Network Attack Traffic Detection for Calculating Correlation of the Flow

Jee-Tae Park, Young-Hoon Goo, Kyu-Seok Shim, Ui-Jun Baek, Myung-Sup Kim
Dept. of Computer and Information Science, Korea University
Korea
{pjj5846, gyh0808, kusuk007, pb1069, tmskim}@korea.ac.kr

***Abstract*— As the propagation of high-speed Internet and the rapid development of the network environment have led to an increase of various types of attack traffic. To cope with the various types of attack traffic, it is essential to detect these traffic accurately. There are many ways to detect traffic, and the most common methods are signature-based analysis and machine learning-based analysis. Both methods have the advantage of being able to detect with high accuracy, but both methods have several limitations in processing. In this paper, we propose a classification method with sequential grouping based on correlation of the flow. The proposed method is a method of calculating the correlation of two flows with the attack flow information, and then detecting the related attack flow based on that. As a result of applying the proposed method to real attack traffic, we could detect with high accuracy.**

Keywords—Seed, Guideline, Flow Correlation Index, Attack Traffic Detection, Flow Correlation

I. INTRODUCTION

An emergence of high speed Internet and rapid development of today's network environment have led to network traffic more diverse and complex. On this trends, various types of attack traffic are emerged, and the damage amount of attack traffic is also increasing rapidly. In order to prevent from these attack traffic, an precise analysis and detection method for the attack traffic is needed[1]. Various classification methods are being studied to detect attack traffic, and the most commonly used methods are signature-based classification and machine learning-based classification[5-7].

Signature based detection method usually use automatic signature generation system on recently. It extracts signatures of traffic automatically and detect attack traffic by using these signatures[4-6]. There are two methods to use header signatures and payload signatures as signature based classification method. First, the header signature based method use the IP addresses and the port number of server[2]. However it is not reliable because of using dynamic and changeable port numbers. Second, the payload signature based method results highest performance on accuracy. However there are also several problems in the method. It cannot react quickly to application changes. It also needs to extract signature, but extraction process is complicated and requires a lot of time.

The other commonly used detection method is machine learning based classification. Machine learning has been used in many fields as a well-known method in recent years. It is easier and more convenient than other classification methods[3,8-10]. It has also high performance on accuracy. However there are also several problems on the machine learning based classification method. Machine learning method depends on the quality of the training traffic or learning features. Therefore this method needs to select appropriate training data and learning features for the high performance. It also cannot classify the unknown application traffic. It is necessary for additional learning for the unknown traffic.

For these reasons, we use the classification method with sequential grouping based on correlation of the flow. This method calculates the correlation with the flow in unlabeled traffic set based on the information of attack traffic flow. This value is defined as Flow Correlation Index(FCI). Based on the FCI, we detect attack traffic and its related traffic by grouping them sequentially.

FCI consists of similarity index and connectivity index. Similarity index is calculated with the statistical information of packets in flow. Connectivity index is calculated with the header information of flow. In this previous method, both indexes were used to calculate a FCI. However, it is hard to calculate similarity index because it needs lots of time and cost to get statistical information of packets in flow. As this reason, it is hard to apply previous FCI system.

Therefore, we propose a method of advanced FCI system to solve this problem of previous FCI system. This method use only connectivity index to calculate the FCI value. The proposed method has an advantage that it can detect the attack traffic by using the minimum attack traffic information without complicated procedure and statistical traffic information.

In this paper, we define the correlation of network flow and propose the method of sequential grouping based on the flow correlation. The remainder of this paper is organized as follows. In Section 2, we explain a definition of three keywords of the system and the entire system. In Section 3, we will explain several algorithm of the proposed system. We perform two evaluation experiments to verify the effectiveness of the proposed method in Section 4. Finally, we conclude and remark the future work in Section 5.

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No. 2017-0-00513, Developing threat analysis and response technology based on Security Analytics for Heterogeneous security solution) and This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea Government(MSIT) (No.2018-0-00539-001,Development of Blockchain Transaction Monitoring and Analysis Technology)

II. PROPOSED METHOD

A. Seed Information

In this paper, we define a seed information. Seed information indicates a minimal information of attack flow. Seed information can be obtained from IDS or IPS of external firewall. Seed information can be defined differently for the people. However, it should contain an essential seed information which indicates 5-tuples of flow. Seed flow is a flow corresponding to the seed information. Seed group indicates a group of seed flows corresponding to the single seed information. Grouping detection is performed through sequential grouping from the seed group.

