NETWORK TRAFFIC ANALYSIS - A DIFFERENT APPROACH USING INCOMING AND OUTGOING TRAFFIC DIFFERENCES

RENATO PREIGSCHADT DE AZEVEDO, DOUGLAS CAMARGO FOSTER,
RAUL CERETTA NUNES, ALICE KOZAKEVICIUS
Universidade Federal de Santa Maria - UFSM,
Av. Roraima,1000 - Camobi - Santa Maria,RS
E-mail: rpa.renato@gmail.com, dcfoster@mail.ufsm.br,
ceretta@inf.ufsm.br, alicek@smail.ufsm.br

ABSTRACT

The network traffic analysis is a fundamental area on network management because the network anomalies may affect the network quality of service. However, the data network traffic anomalies are still a critical issue. On last years signal processing methods like wavelet-based ones have been used to detect anomalies on network traffic, specially because wavelet transforms allow the selection of signal characteristics via a combined time-frequency representation. This paper explores a simple and fast wavelet transform for analyzing the network flow, considering the difference between incoming and outgoing traffic data, for improving identification of deny of service attacks.

1 - INTRODUCTION

According to Barford et al [1], computer networks without traffic analysis can not operate with efficiency and security, therefore a fundamental area of network management is related to this issue. Traffic analysis have been used to gather network information, such as the real nature of traffic and the Quality-of-Service (QoS) being provided, and to detect network anomalies, such as deny of service (DoS) attacks (a type of flooding attack). This paper looks at network anomaly detection methods. On last years signal processing derived methods, like wavelet-based ones, have been explored to detect anomalies [1]. Wavelet transforms allow the selection of signal characteristics via a combined time-scale representation, which has been used in many traffic analysis to detect anomalies in one-dimentional trace [2] [3]. In order to help the correct identification of DoS attacks, this paper uses wavelets transforms and signal pre-processing to analyze the network flow considering the incoming and outgoing traffic differences.

The IP traffic can be separated into some primitive protocols, such as TCP, UDP, ICMP, ARP, TCP-SYN, TCP-ACK, etc [4]. When analyzing them separately, it is possible to observe that some attacks are better recognized according to a specific protocol, which composes the entire IP traffic data. Thus, the main idea proposed in this paper is the analysis done after a protocol separation, considering the difference between the incoming and outgoing traffic data, for different primitives. To this new set of information the Haar wavelet transform (a simple and fast wavelet [5]) is applied. After this previous processing, the most significant wavelet coefficients are selected through a threshold strategy in order to identify the starting point of the attack, as well as its duration.

The paper is organized as follows. In the section 2 a brief introduction of the discrete wavelet transform is presented. In section 3 the proposal of this paper is presented . In section 4 some relevant information about the data set considered is presented. In section 5 numerical simulations are discussed and their results are shown. Finally some remarks and conclusions are set up in section 6.

2 - WAVELET TRANSFORM

Wavelets are a mathematical tool for hierarchically decomposing functions [6]. They can be applied to extract information from different kinds of data in different resolution levels. They allow a function to be described in terms of a coarse overall shape, plus a family of details, which corresponds to the complementary information, necessary to recover the original data from one level to the other, until the finest resolution level is achieved.

Technically, wavelets are mathematical functions used to divide a given function or continuous-time signal into different scale components [5]. Because of their inherent multi-resolution nature, wavelet-coding schemes are especially suitable for applications where scalability and tolerable degradation are important. Wavelet analysis has gained widespread acceptance in signal processing and image compression. In this work only discrete wavelet transforms are considered for the traffic analysis.

