Pattern Recognition Approach for Anomaly Detection of Web-based Attacks

Aruna Jamdagni[1,2], Zhiyuan Tan[2], Ren Ping Liu[1], Priyadarsi Nanda[2], Xiangjian He[2]

[1]CSIRO ICT Centre, [2]Faculty of Engineering & Information Technology, UTS

1. Introduction

The universal use of the Internet has made it more difficult to achieve a high security level. Attackers target web applications instead of Telnet ports. Cyber-attacks and breaches of information security are increasing in frequency. The goal of Intrusion Detection Systems (IDSs) is to monitor and detect web-based attacks. The commonly used IDSs are: signature based IDSs and anomaly based IDSs. Signature based IDS is unable to detect novel attack (i.e., zero-day) or polymorphic attacks, until the signature database is updated. On the other hand, an anomaly-based IDS can detect new attacks and polymorphic attacks. However, anomaly based system has a relatively high number of false positives.

2. GSAD Model

A brief description of our GSAD model [1] is given below. The model is extended for HTTP environment to detect web based attacks using HTTP service. The model uses pattern recognition technique used in image processing to calculate the correlations between various payload features. The key components of GSAD model are 1-gram frequency model and Geometrical Structure Model (GSM). The GSM detects similarity between the behavior of new input traffic profile and that of the developed normal traffic profile using Mahalanobis Distance Map (MDM). The correlations among different features (256 ASCII characters) are calculated using equation (1) to (3). Here, μ is the average frequency of each ASCII character presenting in the payload. $\sum_i$ is the covariance value of each feature, and *D* is the MDM of a network packet. Weight factor *w* is calculated using eq. (4) to detect an intrusive activity. In eq.(4) $\bar{d}_{nor(i,j)}$ and $\sigma^2_{nor(i,j)}$ are the averages and variances for all elements *(i, j)* of the distance map $D = [d_{(i,j)}]_{256x256}$ of the normal profile, and $d_{(i,j)}$ is the element of the distance map $D_{obj} = [d_{(i,j)obj}]_{256x256}$ of the new incoming object.

$$\sum_i = (x_i - \mu)(x_i - \mu)' \; (0 \le i \le 255) \quad (1)$$

$$d_{(i,j)} = \frac{(x_i - x_j)(x_i - x_j)'}{\sum_i + \sum_j} \quad (0 \le i, j \le 255) \quad (2)$$

$$D = \begin{bmatrix} d_{(0,0)} & d_{(0,1)} & \cdots & d_{(0,255)} \\ d_{(1,0)} & d_{(1,1)} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \\ d_{(255,0)} & d_{(255,1)} & \cdots & d_{(255,255)} \end{bmatrix} \quad (3)$$

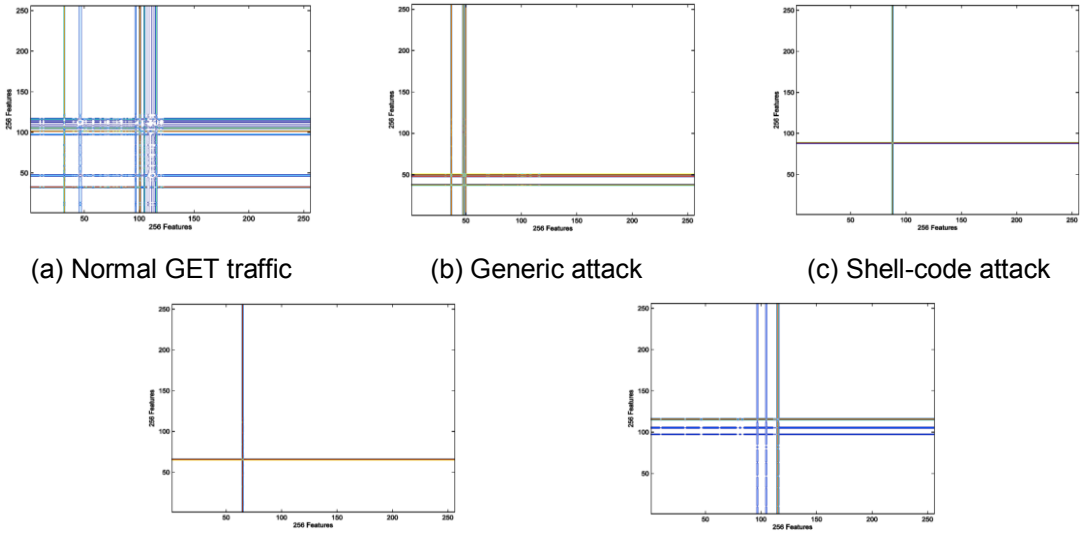
$$w = \sum_{i,j=0}^{255} \frac{(d_{obj(i,j)} - \bar{d}_{nor(i,j)})^2}{\sigma^2_{nor(i,j)}} \quad (4)$$

