A LOW COST WORM DETECTION TECHNIQUE BASED ON FLOW PAYLOAD SIMILARITY

Youhei Suzuki, Yuji Waizumi, Hiroshi Tsunoda and Yoshiaki Nemoto
Graduate School of Information Sciences, Tohoku University
6-6-05, Aramaki-Aza-Aoba, Aobaku, Sendai-shi, Miyagi, 980-8579, Japan

Keywords: Worm, Similarity of Flow Payloads, Clustering, Intrusion Detection.

Abstract: Recently, damages of information systems by worms have been reported at global level. Signature based Intrusion Detection Systems (IDSs) are widley used to prevent these damages. To handle newly created worms, automatic signature generation techniques based on common strings in the payloads of multiple worm flows of the same kind have been proposed. Because these techniques need to use multiple strings as a signature for each kind of worm to acheive high detection accuracy, the calculation cost to detect worms is a serious issue. In this paper, we propose a novel scheme that does not use common character strings. The proposed scheme uses a 256-dimensional vector based on the appearance frequencies of 256 character codes. This vector is generated automatically and used as a mean to detect worms with low cost. In addition, we construct a cheap worm detection system by using the proposed method as the first stage analysis of conventional IDS. We evaluate the proposed scheme through experiments and present its performance.

1 INTRODUCTION

Internet worms are one of the most serious threats in the Internet. With improvement in the speed of networks and computers, the diffusion speed of worms is also increasing vigorously. Worms are responsible for a large bulk of damages caused to information systems (Yaneza et al., 2005). In order to control the damage by these worms, highly accurate Intrusion Detection Systems (IDSs) need to be implemented. Most IDSs like Snort (Snort, 1998) adopt signature matching techniques to detect worms. Although this approach can detect known worms with high accuracy, the signature matching process is computationally expensive. Since a new signature has to be added to the signature database of the IDSs in order to detect a new kind of worm and its subspecies, the computational cost for searching signatures constantly gets increased. This can be a serious problem as new kinds of worms are created every day.

Same kinds of worms carry similar payloads (a set of the payload of all the packets contained in a flow) reported in (Akritidis et al., 2005). Because a worm may spread by its own copy to other hosts at the time of diffusion, payloads of flows transmitted from the same kind of worms can have high similarity. Based on this fact, systems to generate signatures automatically from common strings in the payloads of multiple worm flows are proposed in (Kim and Karp, 2004) (Simkhada et al., 2005) (Newsome et al., 2005) (Singh et al., 2004) (Wang et al., 2005). Although these systems can shorten the time to generate signatures, they can not reduce the detection cost because they use multiple strings as signatures for each kind of worm. The calculation cost to detect worms remains a serious issue for network security.

2 BACKGROUND

Most signature based IDSs detect worms by matching the signature strings to worm payloads. Since most worms can be detected by using only one string, of the computational cost of these IDSs to detect a worm flow is $O(LN)$, where L and N denote the lengths of payloads of a worm flow and the number of strings of a signature, respectively. To reduce the signature generation time, automatic signature generation tech-

Suzuki Y., Waizumi Y., Tsunoda H. and Nemoto Y. (2007).
A LOW COST WORM DETECTION TECHNIQUE BASED ON FLOW PAYLOAD SIMILARITY.
In *Proceedings of the Third International Conference on Web Information Systems and Technologies - Internet Technology*, pages 414-417
DOI: 10.5220/0001279704140417

niques based on similarities of worm flows are proposed in (Akritidis et al., 2005),(Singh et al., 2004). Because these approach need two or more common signature strings to achieve a high detection accuracy, the computational cost gets significantly high.

(Singh et al., 2004)(Wang et al., 2005)(Kruegel et al., 2002)(Tsuji et al., 2005) have shown that it is possible to evaluate similarities between flow payloads in terms of a 256-dimensional vector based on histograms of the appearance probabilities of 256 byte codes. We call this vector $\vec{h}$ vector and express it as

$$\vec{h} = (h_0, h_1, h_2, \cdots, h_{255}) \quad (1)$$

where h_i is the appearance probability of code i. The vector exhibits the feature of whole flow payloads of a flow, and is of fixed length 256. Consequently, by using the $\vec{h}$ vector as the signature, we can detect worms with a lower calculation cost. The computation cost to extract $\vec{h}$ vector from a flow payload is proportional to the length of the flow, i.e. $O(L)$. The cost of evaluating the similarity between flows by using $\vec{h}$ vectors is constant, which is equal to the dimension of $\vec{h}$ vector. Thus, the total calculation cost for detecting worms by methods using $\vec{h}$ vectors is $O(L+256C_i)$, where C_i is the average number of signatures.

