A Study of IDS using Discrete Fourier Transform

Enkhbold Chimetseren , Keisuke Iwai , Hidema Tanaka and Takakazu Kurokawa
Dept. of Computer Science
National Defense Academy of Japan
Kanagawa, Japan
e-mail:{em52045,iwai,hidema,kuro}@nda.ac.jp

***Abstract*—Intrusion Detection System (IDS) detects attacks using pattern files which are known as "signature". Effectiveness of detection depends on the kind of signature. In this paper, we propose a signature generation method using Discrete Fourier Transform. Our method regards payload between client and server as discrete waveform. Regarding normal communication spectrum as noise, we can clarify the characteristics of attack sessions. From the viewpoint of spectrum analysis, our method detects attack sessions. Furthermore, it has dynamic analysis features like anomaly type of IDS and will be able to detect unknown attack session. Our proposal method simulated using a Kyoto2006+ data set which is currently used as an intrusion detection evaluation. As the result, we have 5% of false positives for detecting attacks.**

Keywords—IDS, false positive, signature, Kyoto 2006+ Data set

I. INTRODUCTION

Intrusion detection system (IDS) is a system for monitoring unauthorized access to network services. IDS monitors the host which provides the network services, and detects malicious penetration. Currently, many methods and techniques are proposed, and produced in practical use. There are many types of IDS construction such as signature based, machine learning, simulation based and so on. Almost IDS is classified into signature based which uses a signature file generated from analysis of known attack schemes. Therefore, signature based IDS can detect almost known attack scheme, however, they cannot detect unknown attack schemes. On the other hand, for example, machine learning IDS will be able to detect unknown attacks, but there is problems in the generation schemes for training data and poor detection rate against both known and unknown attacks. In this paper, we propose a signature generation method using Discrete Fourier Transform (DFT) to improve detection rates against unknown attacks.

Conventional signatures are generated by using methods of distribution or frequency of the packet content or the frequency of character occurrence. On the other hand, the proposed signature is generated by analyzing the DFT spectrum of the each communication payload. The advantage of using DFT spectrum, there is no influence of the contents, packets and changes of the protocol order. Furthermore, it can be evaluated quantitatively difference between normal and attack sessions. Therefore, it is possible to detect unknown attacks by comparison with the spectrum of normal session. We used a Kyoto2006+ data set to evaluate the effectiveness of our proposal method. As the result, we have 5% of false positives for detecting unknown attacks. We conclude that our method is more effective than existing schemes.

II. OVERVIEW OF IDS

A. Classification of IDS

The construction scheme of IDS can be categorized into network-based and host-based type from the viewpoint of monitoring method. And intrusion detecting scheme can be categorized into signature type and anomaly type.

Network-based type (NIDS: Network Intrusion Detection System) monitors traffic flows on the network by analysis of the protocol header. Host-based type(HIDS: Host Intrusion Detection System) monitors the status of the particular host by analyzing received files, access attempts, system logs or other defining characteristics which contain suspicious activities.

Signature type checks packet, and judges whether the characteristic string used for unauthorized access is contained or not. Signature type judges the unauthorized access, when the extracted feature from the packet which matches malicious pattern in the signature [1][10]. The signature always needs to update for discovered new attacks. However, an update of the signature takes a time. Therefore the time-lag before updating will diffuse discovered new attack and its damage.

Anomaly type creates behavior model between user and systems for the monitored infrastructure. Anomaly type defines the normal behavior and deviation threshold between monitored behavior and normal. When deviation of monitored behavior is greater than defined threshold, anomaly type judges it malicious [2][3].For example, frequent occurrence of ping command is judged as suspicious behavior.

B. False Positive and False Negative

False positive is an event that the IDS identifies as an intrusion when none or normal action has occurred. Most current IDS have very high rate of false positives, as they cannot yet make wise decisions on whether the traffic is unauthorized access or normal access.