We parse 150 bytes of HTTP GET request payload by using a sliding window of 1 byte length and count the occurrence frequency of each feature in the payload. The HTTP GET request payload is represented by a pattern vector in a 256–dimensional feature space. A profile is created for HTTP GET request payload using equations (1) to (3).

3. Experimental Results

We Train the GSAD model on 10 days normal HTTP GET request traffic (DARPA 1999 IDS dataset [2]) and generate average normal profile for the HTTP GET request using equations (1) to (3). The total numbers of packets used for training of the model after filtering are 13,933 for host marx. Fig. 1 (a) shows the MDM result of normal HTTP traffic behavior for host marx.

We test our GSAD model on GATECH attack dataset [3] and evaluated similarity between the MDM results of new incoming packet profile and the normal profile using Weight factor and threshold value. The incoming request is considered as an attack or a threat if the Weight factor is more than +3δ or less than -3δ. The attacks are divided into four groups, namely Generic attacks, Shell-code attacks, CLET attacks and Polymorphic Blending attacks.



(a) Normal GET traffic (b) Generic attack (c) Shell-code attack

(d) Polymorphic attack (e) polymorphic blending attack

Fig 1: Patterns for normal and attack traffic

5. Analysis of Results

The test patterns for various attacks are shown in Figs. 1 (b) – (e). The figures show clear differences between the behavior of the various attack profiles and the normal HTTP request profile. Furthermore, the correlations between the features in these attacks are different from the correlations between the features of normal HTTP request on the host marx. X axis and Y axis show the 256 possible features (ASCII characters) present in a packet payload. The cross points in the figure represents the correlation between two features. The results show that the Geometrical Structure Anomaly Detector (Mahalanobis Distance Map) can detect new attacks including polymorphic attack and Polymorphic Blending attack without prior knowledge of the attacks with high accuracy and low false alarm rates.

6. Conclusion

In this paper, we use our GSAD model for web-based HTTP traffic. We evaluated the performance of our GSAD model in real environment. The results show high detection rates and low false positive rates with the proposed GSAD approach. The best performance is achieved for threshold values selected between 3 standard deviations (-3δ and +3δ). .

6. References

[1] A. Jamdagni, Z. Tan, P. Nanda, X. He and R. Liu, Intrusion Detection Using Geometrical Structure, in: 4th International Conference on Frontier of Computer Science and Technology, 2009 pp. 327 - 333

[2] MIT Lincoln Laboratory. 1999 DARPA Intrusion Detection Evaluation Data Set. http://www.ll.mit.edu/IST/ideval/data/1999/1999_data_index.html, 1999

[3] R. Perdisci, D. Ariu, P. Folga, G. Giacinto, W. Lee, McPad: A multi classifier system for accurate payload-based anomaly detection, in: Science direct, Computer Networks 53 (2009), pp 864-881

Aruna Jamdagni Zhiyuan Tan Ren Ping Liu Priyadarsi Nanda Xiangjian He