In this paper, we use $\vec{h}$ vectors to reduce the calculation cost for detection, and build a low cost worm detection system which consists of two detection stages. The first stage uses $\vec{h}$ vectors to detect worms and reduce the number of flows required to be analyzed during the second stage. The second stage adopts common string signatures to detect worms whose signature $\vec{h}$ vectors do not exist in the signature database.

3 WORM DETECTION USING $\vec{H}$ VECTORS

In this section we use $\vec{h}$ vector for worm detection and investigate its performance. We adopt a clustering technique to extract these worms. Similar worm flows are clustered in the 256-dimensional space. The average position of the flows in each cluster as used as the $\vec{h}$ vector of the corresponding cluster. (Waizumi et al., 2005) reports that $\vec{h}$ vectors of a kind of worm can be present in multiple clusters. Signature $\vec{h}$ vectors are then calculated. Worms are detected by using these signature vectors.

3.1 Signature Vector Generation

Let, $\vec{h}_{i,j}$ denote the $\vec{h}$ vector of flow j of a kind of worm i. And $\vec{m}_{i,c}$ represent the $\vec{h}$ vector of worm cluster c. The clustering algorithm is shown as follows:

Begin
$m_{i,1} \leftarrow h_{i,1}$
do $j \leftarrow j+1$
$w \leftarrow \arg\min_{c'}(D(\vec{h}_{i,j}, \vec{m}_{i,c'}))$
if $D(\vec{h_{i,j}} - \vec{m_{i,w}}) < \theta_1$
then $\vec{m_{i,w}} \leftarrow \vec{m_{i,w}} \cdot (n_{i,w}-1)/|w| + \vec{h_{i,j}}/|w|$
else $\vec{m_{i,c_i}} \leftarrow \vec{h_{i,j}}$
$c_i \leftarrow c_i + 1$
until $\vec{h_{i,j}}$ == NULL
end

where, $|w|$ is a number of elements included in cluster w, and c_i is number of clusters. Moreover, the distance $D(\vec{h_{i,j}}, \vec{m_{i,c}})$ between $\vec{h_{i,j}}$ and $\vec{m_{i,c}}$ is calculated as,

$$D(\vec{h_{i,j}}, \vec{m_{i,c}}) = \sum_{k=0}^{255} (h_{i,j,k} - m_{i,c,k})^2 \quad (2)$$

where, $h_{i,j,k}$ and $m_{i,c,k}$ are the elements of the k^{th} dimension of $\vec{h_{i,j}}$ and $\vec{m_{i,c}}$.

Two or more clusters with radius is θ_1 are generated by this clustering algorithm. Flows whose $\vec{h}$ vectors are far from each other are clustered into different clusters. If the same kind of worm has multiple $\vec{h}$ vectors, multiple signature vectors are generated for the worm.

3.2 Detection By Signature Vectors

Observed flows which are significantly near the signature vector are detected as worms. The criterion of detecting worm is defined by a threshold distance θ_2. If a newly observed flow is less than θ_2 from a signature vector, the flow is detected as a worm flow.

3.3 Performance Evaluation

3.3.1 Experimental Environment

In this experiment, we use an off-line real network traffic containing worm flows. By using a signature provided by Bleeding threats (Bleeding Edge Threats, 2004), Bagle, MyDoom and Netsky.P worm flows are extracted from the traffic and are used as test flows. In the same way, about 13,000 normal flows are extracted from the traffic in one day and are used for evaluating false alarms.