False negative is an event that the IDS fails to identify as intrusion when unauthorized access has occurred. In other words, unknown attacks were included in the access which failed to identify with high probability.

IDS with huge false negative is useless from the viewpoint of detecting unknown attack. Therefore, in this paper our primary goal is to develop the scheme the detecting unknown attacks with by low rate false positive. In our study, our target is a signature type of IDS.

978-1-4799-6956-2/14/$31.00 ©2014 IEEE

III. Proposed Method

A. Outline

The existing IDS has the following problems.

1. Definition of "normal" behavior
2. Detect unknown attacks

In this study, we will attempt to solve the above problems by regarding the communication status between the server and clients as discrete waveform. In addition we use DFT to perform spectrum analysis. Our method will be able to solve followings.

1. Since normal users have various demands, the normal communication packet exchange will have various forms. Thus, it is difficult to categorize normal behavior into a certain form. However, from the viewpoint of DFT spectrum analysis, normal communication can be regarded as white noise. In other word, we can define normal behavior as noise.
2. Some features can be discovered comparing normal session with attack session. Existing signature generation methods discovers such features using packet analysis [5][7]. In our proposal method, we can find features as difference of peak point in spectrums between normal and attack session.

In addition, the advantage of using DFT spectrum, there is no influence of the contents, packets and changes of the protocol order. FIG1 shows an overview of the proposal method.

B. Proposal scheme for signature generation

The procedure of our proposal scheme is as follow.

Step1. Making "Connection State"
Step2. Adapting DFT of Connection State
Step3. Signature generating from spectrum analysis

In Step1, we regard packet exchange between client and server as discrete waveform. At that time, we define the amount of packet from client to server as plus value and amount of packet in another direction as minus. Then we can draw the discrete waveform such as "Connection State" in FIG1.

In Step2, we apply DFT to Connection state. DFT is defined as follows.

$$X(k) = \sum_{n=0}^{N-1} X(n) e^{i\frac{2\pi k}{N}} \qquad (1)$$

Where $k=0,\ldots,N-1$ sample number,
$X(k)$ N complex sequence,
i imaginary unit,
e Napier number,
π Pi.

As the results, we can get the resultant spectrum form as "Connection Spectrum" in FIG1.

In Step3, from the analysis of resultant spectrum, we define signatures.

IV. Experiment

A. Kyoto2006+ Dataset

In this paper, we used the Kyoto2006 + dataset [11].This dataset was created by Kyoto University project from actual data traffic (November 2006 ~ August 2009) by using the honey pot which is installed in some organizations. Kyoto2006+ dataset is composed 14 statistical characteristics, 10 additional features. It can easily verify the effectiveness of the proposed method.

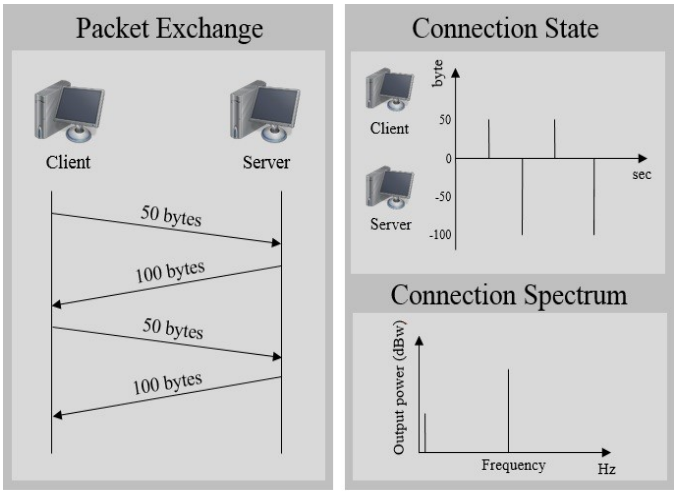

FIG1. Outline of algorithm

TABLE I shows the entire summary of the data set. TABLE II shows the additional features and statistical features.