Generally, in a Discrete Wavelet Transform (DWT) a wavelet representation of a function $f(t)$, defined for all $t \in \mathbb{R}$, (signal) can be given by

$$f(t) = \sum_{k=-\infty}^{+\infty} c_{0,k}\phi(t-k) + \sum_{k=-\infty}^{+\infty}\sum_{j=0}^{J-1} d_{j,k}\psi(2^j t - k). \tag{1}$$

In the discrete case the scale coefficients $\{c_{0,k}\}_{k\in\mathbb{Z}}$ and the wavelet coefficients $\{\{d_{j,k}\}_{k\in\mathbb{Z}}\}_{j=0,..,J}$ can be obtained by the fast cascade algorithm [7]. Considering x_n the discrete time signal, sampled with $n = 2^T$ points as: $x[n] = \{x_0, x_1, x_2, ..., x_{n-1}\}$, the starting point of the fast wavelet transform algorithm is the assumption that scale coefficients $\{c_{J,k}\}$ at the finest initial level J are given exactly by the discrete function (signal) values, therefore for $k = 0, ..., n-1,\ c_{j,k} = x_k$.

Assuming the wavelet function defined by its filters L and H, the fast direct wavelet transform is represented by the following equations, where $\{c_{j-1,i}\}$ and $\{d_{j-1,i}\}$ have the half amount of points than the vector $c_{j,k}$.

$$c_{j-1,i} = \sum_{k=0}^{2N-1} L_k c_{j,2i+k}, \tag{2}$$

$$d_{j-1,i} = \sum_{k=0}^{2N-1} H_k c_{j,2i+k}. \tag{3}$$

The Figure 1 represents a DWT of a signal with 4 levels. The low-pass filters are represented by L, and the high-pass filters are represented by H.

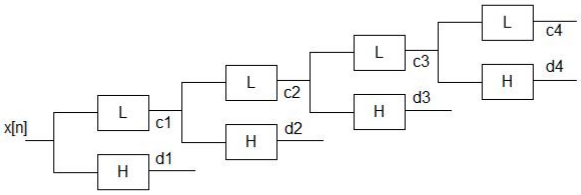

Figure 1: Decomposition of an input signal $x[n]$ with 4 levels.

Before applying the inverse wavelet transform, a threshold operation can be used in order to select the most significant wavelet coefficients, discarding irrelevant information. Thus, the wavelet is a good tool to expose the characteristics of anomalies combining time-frequency representation.

3 - METHOD

The use of wavelets on network anomalies analysis is not new. However, considering the non-stationary feature of a network traffic, the choice of the correct wavelet family is still a challenge [8]. Despite in [8] two wavelets families have been identified as good ones (Coiflet and Mexican) to analyze the network traffic, our hypothesis is that a more simple wavelet family can be used. Thus, in this paper we explore a new approach: the use a simple and fast wavelet not yet explored, the Haar family [5].

The Haar wavelet is the simplest possible wavelet and presents a good capability to identify data variations, as well as to preserve the variations location [5].

For the simulations presented in the section 5, it is considered the Haar wavelet with one null moment to detect the borders of the *neptune* attack [8]. The scale function $\phi(x)$ and the corresponding Haar wavelet function $\psi(x)$ are given by:

$$\psi(t) = \begin{cases} \frac{1}{\sqrt{2}}, 0 \leq t \leq 1/2 \\ \frac{-1}{\sqrt{2}}, 1/2 \leq t \leq 1 \\ 0, otherwise \end{cases} , \quad \phi(t) = \begin{cases} \frac{1}{\sqrt{2}}, 0 \leq t \leq 1 \\ 0, otherwise \end{cases}$$

Note that it operates on data by calculating the sums and differences of adjacent elements.

In our analysis we are going to consider a threshold value of 70% of $Max = Max|d_{jk}|_{,\forall k=0,\ldots,n_j-1}, n_j = 2^{T-j}$ per level. This operation will allow a faster data manipulation, and consequently a faster data analysis, in order to identify the location of strong variations inside the signal.

4 - NETWORK DATA

In this paper a MIT DARPA (Defense Advanced Research Projects Agency) [9] intrusion detection data set is used, as in [8]. This data set contains a collection of five weeks of network activity. The data related to the first and third week of analysis do not contain any attacks, simplifying the calibration of the Intrusion Detection System (IDS). The second week contains a labelled set of attacks. The fourth and fifth weeks are formed by realistic data, which contains several attacks to the servers, without any label.

