient for the reconstruction process.

4. *Anomaly Detection*

The signal is reconstructed only from those wavelet coefficients that reach the alert thresholds. We can get better performance in time localization if we detect the reconstructed signal. If the alert threshold is reached, by measuring the peak height and peak width, one is able to identify anomalies, their duration, and their relative intensity.

V. RESULTS

The network traffic dataset was collected on a public server in our university network. This server provide the following services: http (for our web sites, students and staff resources), ftp (documents sharing), VPN (Virtual Private Network – for staff only), SSH (Secure Shell – for administrative tasks).

The network interface was put in promiscuous mode, in order to capture all data traffic from the network.

There are five major categories of attacks:

1) Denial of service (DOS): an unauthorized attempt to make a computer (network) resource unavailable to its intended users, for example, SYNFlood.

2) Probe: unauthorized probing of a host or network to look for vulnerabilities, explore configurations, or map the network's topology, for example, port scanning.

3) Remote-to-local (R2L): unauthorized access from a remote machine, for example, guessing password.

4) User-to-root (U2R): unauthorized access to local super-user (root) privileges, for example, various buffer overflow attacks.

5) Data: unauthorized access or modification of data on local host or remote host.

Fig. 2 illustrates the feature "number of TCP flows per minute" over one day. Examining the TCP traffic flows from this figure, is hard to detect any attacks at all.

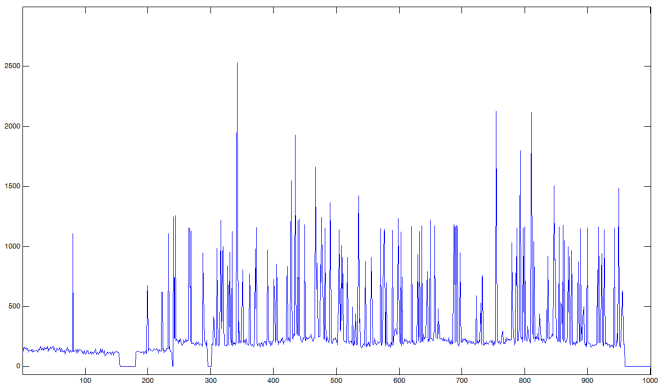

Fig. 2. Number of TCP flows per minute over one day.

The result of the detection from the reconstructed signal for the first day which contains several attacks is illustrated in Fig. 3.

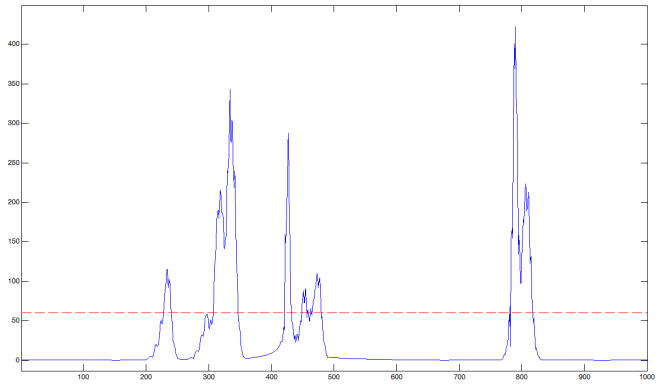

Fig. 3. The detection result based on forth-order cumulant of the reconstructed signal, representing the TCP flows per minute over one day.

It can be observed from Fig. 3, that 6 attacks have been identified: the first two attacks are: *nmap port scan* and *nmap Operating System (OS)/Service detection scan*. All the next four attacks are *proxy scan* attacks on port 445 and 8080, as well as *worms* attack on port 135.

Next, in Fig. 4 is shown the feature "number of UDP flows per minute" over one day. Also, the detection results is depicted in Fig. 5.

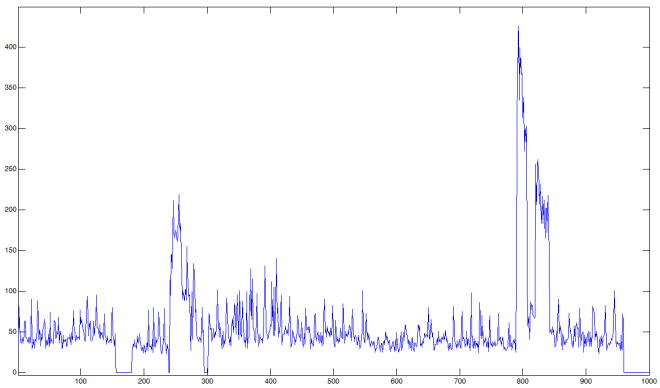

Fig. 4. Number of UDP flows per over one day.