B. Flow Correlation Index (FCI)

We define a numerical value of flow correlation. As we mentioned before, connectivity index is only used in the system. This connectivity index is numerically calculated with the header information of two flow. By using the FCI value, we set a threshold which used in sequential grouping. It is import to set an elaborate threshold to get a high performance of results.

Connectivity index features originally consists of 4 features (Start Time, IP Address, Port number Protocol). However, if only these 4 features are used, the performance of detection will be reduced because similarity index is not used. Therefore we add 2 more features (Ratio of Forward Packets and Bytes) to calculate the connectivity index. Each of features is shown in Table 1.

Table 1. Connectivity Index Features Description

Feature	Explanation	Value Range
ST	Start Time	0~1
IP	Source & Destination IP Address	0~1
PT	Source & Destination Port Number	0~1
PR	L4 Protocol	1 or mean
RP	Ratio of Forward Packet	0~1
RB	Ratio of Forward Byte	0~1

Table 2. Connectivity Index Feature Function

Feature	Function
ST	$f_{ST}(f_x, f_y) = 1 - \frac{\lvert(ST_x - ST_y)\rvert}{MAX\_INTERVAL\_TIME}$
IP	$f_{IP}(f_x, f_y) = \frac{\left(\left(\frac{PF\left(src_{P_{X,}} src_{IP_y}\right)}{32}\right)^2 + \left(\frac{PF\left(Dst_{IP_{X,}} Dst_{IP_y}\right)}{32}\right)^2\right)}{2}$
PT	$f_{IP}(f_x, f_y) = \frac{\left(\left(\frac{PF\left(src_{PT_{X,}} src_{PT_y}\right)}{16}\right)^2 + \left(\frac{PF\left(Dst_{PT_{X,}} Dst_{PT_y}\right)}{16}\right)^2\right)}{2}$
PR	$f_{Pr}(f_x, f_y) = \begin{cases} \textbf{mean}: f_x.PROT \neq f_y.PROT \\ \textbf{1}: f_x.PROT = f_y.PROT \end{cases}$
RP	$1 - $ The Difference of Forward Packet Ratio in Flow $(f_x, f_y$
RB	$1 - $ The Difference of Forward Byte Ratio in Flow (f_x, f_y)

Each connectivity index features function is shown in Table 2. Start Time indicates the similarity of each flows occurrence time. This value is subtracted a difference between the two flow occurrence times from 1. IP indicates the similarity of each flows IP address and it is calculated with the prefix of source and destination IP addresses. PT indicates the similarity of each flows port number and the calculation of port number is same as IP address. PR checks each flow of protocol. If the protocols are same, it indicates 1 and if not, it is set to the average of the other features. RP and RB indicates the ratio of forward packets and bytes. RP is subtracted a difference between two flow of forward packet ratio from 1. RB is calculated same as RP.

$$\begin{aligned} \mathrm{CI}(f_x, f_y) = &\left(w_{ST} \times f_{ST}(f_x, f_y)\right) + \left(w_{IP} \times f_{IP}(f_x, f_y)\right) \\ &+ \left(w_{Pt} \times f_{Pt}(f_x, f_y)\right) + \left(w_{Pr} \times f_{Pr}(f_x, f_y)\right) \\ &+ \left(w_{RP} \times f_{RP}(f_x, f_y)\right) + \left(w_{RB} \times f_{RB}(f_x, f_y)\right) \\ &(\text{where}, \textstyle\sum_{i=1}^{6} w_i = 1) \quad (1) \end{aligned}$$

Weights are set to 0.05~ 0.7 and the sum of weights is 1. We consider all cases of weights and calculates connectivity index by using weights and feature values. Each features are multiplied with each features of weights. Connectivity index is sum of these calculated values as shown in Eq (1).

C. Guideline (GL)

In this paper, we define a Guideline which can help for grouping and detecting attack flow. Guideline is a excel file which contains thresholds and weights in grouping sequence. Grouping sequence means the number of grouping number. An elaborate guideline enables to make more precise and improved detection results. Therefore, it is important to set a sophisticated threshold and make a well-designed guideline.

