Payload Signature Structure for Accurate Application Traffic Classification

Young-Hoon Goo, Kyu-Seok Shim, Su-Kang Lee, and Myung-Sup Kim
Dept. of Computer and Information Science
Korea University
Sejong, Korea
{gyh0808, kusuk007, sukanglee, tmskim}@korea.ac.kr

***Abstract*— Emergence of high-speed Internet and various smart devices has led to a rapid increase of applications on the Internet. In order to provide reliable services and efficient management of network resources, accurate traffic classification of various applications is essential. Through various methods of extraction when payload signatures are extracted, most of these payload signature formats are just strings or hex values which appear frequently within payloads. Thus, it is difficult to extract unique signatures for a specific application, because redundant signatures extraction is in most cases unavoidable. In this paper, we propose a more elaborative payload signature structure for accurate classification of each specific application. The formats of this signature structure is composed of three level signatures. These are Content signature which is single contiguous substring in payloads, Packet signature which is the sequence of Content signatures that appear in the same packet, and the Flow signature which is a sequence of Packet signatures that appear in the same flow. By applying and comparing the existing signature format and proposed signature format to the actual application traffic classification, we demonstrate the effectiveness of the proposed signature structure.**

Keywords—signature structure; Content signature; Packet signature; Flow signature; completeness; false positives

I. INTRODUCTION

Today's emergence of high-speed Internet and various smart devices, has led to a rapid increase of developed applications on the Internet. As a result, a lot of researches are being done to provide reliable services to users and maximize the utilization of network resources. In order to achieve this, a method that can accurately classify various applications traffic is necessary.

Among various existing traffic classification methods, the highest-performing method in terms of accuracy and completeness is a payload signature-based method [1,2,7,8]. The reason behind is that the method accurately classifies an application from extracted traffic payload information. However, there is a possibility of incorrect classification by signature redundancy when applied to raw traffic of today's increasing various applications.

Most of the payload signature formats in the previous researches are simple substrings which emerge frequently within payloads [3,5,10]. Therefore, still there is a possibility that extracted payload signatures may not be specific to a particular application, some might be of another application, so-called signature redundancy. This lowers the reliability in network management resulting to improper network policies, capacity planning, trouble shooting, etc. Particularly in network security field, false positives and false negatives of malicious traffic can lead to great loss.

As Internet application traffic drastically increase, the process of network managers to manually extract payload information has become time consuming and demands high expertise. Although various research methods for automatic payload signatures extraction to solve this problem are in progress, the payload signatures automatically extracted cannot guarantee whether the signature only can detect its application accurately without the false positives or false negatives. Thus, we need a much more unique structure for particular payload signature formats to clearly distinguish one application from another.

In this paper, we propose a much more elaborative payload signature structure which is specific for classifying a single application. The formats of this signature structure is composed of three level signatures. These are the Content signature which is single contiguous substring in payloads, the Packet signature which is the sequence of Content signatures that appear in the same packet, and the Flow signature which is a sequence of Packet signatures that appear in the same flow. By applying the existing signature format and proposed signature format to the actual application traffic classification, we demonstrate the effectiveness of proposed signature structure by maintaining completeness and reducing the false positives.

The rest of this paper is organized as follows. Section II presents previous researches on payload based traffic classification. Section III presents the limitations of previous researches. Section IV presents the definition and proposed for structure of payload signature formats. Section V presents an experiment to demonstrate the feasibility of the proposed payload signature format by applying to the actual application traffic classification. Finally, Section VI presents conclusive remarks and a brief look for the future research directions.

II. RELATED WORK

Payload signatures based traffic classification is a method for extracting features that are distinguishable and unique patterns of a particular traffic application, then applying it to target networks traffic for classification of particular applications by

inspecting whether the features are inclusive in the payloads or not. Payload signatures extraction methods have been studied in a variety of ways.

Kim et al [3] used a common substring that emerges within payloads as a signature which can detect worms by using COPP (Content-based Payload Partitioning) method. Cheng et al [10] used a common bit sequence that emerges within same offset of payload as a signature by using fixed bit offset method. Mingjiang et al [11] defined a substring of units called Shingle and used a Shingle which satisfy certain threshold as a final common substring. Park et al [5] used longest common substring which emerges in payloads as a signature by using LASER (LCS-based Application Signature ExtRaction) algorithm that is one of the LCS(Longest Common String) algorithm. Newsome et al [4] and Feng et al [6] used a set of common substrings which emerge in payloads under Smith-Waterman Algorithm.