Several log files are included in this data set, but only incoming and outgoing traffic are considered in this paper. Each week of analysis was divided in 5 days, and each day was logged for 22 hours. In our experiments the following counters of traffic data are considered: total IP packets, TCP packets, ICMP packets, UDP packets, and TCP-SYN packets. The Table 1 shows a sample of labelled attacks from week 2, including the day and the hour of the attacks.

	Date	Time	attack
1	03/08/1999	08:50:15	ping-of-death
2	03/10/1999	12:02:13	satan
3	03/11/1999	09:33:17	satan
4	03/12/1999	09:18:15	ping-of-death
5	03/08/1999	15:57:15	land
6	03/09/1999	08:44:17	portsweep
7	03/10/1999	13:44:18	mailbomb
8	03/11/1999	09:33:17	satan
9	03/11/1999	11:04:16	neptune
10	03/12/1999	09:18:15	ping-of-death
11	03/12/1999	11:20:15	neptune
12	03/12/1999	17:13:10	portsweep

Table 1: Attacks on DARPA second week

5 - EVALUATION RESULTS

We have used a simple and fast wavelet transform to evaluate the traffic data in three different ways. Firstly the incoming traffic is analyzed, secondly the outgoing traffic is analyzed, and finally the difference between both incoming and outgoing traffic is analyzed with the same wavelet basis. Therefore this paper concentrates its analysis on an attack of TCP-SYN flood DoS, called Neptune on DARPA report. According to [8] when a TCPD server receives a TCP-SYN message, it allocates some resources for the expected connection. The *neptune* attack floods the server with TCP-SYN messages resulting in overflow in TCPD resources for half-open connections. According to [10] DoS attacks are made by putting a large number of requests to one resource, thus keeping the resource too busy or too full to be able to handle with legitimate requests. Since the attacker needs to establish large number of requests, the traffic activity increases.

In the graphs of Figure 2 are showed network traffic flows, with 512 samples (one sample/sec), being the Y axis the number of packets per second. The X axis presentes the associated time stamp of network traffic. The boxes over