D. Entire System

The entire system of this method is shown in Figure 1. It consists of two modules (Guideline Generation & Sequential Grouping).

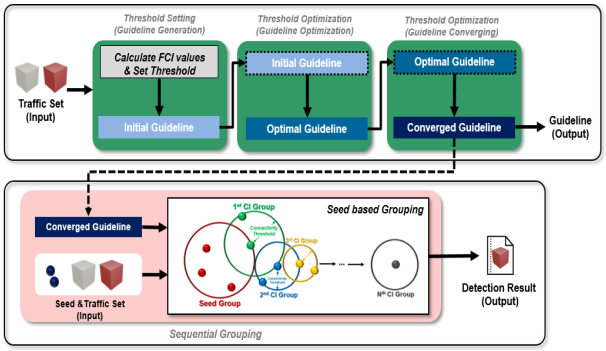

Figure 1. Entire System of Grouping Detection

First a guideline is generated in Guideline Generation module. Before the detection, we should define a guideline which is the criteria of sequential grouping detection. The guideline consists of threshold and weights. Threshold is set through using a Threshold Setting Algorithm. When the guideline is generated, its threshold and weights are optimized through using a Threshold Optimization Algorithm. Finally, optimal threshold and weights are converged through using a Threshold Converging Algorithm.

Second, a grouping detection performed in Sequential Grouping module. Based on the previously created seed group, the grouping of the flows related to the attack flow is performed continuously. The FCI value between the two flows is compared with the threshold value of converged guideline. Then the grouping is performed if it is larger than

the threshold value, and if not, the grouping is not performed. Therefore, it is important to set an elaborate and precise threshold to get a high accuracy results.

Grouping is performed until grouping is no longer in progress. When grouping is complete, all attack flow groups that have been grouped before are analyzed and the detection results are derived. A detection result contains 3 evaluation indicators (Recall, Precision, F-measure). These 3 evaluation indicators are explained in detail in Section Ⅳ. Figure 2 shows entire system of proposed method with input and output.

III. Algorithm of the Proposed System

A. Threshold Setting Algorithm

As we mentioned before, it is important to get a high performance of detection. Therefore we set an algorithm to set an appropriate threshold to detect as shown in Figure 2 and Figure 3. In Figure 2 and 3, a red circle indicates a noise flow a blue circle indicates an attack flow. Each of flows are calculated with the correlation of seed flow. There are two ranges of connectivity index. Threshold can be set any value of these range.

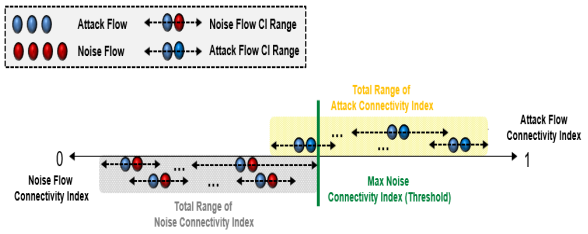

Figure 2. Threshold Setting Algorithm

However, we set a threshold as a value of max noise connectivity index in total range of noise connectivity index. Because, the grouping precedes if connectivity index of flow is larger than threshold. If we set a threshold as max noise connectivity index, we can guarantee 100% precision of the results.

B. Threshold Optimization & Converging Algorithm

A guideline is generated from a seed flow. In other words, there are many guidelines are generated in one attack traffic. Each of these guidelines are called an initial guideline. When proceeding with grouping detection, we apply only one guideline. Therefore, initial guidelines generated as many as the seed flows should be made as one optimal guideline.

$$GL_{i_{TH_{Rating}}} = G{:}Flows_A \times Precision = G{:}Flows_A \times \frac{G{:}Flows_A}{G{:}Flows_A + G{:}Flows_N}$$

$$G{:}Flows_A = Grouped\ Attack\ GT\ Flows\ \ G{:}Flows_N = Grouped\ Noise\ GT\ Flows \quad (2)$$

$$GL_i\_TH_{Rated} = GL_i\_TH\_Rating \times GL_i\_Threshold \quad (3)$$

$$GL_i\_TH_{Optimal} = \frac{\sum_{i=1}^{Numbers\ of\ GL} GL_i\_TH_{Rated}}{\sum_{i=1}^{Numbers\ of\ GL} GL_i\_TH_{Rating}} \quad (4)$$