Payload signature formats of the above-mentioned researches are simply common substrings within payloads, which still indicate extraction of redundant signatures belonging to two or more applications. Therefore, to solve such problems we need a much more elaborative payload signature structure which is only specific to a particular application.

III. LIMITATIONS OF PREVIOUS RESEARCH

In this section, we describe the limitations of payload signature formats of previous researches. Payload signature formats of the papers mentioned above, are simply partial strings which are commonly found in payloads, except the Smith-Waterman algorithm. Such payload signature formats cannot guarantee that this signature is only distinct from signature of other applications, even if the signature is extracted from a ground-truth traffic of particular application.

Table 1 and Table 2 are examples of signatures that can be duplicated with a high probability.

TABLE I. STRINGS USED AS METADATA OF THE SPECIFIC PROTOCOLS

	Example
common substring in payloads	HTTP /1.1
	200
	Error 201
	Error 202
	GET /
	Content-Type
	User-Agent
	No-cache
	Referrer

As shown in Table 1 metadata for a particular protocol may be all of the same signatures that are extracted from all applications using the same protocol. For example, a HTTP Request/Response command or HTTP header field found in several packets is likely to be extracted as the payload signature. This signature is not worth for traffic identification because this signature can be extracted from many applications using HTTP protocol.

Also, if a string with dictionary words is extracted as the payload signature, this signature is very likely to be found in the traffic of other applications. Table 2 is the case showing that strings which have dictionary words are extracted as payload signatures.

TABLE II. STRINGS HAD DICTIONARY MEANINGS

	Example
common substring in payloads	Image
	Album
	Data
	Music
	Server
	Video-streaming

For example, the string ‘Album’ can be extracted from the traffic of webcam application and can be extracted from the traffic of music related applications, too. Although the string ‘Music’ can be extracted from music related applications, this string cannot accurately distinguish what kind of particular music application among several music related applications.

Figure 1 shows the payload signature formats extracted under the Smith-Waterman algorithm [4] and LCS(Longest Common String) algorithm [5]. The payload signature format under the Smith-Waterman algorithm is not simply a common substring, but a set of common substrings which appear in two payloads. This payload signature format can increase the uniqueness of the application rather than the payload signatures formats of the papers mentioned above, since they use payloads of two packets as inputs to find a set of common substrings. This signature format can be called the signature of one Packet unit. Hence it is much more elaborative than the payload signatures formats using common substring as mentioned above.

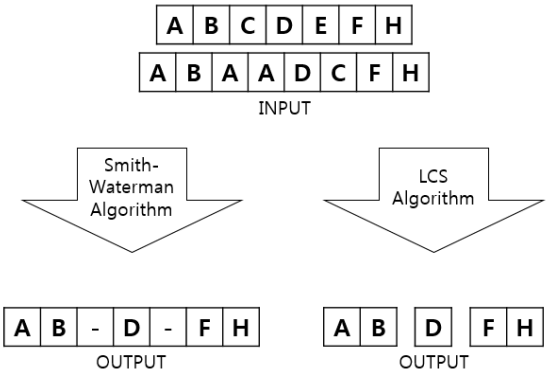

Fig. 1. Payload signature formats of Smith-Waterman and LCS algorithm

However, Smith-Waterman algorithm has big time complexity and computational complexity. Its calculations should be performed as the number of packet of squares in worst case because the algorithm always use payload of only two packets in order to make one set of common substrings. In addition, in order to search for sets of common substrings, more calculations are needed to certain frequencies or thresholds with constant higher values.

In the next section, we describe the payload signature structure which is specific to a particular application and is not redundant to other applications.

IV. Proposed Payload Signature Structre

In this paper, we propose a new payload signature structure consisting of three levels. Figure 2 is extraction process of this structure. The first level part extracts common substrings that satisfy certain frequencies as payload signatures. These are contiguous characters, hex values or combination of them, in traffic payload. The signature in this part is named as 'Content signature'. It is the same as the payload signature formats of the majority of previous researches. The second part extracts a series of Content signatures which satisfy certain frequencies appearing in the same packet. The signature in this part is named as 'Packet signature'. Packet signature increases the accuracy by reducing false positives because it must be matched to a number of Content signatures on one packet. Thus, Packet signature is more specific to a particular application than Content signature. The third part extracts a series of Packet signatures which satisfy certain frequencies appearing in the same flow. This signature in this part is named as 'Flow signature'. Flow signature is much more specific to a particular application than Packet signature and highly increases the accuracy because it must be matched to a number of Packet signatures on one flow.