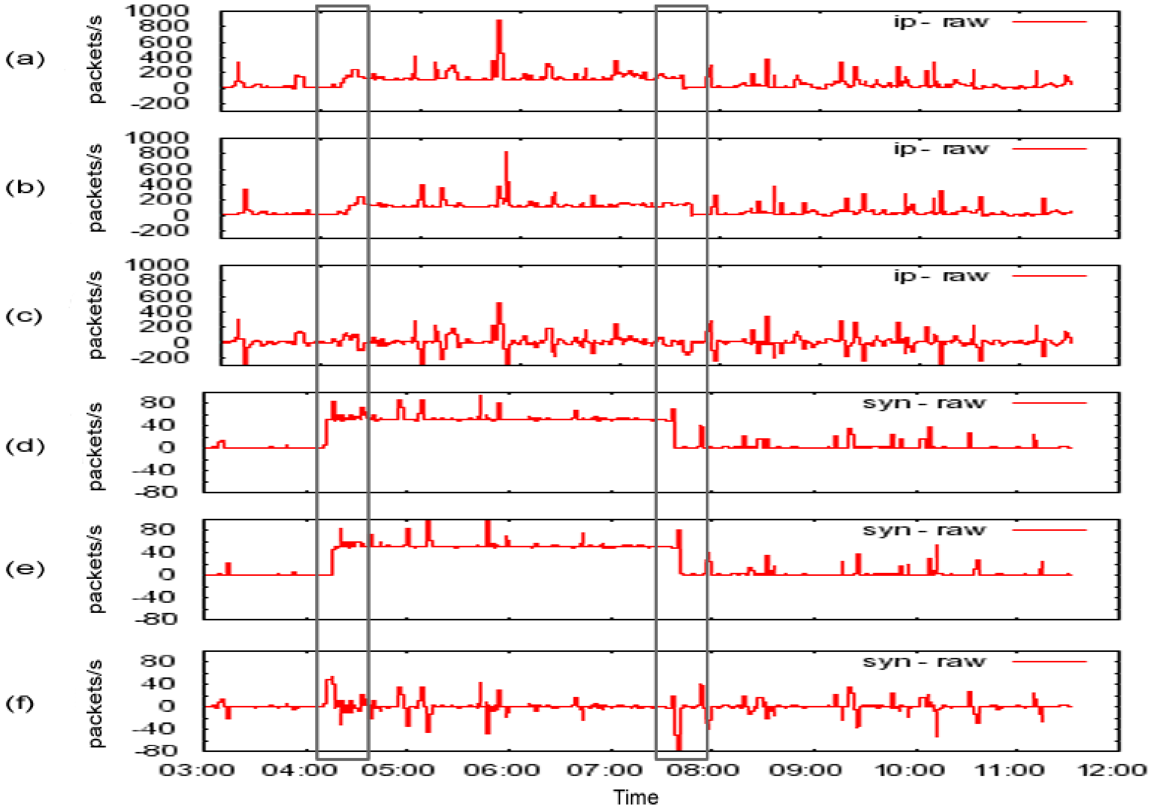

Figure 2: *neptune* attack occurred in 11/03/1999. (a) Input, (b) Output and (c) Variance (input-output) in IP flow. (d) Input, (e) Output and (f) Variance (input-output) in TCP-Syn flow.

the graphs of Figure 2 point out the starting and ending time of the *neptune* attack from a labelled DoS attack, whose occurrence is mentioned in the DARPA report as shown in Table 1. In Figure 2(a),(b),(c) are plotted, respectively, the incoming, outgoing and difference of IP flow of network traffic. In Figure 2(d),(e),(f) are showed, respectively, the incoming, outgoing and difference of TCP-SYN flow of network traffic. In Figure 2(c) the difference between incoming and outgoing traffic data did not show clearly the starting and final instants of the *neptune* attack. It has a lot of spikes instead of the spikes on the moments of the attack. Figure 2(f) shows the difference between incoming and outgoing TCP-SYN network traffic, and shows that the start and end points of attack has suffered relevant variance in the signal strength.

The Haar wavelet transform was then applied in the signal resulting from difference of both network traffic flows (IP and TCP-SYN). The numerical experiments are made considering data from the second week of the training sample. In the neighborhood of the attack occurrence, a 512 sample was chosen for the wavelet analysis, corresponding to 512 seconds of the traffic flow (one sample/sec). Associated to the Haar wavelet transform, a threshold operation is applied to the wavelets coefficients of the decomposition. For each level the threshold value is chosen as 70% of the maximum detail value.

In the first numerical experiment, the input signal is the difference between incoming and outgoing network traffic from IP flow (Figure 3(a)). In Figure 3(b),(c),(d),(e),(f),(g) the levels 1 to 6 of wavelet coefficients are shown. They correspond to details after threshold application. The scale coefficients in level 6 (Figure 3(h)) show that exploring details in more levels do not contribute for the analysis. Looking the graph in Figure 3(b), that corresponds to wavelets coefficients at level 1, we do not recognize the borders of the attack. Note that the borders can not be recognized on levels 2,3,5,6 (see Figure 3(c),(d),(f) and (g)). In level 4 (Figure 3(e)) the wavelet coefficients identifies only the end time of *neptune* attack. Thus we noted that in no one of the levels of wavelet coefficients presented in Figure 3 is identified both the starting and ending points of the *neptune* attack. As result it is difficult to localize the *neptune* attack borders looking only the incoming and outgoing difference from IP flow.