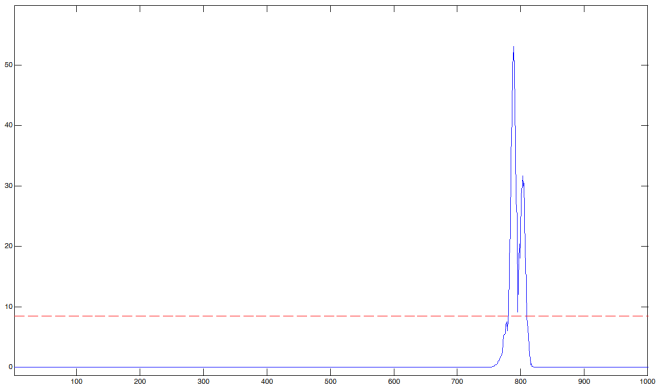

Fig. 5. The detection result based on forth-order cumulant of the reconstructed signal, representing the UDP flows per minute over one day.

From Fig. 5 we can see that an *udp scan* attack have been successfully detected.

An important issue in detecting traffic anomalies is the relationship between the strength of an anomaly's signal in a large amount of other traffic. An anomaly measured close to its source should be very evident while the same anomaly would be less evident if its signal were aggregated traffic. We address this issue, by separating the above captured traffic. Because, the server is an important node in our university network, we filter the sniffed packets that has only or the source or destination IP of the server.

In the following Fig. 6, we illustrates the feature "number of TCP flows per minute" over one day.

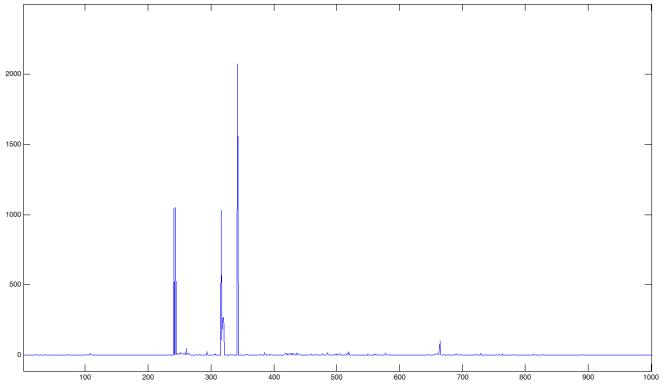

Fig. 6. Number of TCP flows per minute over one day.

The result of the detection from the reconstructed signal for the first day is shown in Fig. 7.

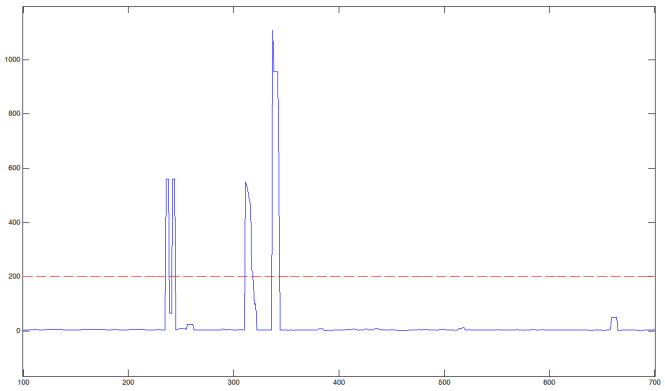

Fig. 7. The detection result based on forth-order cumulant of the reconstructed signal, representing the TCP flows per minute over one day.

It can be noticed, that the proposed detection technique has identify four attacks: the first three attacks are *nmap port scan* and the last one is *nmap Operating System (OS)/Service detection scan* attack. Compared to Fig. 3 the *nmap scan probe* attacks are more precisely identified.

Fig. 8 illustrates the feature "average number of TCP packets per flow over 1 minute".

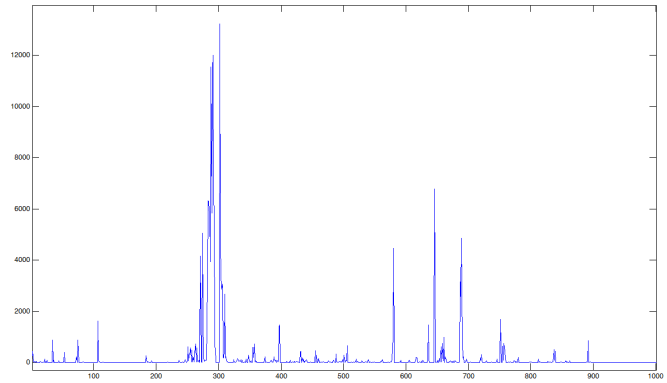

Fig. 8. Average number of TCP packets per flow over 1 minute, in one day.

The detection result is presented in Fig. 9.

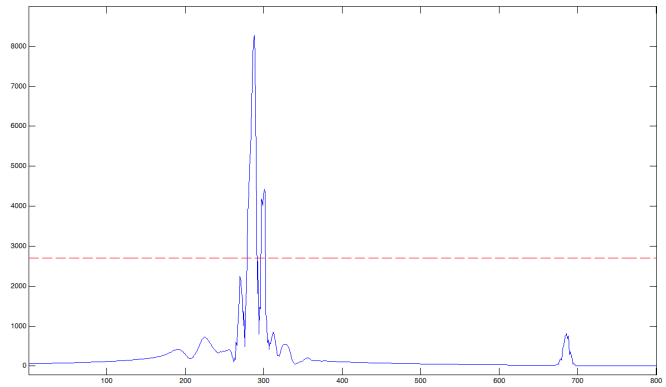

Fig. 9. The detection result based on forth-order cumulant of the reconstructed signal, representing the average number of TCP packets per flow over 1 minute, in one day.