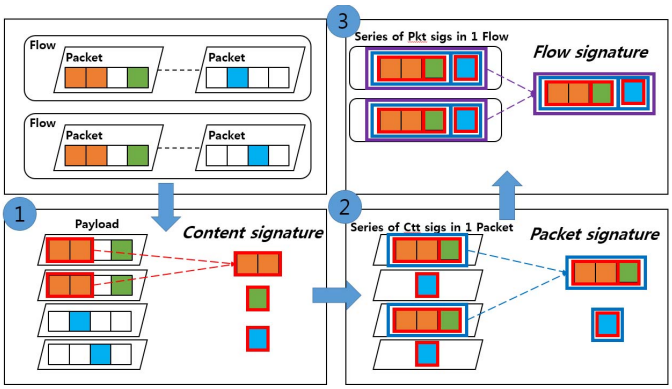

Fig. 2. Extraction process of proposed payload signature structure

Below, the equations are expressions of the proposed payload signature format, whereby C stands for 'Content signature', P for 'Packet signature', and F for 'Flow signature'.

$$C = \{c \mid c \text{ is single substring in a payload}\} \quad (1)$$

$$P = \{p \mid p \text{ is a subset of C; p appears in a packet}\} \quad (2)$$

$$F = \{f \mid f \text{ is a subset of P; f appears in a flow}\} \quad (3)$$

When 9 Content signatures satisfying certain frequencies are extracted, we express this as C = {c1, c2, c3, c4, c5, c6, c7, c8, c9}. Out of these, if 4 Packet signatures are extracted and are in a series of Content signatures in one packet, can satisfy certain frequencies, and we can express it as P = {p1, p2, p3, p4} = { {c1, c3, c9}, {c2, c5, c9}, {c8,c4}, {c6} }. In the same way, if there are 2 Flow signatures extracted and are in a series of Packet signatures in one flow, can satisfy certain frequencies, and we can express it as F = {f1, f2} = { {p1,p3}, {p4} } = { {{c1,c3,c9},{c8,c4}} , {{c6}} }. Table 3, Table 4, and Table 5 are examples of Content signatures, Packet signatures, and Flow signatures, respectively. These are the signatures of eBay which is a multinational e-commerce services via Internet. We extract these signatures by using our automatic signature extraction module of SnorGen. Its payload signature structure also uses additional header and position such as offset and depth information to improve the accuracy. In the following tables, this information is omitted.

TABLE III. Example of Content signatures of eBay

Signature$_{ID}$	Content signature
1	<bay|03|com>
2	<200>
3	<K|0d||0a|Server:>
4	<.ebay>
5	<|01||00||00||01|>
6	<7g64%60%28>
7	<Va TAI>
8	<no-cache>
9	<s|0a|ebaystatic|03|com>

TABLE IV. Example of Packet signatures of eBay

Signature$_{ID}$	Packet signature
1	< <bay|03|com> <K|0d||0a|Server:> < s|0a|ebaystatic|03|com> >
2	< <200> <|01||00||00||01|> < s|0a|ebaystatic|03|com> >
3	< <no-cache> <.ebay> >
4	< <7g64%60%28> >

TABLE V. Example of Flow signatures of eBay

Signature$_{ID}$	Packet signature
1	<<<bay|03|com> <K|0d||0a|Server:> <s|0a|ebaystatic|03|com>> <<7g64%60%28>>>
2	< <<7g64%60%28>> >

From the above tables, although signature ID 6 of Table 3, signature ID 4 of Table 4, and signature ID 2 of Table 5 are of the same string, Packet signature 4 is more elaborative signature than Content signature 6 while Flow signature 2 is the most elaborative signature of the three for identification of a particular application.

V. Experiment and Result

In this Section, by applying and comparing the existing signature format and the proposed signature format to the actual traffic classification, we demonstrate the effectiveness of proposed signature structure.