In the second numerical experiment, the input signal is the difference between incoming and outgoing network

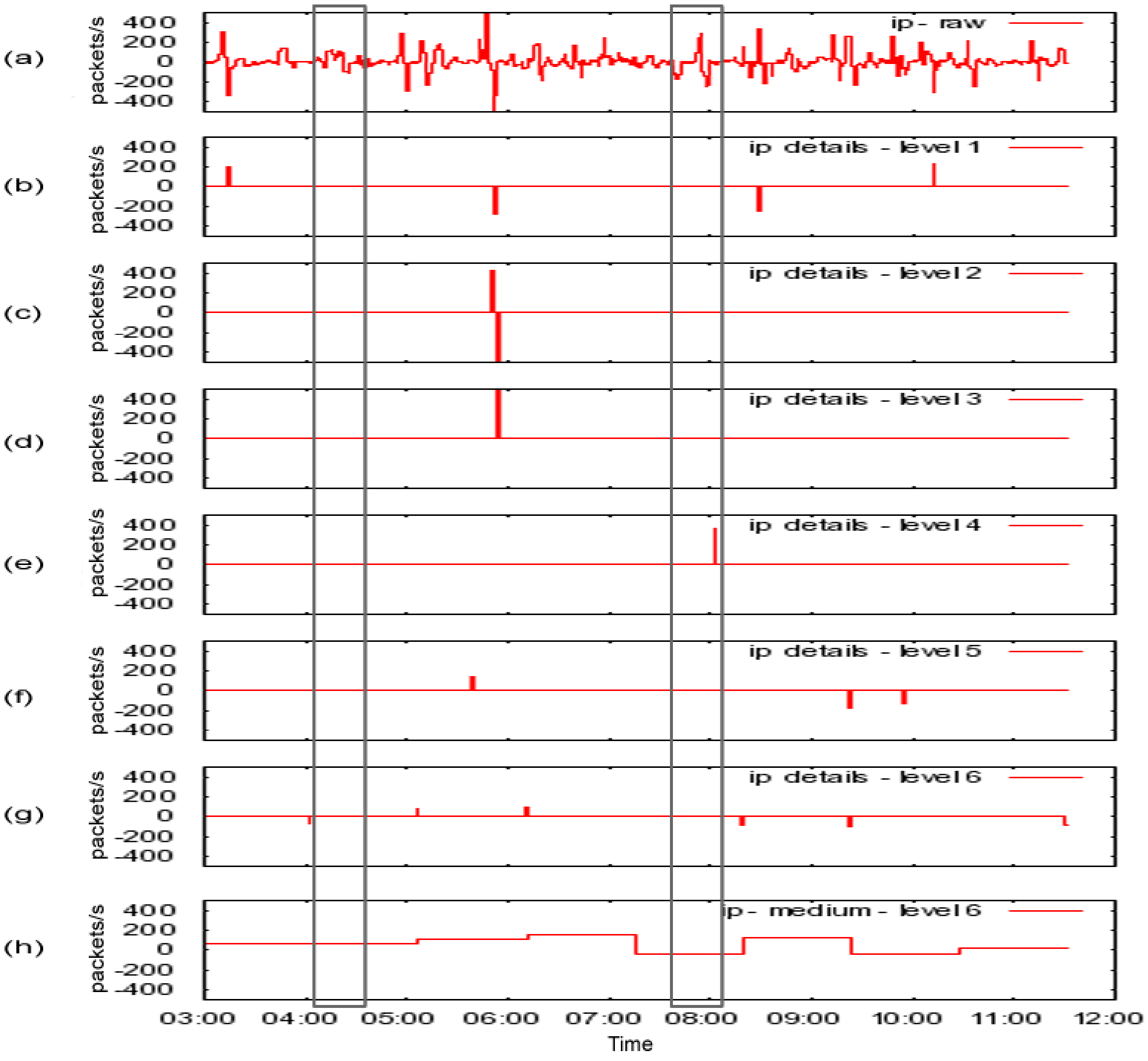

Figure 3: Numerical experiment - *neptune* attack occurred in 11/03/1999. Original Signal - IP, and wavelet coefficients for level 1 to 6 and scale coefficients for level 6.