TABLE I. Summary of data

	Number of session	Average number of session per day
Total	93,076,270	93,638
Normal	50,033,015	50,335
Known attack	42,617,536	42,874
Unknown attack	425,719	428

TABLE II. Feature vectors

	Statistical feature	Additional features
1.	Duration	IDS_detection
2.	Service	Malware_detection
3.	Source bytes	Ashula_detection
4.	Destination bytes	Label
5.	Count	Source_IP_address
6.	Same_srv_rate	Source_Port_number
7.	Serror_rate	Destination_IP_address
8.	Srv_serror_rate	Destination_Port_number
9.	Dst_host_count	Start time
10.	Dst_host_srv_count	Duration
11.	Dst_host_same_src_port_rate	
12.	Dst_host_serror_rate	
13.	Dst_host_srv_serror_rate	
14.	Flag	

In this paper, we used the occurred time and payload which are transmitted through source IP address and destination IP address. Furthermore, we referring to the label in additional feature that has been classified as normal session, known attack session and unknown attack session.(shaded area in TABLE II)

B. *Our dataset*

We did not use all the Kyoto2006+ dataset because some attack sessions can be detected easily by other existing techniques. Typical easily detectable attack sessions are as follows.

1. 0 byte of payload
2. 0 second of session time
3. Huge payload
4. Too long session time

These are considered by unauthorized accesses which send MSSQL stack BO, lot of SYN and packet of big size as shown in [8]. And, they are easily detectable, thus we excepted such session data. From the viewpoint of calculation of DFT, we need to define the threshold value of session time. In this paper, we set the threshold session time as 2 seconds. And we excepted more than 2 seconds session data. In this way, we create our dataset to evaluate proposal scheme. Sessions contained in our data set can be classified into following three types.

1. Normal session
2. Known attack session
3. Unknown attack session

The ratio of in our dataset is shown in FIG 2.

C. *Classification of communication form*

From the viewpoint of communication conditions, we can classify dataset into 2 types; transmission relation and amount of payload. Transmission relation has four categories.

1. One server – one client
2. One server – multiple clients
3. Multiple servers – one client
4. Multiple servers – Multiple clients

Since type 3 and 4 is not adequate for our purpose, we excepted these conditions. Amount of payload has two categories.

1. Constant payload
2. Various payload

In the followings, we consider the spectrum analysis according to these classifications.

D. *Result of spectrum analysis*

The classifications shown in Section IV.C influence the form of resultant spectrum. So, we need to prepare normal spectrum signatures according to the communication conditions. FIG 3~5 shown the resultant spectrums. Note that our data set does not contain "one server-multiple clients with constant payload". From these resultant figures, we can find following common features.

1. The maximum spectrum peak exists at the center frequency.
2. The second spectrum peak exists at 1[Hz].
3. The spectrum has symmetrical form to the center frequency.

FIG 3 and 4 show the comparison of spectrums of normal session and known attack session. Note that, in these cases, there is no data for unknown attack session. From FIG 3, we can find the difference between normal session and attack session easily. Because the values of known attack session is greater than these of normal session. About 40% of known attack session in our data is classified into these case. In the same way, we can detect known attack session from FIG 4.