Experimental environments are as follows. By using two applications Facebook and YouTube, we first generate ground-truth traffic traces. We extract Content signatures and Packet signatures of each two applications from the ground-truth traffic traces. We apply Content signatures and Packet signatures extracted from the ground-truth traffic of one application to the ground-truth traffic of the other application. We then calculate false positives of each application. Completeness experimental environments for the two applications are shown in Table 6, (1) and (4), while (2) and (3) are false positives of the two applications. We finally, analyze the results of false positives and completeness of Content signature and Packet signature. In this experiment, we took an assumption that Content signature is a signature format of previous researches

TABLE VI. EXPERIMENTAL ENVIRONMENT

	Signature of **Facebook**	Signature of **YouTube**
Ground-truth traffic of **Facebook**	(1) TP of Facebook = TN of YouTube	(2) FN of Facebook = FP of YouTube
Ground-truth traffic of **YouTube**	(3) FP of Facebook = FN of YouTube	(4) TN of Facebook = TP of YouTube

Table 7 shows experimental results. All of the values are in flow unit. From the results, we discovered that false positives of Packet signature are lower than those of Content signature and that completeness which was found to have small true positives changes, appeared the same to both applications, Facebook and Youtube.

TABLE VII. EXPERIMENTAL RESULT

	Content Signature		Packet Signature	
	TP	FP	TP	FP
Facebook	6,910/7,530 ***92%***	381/6,315 ***6%***	6,848/7,530 ***91%***	91/6,315 ***1%***
YouTube	5,770/6,315 ***91%***	387/7,530 ***5%***	5,610/6,315 ***89%***	108/7,530 ***1%***

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel payload signature structure for accurate application traffic classification. By using this payload signature structure (Content, Packet, and Flow signatures), we managed to improve the accuracy in identifying a particular application, also managed to prevent redundancy of signatures of applications, and reduced false positives. Therefore, we concluded that it is possible to efficiently manage a network by providing reliable classification results. Through experiments, we validated the effectiveness of proposed payload signature format by comparing to other payload signature format of previous research. As our future work, we will study and consider more methods to correct the idea and applying the proposed method in real-time network management environment so as to correctly identify various applications.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2015R1D1A3A01018057).and KREONET-Emulab Testbed of Korea Institute of Science and Technology Information (K-16-L01-C02-S03).

REFERENCES

[1] F. Risso, M. Baldi, O. Morandi, A. Baldini, and P. Monclus, “Lightweight, Payload-Based Traffic Classification An Experimental Evaluation,” IEEE International Conference on Communications, Beijing, China, May. 19-23, pp. 5869-5875, 2008.

[2] Liu, Hui Feng, Wenfeng Huang, Yongfeng Li, Xing “Accurate Traffic Classification”, Networking, Architecture, and Storage, NAS 2007.

[3] H.-A. Kim and B. Karp, “Autograph: Toward automated, distributed worm signature detection,” in USENIX Security Symp., vol. 286, 2004.

[4] J. Newsome, B. Karp, and D. Song, “Polygraph: Automatically generating signatures for polymorphic worms,” IEEE Symp. Security and Privacy, pp. 226-241, 2005.

[5] B.-C. Park, Y. J. Won, M.-S. Kim, and J. W. Hong, “Towards automated application signature generation for traffic identification,” IEEE Network Operations and Management Symp. (NOMS 2008), pp. 160-167, 2008.

[6] X. Feng, X. Huang, X. Tian, and Y. Ma, “Automatic traffic signature extraction based on Smith-waterman algorithm for traffic classification,” IEEE Int. Conf. Broadband Netw. Multimedia Technol. (IC-BNMT), pp. 154-158, 2010.

[7] J.-S. Park, S.-H. Yoon, J.-W. Park, H.-S. Lee, S.-W. Lee, and M.-S. Kim, "Research on the Performance Improvement of Application-Level Traffic Classification System using Payload Signature," KNOM Review, Vol. 12, No. 2, Dec. 2009, pp. 12-21.

[8] S.-H. Lee, J.-S. Park, S.-H. Yoon, and M.-S. Kim, "High performance payload signature-based Internet traffic classification system ," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2015, Busan, Korea, Aug. 19-21, 2015, pp.491-494.

[9] S.-H. Yoon, J.-S. Park, and M.-S. Kim, "Performance Improvement of a Real-time Traffic Identification System on a Multi-core CPU Environment", The Journal of KICS '12-05 Vol.37B No.5, pp. 348-356, May 2012.

[10] C. MU, X.-h. HUANG, X. TIAN, Y. MA, and J.-l. Qi, “Automatic traffic signature extraction based on fixed bit offset algorithm for traffic classification,” The J. China Universities of Posts and Telecommun., vol. 18, pp. 79-85, 2011.

[11] M. Ye, K. Xu, J. Wu, and H. Po, “AutoSig-Automatically Generating Signatures for Applications”, Proc. Of IEEE 9th International Conference on Computer and Information Technology (ICTI), Xiamen, China, October 11 – 14 , 2009 .