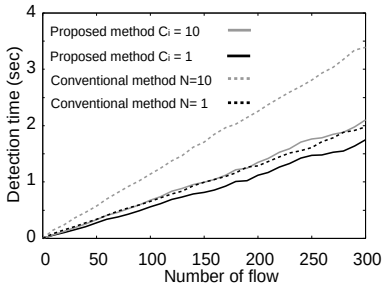

Figure 1: Comparison the detection time of the proposed method and the document (Simkhada et al., 2005).

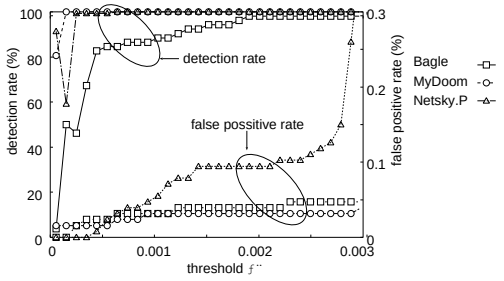

Figure 2: Relation between detection rate and false positive rate by threshold θ change.

3.3.2 Evaluation of Detection Time

We compare the detection time of the proposed method with that of a conventional method which uses common strings as signatures (Simkhada et al., 2005) by using Netsky.P worm flows. Signature vectors and signature strings are generated from 465 flows. 314 flows are used to investigate the efficiency of both methods. The number of signature vectors C_i of the proposed method is set to one and ten by adjusting threshold θ_1.

From Figure 1, it is clear that the proposed method can detect worms with lower calculation cost than existing method. The rate of increase in the detection time in the proposed method is also less compared to the conventional method.

3.3.3 Evaluation of Detection Accuracy

We evaluate the accuracy of proposed technique in detecting worms. In this evaluation, we use the network traffic of two months. The signature vectors are generated from the traffic of the first month and worm flows (Bagle 91 flows, MyDoom 73 flows and Netsky.P 465 flows). Worm flows from a separate database (Bagle 52 flows, MyDoom 62 flows and Netsky.P 314 flows) are used to evaluate the detection accuracy. At the same time, the numbers of false alarms are also evaluated using about 13,000 normal flows.

Figure 2 depicts the detection rate and the false positive rate when threshold θ_1 (to generate signature vectors) and threshold θ_2 (for detection) are set to a same value, θ. Figure 2 shows it is possible to achieve a 100% detection rate with a low false positive rate. Table 1 shows the highest detection accuracy for each worm when threshold θ_1 and θ_2 vary independently. The expression (3) shows the system sensitivity used in the work (Simkhada et al., 2005). The closer the value of S is to 1, better the system sensitivity. From these results we can say that the proposed method can achieve high detection accuracy by selecting an appropriate threshold.

Table 1: Detection accuracy with respect to thresholds θ_1 and θ_2.

Worm	θ_1	θ_2	Det	FP	S	C_i
Bagle	0.002	0.002	98.1%	0.04%	0.977	2
MyDoom	0.001	0.0004	100%	0.02%	0.999	2
Netsky.P	0.001	0.0005	100%	0.00%	1.000	3

$$S = (detection\ rate) \times (100 - false\ positive\ rate)/10000 \quad (3)$$

From Table 1, it is clear that the proposed scheme is capable of achieving a high detection accuracy. Most of the false positive flows were e-mail flows with binary files attached along. Because many elements of $\vec{h}$ vectors extracted from both worm flows and e-mail flows attached binary files tend to be zero, these flows showed similarity and were detected as worm flows.

4 A TWO-STAGE DETECTION SYSTEM TO REDUCE THE CALCULATION COST OF EXISTING IDS

In previous section, we demonstrated that the detection technique using $\vec{h}$ vectors can discriminate worm flows from non-worm flows with lower calculation cost than conventional methods. However, in order to calculate $\vec{h}$ of worm flows and generate signature vectors, some sample flows of worms are necessary. In this section, we propose a two-stage worm detection system which consists of the proposed method in Section 3 as the first stage and the signature string based detection method as the second stage. The second stage sends sample flows to the first stage in order to generate signature vectors (Figure 3).