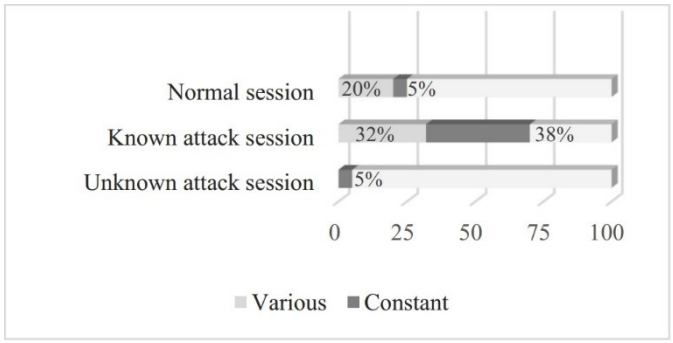

FIG 2. Ratio of each session

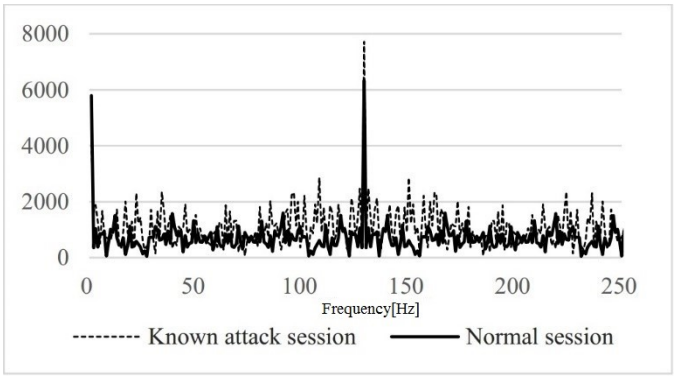

FIG 3. One server – multiple clients with various payload

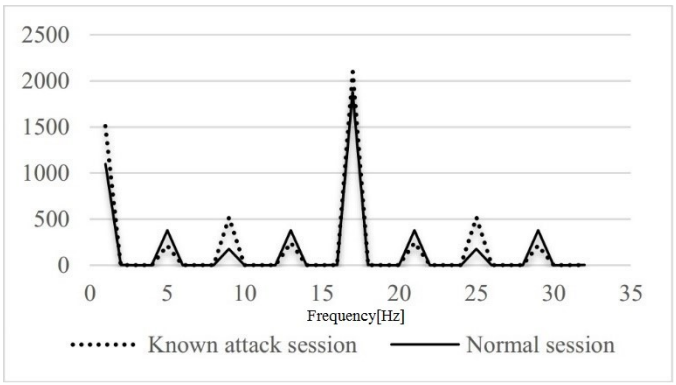

FIG 4. One server – one client with various payload

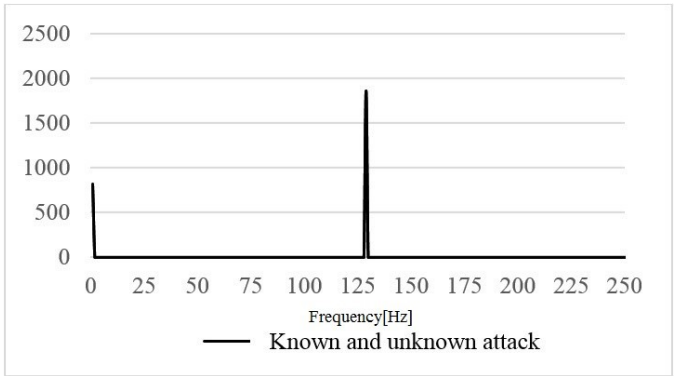

FIG 5. One server – one client with constant payload

In the case of one server- one client with various payload, there is the difference in the second peak position. About 6% of the known attack session in our data is classified into this case. As the result, for the conditions of one server- multiple clients with various payloads and one server – one client with various payloads, our proposal method can distinguish attack sessions from normal sessions perfectly.

FIG 5 shows the resultant spectrum for one server- one client with constant payload. About 54% of the known attack session and all of unknown attack session are classified into this result. And also 5% of normal session is classified into this type, too. In this case, it is very difficult to distinguish normal session from known/ unknown attack session.

As described in Section II.B, since we take false positive strategy, we judge that this type of session is attack session.

E. Summary of experiment

From the results shown in Section IV.D, we can conclude that our method is effective against sessions with various payload. The successful detection for this condition is 100%. On the other hand, for sessions with constant payload, our method is not so effective. In the case of our dataset, the ratio of normal session with constant payload is about 5%. Thus, in the strategy of false positive, the error rate is about 5%. Therefore, we can conclude that our method is with 5% of false positive.