traffic from TCP-SYN flow (Figure 4(a)). The Figure 4(b),(c),(d),(e),(f),(g) shows the wavelet coefficients in levels 1 to 6. In the similar way that in the first numerical experiment the graph in Figure 4(h) shows that it does not make sense exploring more levels of decomposition. Looking the graph in Figure 4(b) (wavelets coefficients at level 1) we observe a lot of spikes disturbing the location of the attack borders. However, in Figure 4 (c),(e), and (f) the graph from wavelet coefficients in level 2,4,5, respectively, identifies the starting and end times of *neptune* attack. The graphs in Figure 4(d) from wavelet coefficients in level 3 identifies the end time of the attack and some other irrelevant points. In the wavelet coefficients in level 6 the attack cant be recognized (Figure 4(g)). Consequently in three levels of wavelet coefficients (2, 4 and 5) we clearly detect the starting and ending instants of the *neptune* attack.

These results show that the use of a simple and fast wavelet family on analyzing the incoming and outgoing traffic differences can localize the instant and duration (difference between start and end attack borders) of the network attacks.

6 - CONCLUSIONS

In the last years signal processing methods, as wavelet-based ones, have been used to detect anomalies in the network traffic data. In this paper we have applied methods for anomalies detection using wavelet transform and signal pre-processing to analyze the network traffic looking at the difference between incoming and outgoing traffic.