Figure 3 depicts the components of the proposed two-stage worm detection system. In the proposed

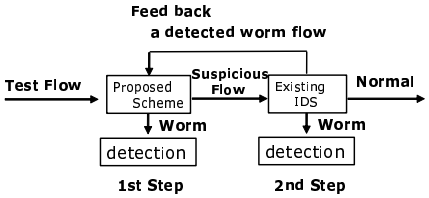

Figure 3: A two-stage worm detection system.

system, the first stage detects worms by using signature vectors at a low calculation cost. The remaining flows are sent to the second stage as suspicious flows. In the second stage, worms are detected from the suspicious flows. Because the number of flows analyzed during the second stage is significantly reduced by the first stage analysis, the total calculation cost of the proposed system is lower than that of the conventional detection system. At least one sample of worm flow is required for the proposed technique to generate a signature vector. Worms detected in the second stage are used as sample flows. The system conducts the process only when the number of the same kind of sample flows exceeds a constant number or if fixed time passes from the last process in order to reduce the signature vector generating cost.

In the same environment as Section 3.3, we evaluated the detection performance of two-stage system using Netsky.P worm. The number of signature vectors finally generated was three. Moreover, when the number of sample flows was 40 or more, all 314 flows could be detected in the first stage. Consequently, the analysis of the 314 Netsky.P flows by the second stage would not be conducted, the calculation cost could be reduced by the proposed system.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a time efficient and low cost worm detection system. The proposed worm detection method evaluates flow similarity by a vector based on the appearance probability of the byte code of flow payloads. The evaluation experiment showed that the method achieves a high detection accuracy while significantly reducing the calculation cost during detection. We also proposed the worm detection system which uses the above-mentioned method as the first stage and existing IDS as the second stage. Through evaluation experiment, we showed that the proposed system is a highly accurate and a low-cost worm detection system.

Future work is to use the proposed method alone to achieve low-cost worm detection. At the same time, a high accuracy is required. A distributed scheme, as introduced by (Staniford et al., 2002), where signatures are shared amongst networks can further enhance the effectiveness of the proposed scheme.

REFERENCES

Akritidis, P., Anagnostakis, K., and Markatos, E. P. (2005). Efficient content-based detection of zero-day worms. In *Proceedings of the International Conference on Communications (I CC 2005)*.

Bleeding Edge Threats (2004). `http://www.bleedingsnort.com`.

Kim, H. and Karp, B. (2004). Autograph: toward automated, distributed worm signature detec tion. In *Proceedings of the 13th USENIX Security Symposium*.

Kruegel, C., Toth, T., and Kirda, E. (2002). Service specific anomaly detection for network intrusion dete ction. In *Symposium on Applied Computing (SAC)*.

Newsome, J., James, B., Karp, B., and Song, D. (2005). Polygraph: Automatically generating signatures for polymorphic worms. In *Proceedings of the 2005 IEEE Symposium on Security and Pri vacy*. IEEE Computer Society.

Simkhada, K., Tsunoda, H., Waizumi, Y., and Nemoto, Y. (2005). Differencing worm flows and normal flows for automatic genera tion of worm signatures. In *Proceedings of the Seventh IEEE International Symposium on Mu ltimedia (ISM)*.

Singh, S., Estan, C., Varghese, G., and Savage, S. (2004). Automated worm fingerprinting. In *Proceedings of the 6th ACM/USENIX Symposium on Operating System Design and Implementation (OSDI)*.

Snort (1998). `http://www.snort.org`.

Staniford, S., Paxson, V., and Weaver, N. (2002). How to 0wn the Internet in your spare time. In *Proceedings of the 11th USENIX Security Symposium*.

Tsuji, M., Waizumi, Y., Tsunoda, H., and Nemoto, Y. (2005). Detecting worms based on similarity of flow payloads. In *IEICE Tech. Rep. NS2005-112*, pages 9–12.

Waizumi, Y., Tsuji, M., and Nemoto, Y. (2005). A detection technique of epidemic worms using clustering of p acket payload. In *IEICE Tech. Rep. CS2005-19*, pages 19–24.

Wang, K., Cretu, G., and Stolfo, S. (2005). Anomalous payload-based worm detection and signature generation. In *Proceedings of the Eighth International Symposium on Recent Adva nces in Intrusion Detection*.

Yaneza, J. L. A., Mantes, C., and Avena, E. (2005). *The Trend of Malware Today: Annual Virus Round-up and 2005 Forecast*. Trend Micro.