V. DISCUSSION

From the result of experiment, we can conclude that our method is effective against complex sessions such as multiple servers – multiple clients, multiple servers – one client, one server – multiple clients with various payload. On the other hand, our method is not so effective against simple sessions. In particular, one server – one client with constant payload is a typical case. Unfortunately, almost unknown attack session in Kyoto2006+ dataset is categorized into this case. Therefore, it is necessary to evaluate using another dataset or newest one. This is our feature work.

To simplify our proposed method, we set session time within 2 seconds. So the result of evolution may have inclined. In addition, we determined "2 seconds" from the average time of Kyoto 2006+ dataset. The following methods can be considered to solve this problem.

1. Normalize the session time
2. Extract the characteristic using window function
3. Develop another method to determine the frequency range which contain characteristics.

The improvement of our proposal method using above techniques is also our future works.

A recent proposal method for detecting unknown attack is shown in[6]. The method has 75.7% of successful detection and 10.7% of false positive for unknown attack session. Note that these results are evaluated using Kyoto2006+ dataset. We confirmed that our method is effective and advantageous comparing with existing methods.

VI. CONCLUSION

In this paper, we propose a new signature generation scheme using Discrete Fourier Transform. As shown in Section IV, we found that our method has 5% of false positive by computer experiments. From the results, we conclude that our proposal method is effective to detect attack session in the complex communication situation. On the other hand, we find that our method is not so effective in the simple situation. In Section IV.E, we discussed the solution against this problem and we take false positive strategy. As the result, we evaluate 5% of false positive. In section V, we discussed the validity of evaluation and method for improvement. From experimental results and discussions, we consider that our method using Discrete Fourier Transform can be improved to be more effective and general scheme.

REFERENCES

[1] A. V. Aho and M. J. Corasick, "Efficient string matching: An aid to bibliographic search" Comm.of the ACM, vol.18(6), 1975, pp.333-340.

[2] P. Barford, J. Kline, D. Plonka, and A. Ron, "A signal analysis of network traffic anomalies" Internet Measurement Workshop, Marseille, pp.71-82, November 2002.

[3] B. Commentz-Walter, "A string matching algorithm fast on the average" Proc. of ICALP, 1979, pp.118-132.

[4] S. Kim, A. L. N. Reddy, and M. Vannucci, "Detecting Traffic Anomalies through Aggregate Analysis of Packet Header Data" Networking 2004, LNCS 3042, 2004, pp.1057-1059.

[5] C.Kreibich and J.Crowcroft "Honeycomb: Creating intrusion detection signatures using honeypots" ACM SIGCOMM Computer Communication Review, vol.34(1), 2004, pp.51-56.

[6] M.Sato, Y.Yamaguchi, H.Shimada, H.Takakura "Detecting anomalous traffic patterns by focusing on sequence of session data" SCIS, 2014 (in Japanese)

[7] S.Singh, C.Estan, G.Varghese and S.Savege "The EarlyBird system for Real-Time Detection of Unknown worms" Tech. Rep. CS2003-0745, CSE Department, UCSD, 2003.

[8] J.Song, H.Takakura, Y.Okabe, M.Eto, D.Inoue and K.Nakao "Statistical Analysis of Honeypot Data and Building of Kyoto 2006+ Dataset for NIDS Evaluation" Workshop on development of large scale security-related data collection and analysis initiatives (BADGERS 2011), ACM, pp.29-36.

[9] A. Valdes and K. Skinner, "Adaptive Model-based Monitoring for Cyber Attack Detection" RAID 2000, LNCS 1997, 2000, pp. 80-93.

[10] S. Wu and U. Manber, "A fast algorithm for multi-pattern searching" Tech. Rep. TR-94-17, Dept. of Comp. Science, Univ of Arizona, 1994.

[11] http://www.takakura.com/Kyoto_data/