To make an optimal guideline, the optimization processing is preceded. Eq (2), (3), (4) shows that an optimization algorithm. In Eq (2), we define a Threshold Rating and Rated Threshold, and Optimal Threshold. First, Threshold Rating is the value multiplied by the number of attacked flows and precision. Precision is the ratio of precisely grouped attack flow. Second, Rated Threshold is the value multiplied by each of thresholds and Rating. Finally, Optimal Threshold is an average of Rated Threshold.

$$GL_{Converged_{TH}} = \frac{\sum_{i=1}^{N_t} GL_i\_TH_{Optimal}}{N_t} \quad (N_t = Numbers\ of\ Traffic\ Traces) \quad (5)$$

Then, one optimal guideline is generated from one type of attack traffic. However, there are many types of attack traffic in the network and the values of each threshold are different. In order to cover these different thresholds, each of optimal guideline should be converged one guideline. As shown in Eq (5), converged threshold is derived from an average of all. optimal thresholds.

C. Multiple Guideline Algorithm

There are many types of attack traffic in the network. Although we use a converged guideline to detect attack traffic, it is hard to cover all types of attack traffic. Because each of them has different flow characteristic. Therefore, we apply the Multiple Guideline Algorithm to detect more elaborate.

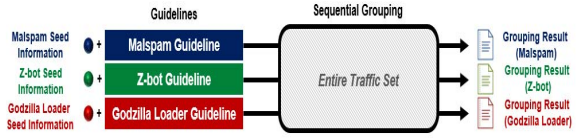

Figure 3. Process of Multiple Guideline

In Figure 3, we make each types of converged guideline before and applying them to corresponding traffic set. For example, if you first create a guideline for Malspam, then that guideline is only applying to related Malspam traffic. This algorithm can cover the limitation of detecting diversity attack traffic.

D. Multiple Seed Algorithm

It is difficult to detect all related attack traffic with only one attack flow information. Therefore we proposed a method of multiple seed algorithm. This algorithm complements the limit of single seed to improve the overall detection results. If we use 2~5 seed information, we can detect all of related attack flow of each information.

However, using multiple seeds can results in relatively long time and overhead than using a single seed. Because there are many combinations of multiple seeds and it will spend a lot of time. Therefore, we first determine which seeds combination should be used by checking a single seed result. After that, we use the combination of seeds which can results the best performance on detection.

IV. Evalution

In this paper, we conducted several experiments to verify the proposed system. As shown in Table 3, we used 4 types of attack traffic (Malspam, Godzilla-Loader, Z-bot, Ransomware) and noise traffic (Skype, Internet Web browser traffic or other application traffic) on experiments.

The evaluation measurement consists of 3 evaluation indicators (Recall, Precision, F-measure). Recall represents the ratio of detected flows in the whole attack traffic flow.

Precision represents the ratio of precisely detected attack flow among the detected flow. F-measure is a numerical value for objectively evaluating the recall and precision. F-measure MS indicates an F-measure which use multiple seeds on experiments.

Table 3. Experiment Traffic Information

Attack Traffic Information				
Trace #	**Attack Method**	**Size**		
		Flow	**Packet**	**Byte**
1	Malspam	71	18,055	15,167,725
2	Godzilla-Loader	69	1,862	1,358,410
3	Z-bot	23	1,232	1,229,269
4	Ransomware	3259	4246	1,169,285
Noise (Normal) Traffic Information				
Trace #	**Application**	**Size**		
		Flow	**Packet**	**Byte**
1	Chrome, IE web, Skype ..	1067	167,800	142,488,591

We conduct two evaluation experiments by using advanced FCI system and SnorGen (Automatic Signature Generation System). As we mentioned before, although it has complicated process to use, signature based detection method results high performance. Therefore, we used signature based detection method to compare and verify the validation of advanced FCI system.