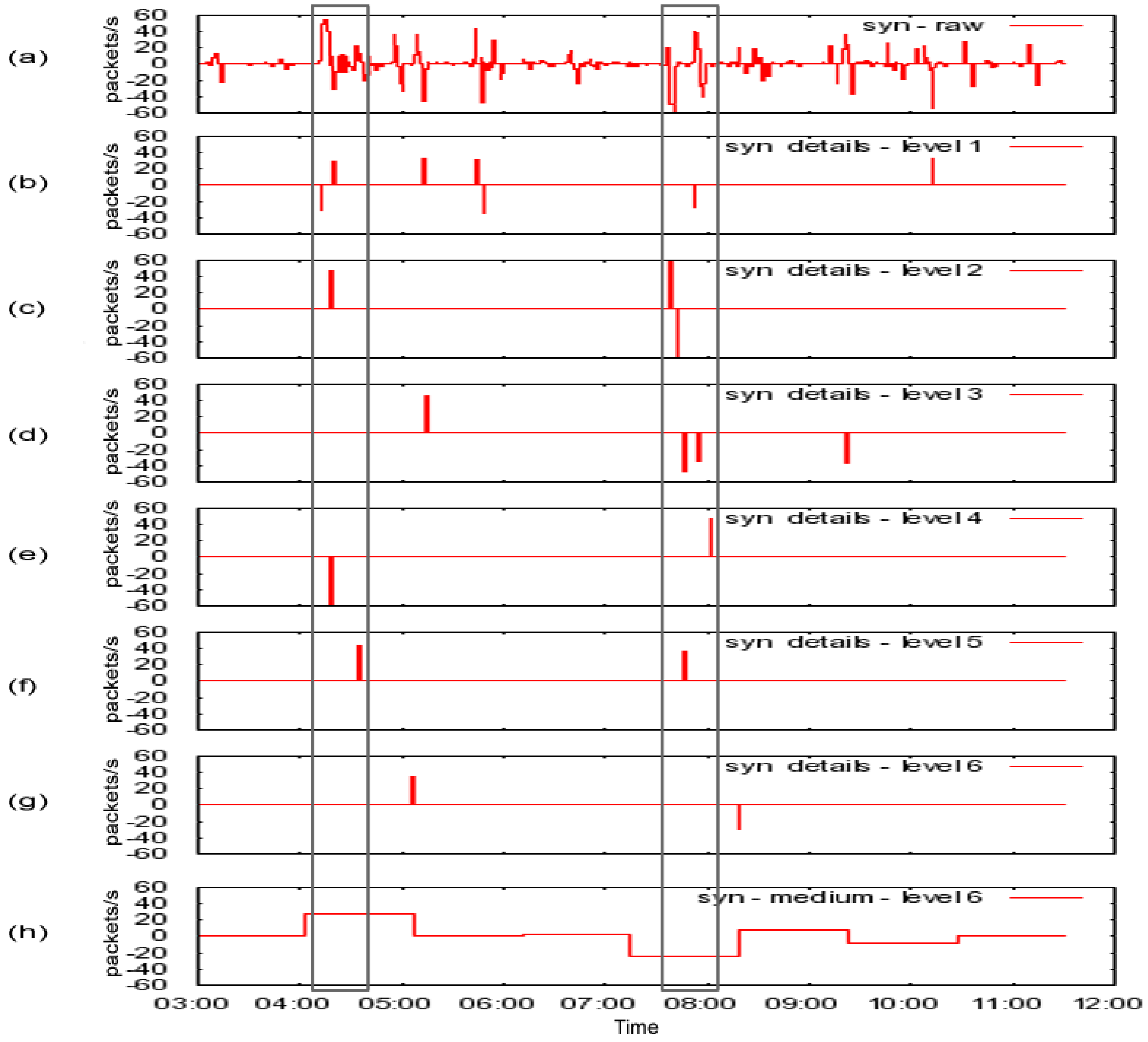

Figure 4: Numerical experiment - *neptune* attack occurred in 11/03/1999. Original Signal - TCP-SYN, and wavelet coefficients for level 1 to 6 and scale coefficients for level 6.

Using the Haar wavelet transform the analysis of difference between incoming and outgoing traffic data in conjunct with a more specific protocol (TCP-SYN), showed that it is possible to clearly identify the starting and ending time of the *neptune*, using the (a simplest and fast wavelet family). This approach offers an alternative for recognizing attack location and can optimize the process to identify the anomaly on network traffic data, providing better information for intrusion detection systems. The Haar wavelet transform (through a very simple and fast algorithm) do not change the time scale, allowing an accurate location of the beginning and the end of a attack.

For future work we will implement an automated anomalies classifier capable to identifies the class of anomalies in network traffic data.

References

[1] P. Barford, J. Kline, D. Plonka, and A. Ron, “A signal analysis of network traffic anomalies,” in *IMW ’02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*. New York, NY, USA: ACM, 2002, pp. 71–82.

[2] S. K. Soule, A. and N. Taft, “Combining filtering and statistical methods for anomaly detection,” in *IMC’05, Oct. 2005*, 2005.

[3] T. M. Lu, W. and A. Ghorbani, “Detecting network anomalies using different wavelet basis functions.”

[4] W. R. Stevens and K. Fall, *TCP/IP Illustrated: The Protocols v. 1*. USA: Addison-Wesley Publishing Company, 2009.

[5] O. Nielsen, “Wavelets in scientific computing,” 1998. [Online]. Available: citeseer.ist.psu.edu/nielsen98wavelets.html

[6] E. Stollnitz, A. DeRose, and D. Salesin, “Wavelets for computer graphics: a primer.1,” *Computer Graphics and Applications, IEEE*, vol. 15, no. 3, pp. 76–84, May 1995.

[7] S. Mallat, *A wavelet tour of signal processing*. Academic Press, 1998.

[8] C. T. Huang, S. Thareja, and Y. J. Shin, “Wavelet based real time detection of network traffic anomalies,” in *Securecomm and Workshops, 2006*, 2006, pp. 1–7.

[9] Darpa, “Mit lincoln laboratory: Information systems technology,” MIT, Tech. Rep., 1999. [Online]. Available: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/

[10] H. Xia and W. Xu, “Research on method of network abnormal detection based on hurst parameter estimation,” in *Int. Conf. on Computer Science and Software Engineering*, 2008, pp. 559–562.