Two incidents happen based on the above detection result: SSH copy of files (probably, due to administrator maintenance tasks).

VI. CONCLUSIONS

In this paper, we have proposed a new detection mechanism of network traffic anomaly based on Analytical Discrete Wavelet Transform (ADWT) and high-order statistical analysis. We evaluate our technique with real traffic dataset, collected over several days period on a university public server. The raw *tcpdump* files were converted into flow logs based on TCP, UDP and ICMP protocols. The statistical detection algorithm is based on the fourth-order cumulant. The test results show that the proposed approach accurately detects a wide range of anomalies.

By setting a smaller threshold at each level, it is possibly that more anomalies could be detected, along with false alarms. In order to reduce the false alarm attacks, further investigations are needed: reading the logs, source ips, the port numbers involved, etc. Furthermore, using the others feature signals, attacks with different characteristics can be detected. In the future we intend to characterize the network traffic behaviour with these features.

The ADWT-based method devised to monitor the traffic activity gives good results, but the proposed implementation may need additional analysis, in particular the issues related to the optimal selection of the alert thresholds need further attention in the future. A solution is to find the thresholds dynamically, according to traffic behaviour over time.

Also, we will focus on applying different wavelet basis functions, and see the impact in detecting the attacks.

REFERENCES

[1] Ioana Firoiu, Alexandru Isar, Jean-Marc Boucher, "An Improved Version of the Inverse Hyperanalytic Wavelet Transform," Proceedings of IEEE International Symposium SCS'09, Iasi, Romania, July 9-10, 2009, ISBN 1-4244-0968-3, 13-16.

[2] I. Adam, M. Oltean, M. Bora, "A New Quasi Sifht Invariant Non-Redundant Complex Wavelet Transform", Scientific Bulletin of the "POLITEHNICA" University of Timisoara, number dedicated to the Symposium on Electronics and Telecommunications ETC 2006 7th Edition, Timisoara, Tom 51 (65), Fascicola 2, 2006 ISSN 1583-3380, pp.14-18, September 2006.

[3] P. Abry, "Transformées en ondelettes-Analyzes multirésolution et signaux de pression en turbulence", Ph.D. dissertation, Université Claude Bernard, Lyon, France, 1994.

[4] N. G. Kingsbury, "Image Processing with Complex Wavelets", Philosophical Transactions of the Royal Society of London A, vol. 357, pp. 2543 – 2560, 1999.

[5] N. G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals", Journal of Applied and Computational Harmonic Analysis, vol. 10, no. 3, pp. 234 – 253, May 2001.

[6] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies," in Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurment (IMW '02), pp. 71–82, Marseille, France, November 2002.

[7] V. A. Siris, F. Papagalou, "Application of Anomaly Detection Algorithms for Detecting SYN Flooding Attacks", IEEE GLOBECOM 2004, Nov. 2004, pp. 2050-2054.

[8] R. B. Blazek, H. Kim, B. Rozovskii, A. Tartakovsky, "A Novel Approach to Detection of Denial-of-Service Attacks via Adaptive Sequential and Batch-Sequential Change-Point Detection Methods", IEEE Workshop Information Assurance and Security, 2001, pp. 220-226.

[9] A. Lakhina, M. Crovella, C. Diot, "Diagnosing Network-Wide Traffic Anomalies", ACM SIGCOMM 2004.

[10] M. Thottan and J. Chuanyi, "Anomaly detection in ip networks," IEEE TRANSACTIONS ON SIGNAL PROCESSING, vol. 51, no. 8, 2003.

[11] C.-M. Cheng, H. Kung, and K.-S. Tan, "Use of spectral analysis in defense against dos attacks," Proceedings of IEEE GLOBECOM 2002, 2002.

[12] W. Allen and G. Marin, "On the self-similarity of synthetic traffic for the evaluation of intrusion detection systems," Proceedings of the 2003 Symposium on Applications and the Internet (SAINT03), 2003.

[13] C.-T. Huang, S. Thareja, and Y.-J. Shin, "Wavelet-based real time detection of network traffic anomalies," in Proceedings of Workshop on Enterprise Network Security and the 2nd International Conference on Security and Privacy in Communication Networks, pp. 1–7, Baltimore, Md, USA, August 2006.

[14] A. Dainotti, A. Pescape, and G. Ventre, "Wavelet-based detection of DoS attacks," in Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '06), pp. 1–6, San Francisco, Calif, USA, November 2006.

[15] J.Gao, G. Hu,X. Yao, and R. K. C. Chang, "Anomaly detection of network traffic based on wavelet packet," in Proceedings of the Asia-Pacific Conference on Communications (APCC '06), pp. 1–5, Busan, Korea, August 2006.

[16] Wei Lu and Ali A. Ghorbani, "Network Anomaly Detection Based on Wavelet Analysis," EURASIP Journal on Advances in Signal Processing, vol. 2009, Article ID 837601, 16 pages, 2009. doi:10.1155/2009/837601.