Table 4. Result of Detection Experiment

Detection Test Result									
Input Traffic		Measurement	Coverage (%)						
			Advanced FCI system			SnorGen			
Attack	Noise		Flow	Pkt	Byte	Flow	Pkt	Byte	
Malspam	all	Recall (%)	67.9	99.9	99.9	100	100	100	
		Precision (%)	100	100	100	78.5	89.1	90.5	
		F-measure(1)	97.8			87.9			
		F-measure MS (2)	100			-			
Godzilla-Loader	all	Recall (%)	92.8	97.0	99.6	94.2	96.1	95.8	
		Precision (%)	100	100	100	90.5	94.5	96.6	
		F-measure(1)	96.24			94.3			
		F-measure MS (2)	100			-			
Z-bot	all	Recall (%)	82.9	99.4	99.9	85.7	92.2	95.7	
		Precision (%)	100	100	100	100	100	100	
		F-measure(1)	90.48			92.2			
		F-measure MS (2)	100			-			
Ransomware	all	Recall (%)	38.9	28.7	15.5	100	100	100	
		Precision (%)	100	100	100	100	100	100	
		F-measure(1)	56			100			
		F-measure MS (5)	100			-			

Results of the experiments is shown in Table 4. It indicates that a detection rate and precision of this experiments. A precision of detection is 100% in all types of attack traffic. However, a detection rate (Recall) is 30~95% on average when we use a single seed. Comparing with SnorGen, the recall of advanced FCI system was relative low when using a single seed. In ransomware, there was a big difference in recall between advanced FCI system and SnorGen. However, when we use a multiple seeds algorithm, the performance of advanced FCI system is higher than SnorGen in all of the traces. Especially, when we use 5 seeds in advanced FCI system, recall and precision results 100% like SnorGen results.

V. Conclusion and Future Work

In this paper, we proposed a method of attack traffic detection based on flow correlation. To verify our proposed method, we conducted an experiment by using 4 types of attack traffic comparing with the previous method. Most of experiments result showed high performance when advanced FCI system applied. When using multiple seeds in advanced FCI system, it has higher performance than SnorGen results.

However, it needs to apply more diverse attack traffic and conduct more experiments to prove validity of proposed system. Therefore, we will conduct additional experiments on other types of attack traffic and set more efficient threshold setting algorithm to enhance the system for the future work.

References

[1] M.-S. Kim, Y. J. Won, and J. W.-K. Hong, "Application-level traffic monitoring and an analysis on IP networks," ETRI journal, vol. 27, pp. 22-42, 2005.

[2] S. H. Yoon, J. S. Park, Baraka D. Sija, M. J. Choi, and M. S. Kim, "Header Signature Maintenance for Internet Traffic Identification." International Journal of Network Management, vol.27, No.1, Jan. 2017, pp. 1-15

[3] S. H. Lee, and M. S. Kim, "Application Traffic Classification using TensorFlow Machine Learning Tool.", In Proc KICS 2016, pp.224-225, ChungAng Univ, Korea, Nov. 2009.

[4] K. S. Shim, S. H. Yoon, S. K. Lee, and M. S. Kim, "SigBox: Automatic Signature Generation Method for Fine-grained Traffic Identification", Journal of Information Sciecnce and Engineering Vol. 33, No. 2, Feb. 2017, pp.541-573

[5] . N. F. Huang, G. Y. Jai, H. C. Chao, Y. J. Tzang, and H. Y. Chang, "Application traffic classification at the early stage by characterizing application rounds," Information Sciences, Vol. 232, 2013, pp. 130-142.

[6] CA. Catania and CG. Garino, "Automatic network intrusion detection: Current techniques and open issues," Computers and Electrical Engineering, Vol. 38, Issue. 5, Sep. 2012, pp. 1062-1072

[7] F. Risso, M. Baldi, O. Morandi, A. Baldini and P. Monclus, "Lightweight, payload-based traffic classification: An experimental evaluation", In Proceedings of IEEE International Conference on Communications ICC, May. 2008, pp.5869

[8] T. Nguyen, G. Armitage, " A Survey of Techniques for Internet Traffic Classification using Machine Learning," IEEE Communications Surveys and Tutorials, to appear, 2008.

[9] S. Zander, T. Nguyen, and G. Armitage, "Automated Traffic Classification and Application Identification Using Machine Learning," Proc. IEEE Ann. Conf. Local Computer Networks, pp. 250-257, 2005

[10] R. Sommer and V. Paxson. Outside the closed world: On using machine learning for network intrusion detection. In Proc. of IEEE Symposium on Security and Privacy, pages 305–316, 2010