Research on Automatic Header-Signature Naming System for Internet Service Identification

Su-Kang Lee, Sung-Ho Yoon, and Myung-Sup Kim
Dept. of Computer and Information Science
Korea University
Sejong, Korea
{sukanglee, sungho_yoon, tmskim}@korea.ac.kr

Abstract— **With the rapid growth of the Internet speed and emergence of new applications, the amount of Internet traffic is continuously increasing. In order to provide stable Internet service, efficient network management based on accurate traffic identification is gaining much importance than ever. Header signature-based identification method for network management can be identified the network traffic quickly more than other methods. In this paper, we propose an automatic header-signature naming system and identification system using the named header-signature. The proposed system provides efficient management of header-signature of each service as well. To prove the feasibility of the proposed systems, we applied the system to the campus network environment. In experimental result, we could find the URI information of actual content providers, which cannot find through IP search such as "whois" or command such as "nslookup". In addition, we can get the characteristics of a network in a short period of time by applying the proposed system.**

Keywords—network management; traffic classification; header-signature; signature management;

I. INTRODUCTION

These days, result of massive growth of Internet speed and rise of new applications, there appear various kinds of Internet traffic which takes up large capacity. In such networking conditions, categorizing and analyzing all the traffic by specific applications are very difficult process. Traffic classification method used for analyzing Internet traffic can be sorted by category of signatures. Signatures are divided into Header signature (HS), Payload signature (PS), Statistics signature (SS), and Behavior signature (BS). Header signature uses selected source/destination IP address, source/destination port number of traffic, and used protocol information as signature. Header signature is relatively simple and has no property values to compute when creating the signature. Also, header signature is proper information of a server where applications and services are provided. So, the one that is accurately created can represent a certain server. Traffic classification systems based on header signature effectively overcame limits of one that are based on Payload signature. But header signature does not provide an intuitive information such as service name because it consist of IP address, port number, and protocol number. Also the header signature is only managed by a group information of IP address, port number, and protocol so it is not easy to manage because there is a case that signatures which have different header information but they represent the same service.

In this paper, we propose an automatic header signature naming system. The method we suggest is naming header signature by extracting the name of application or service represented by the header signature when creating the signature.

This paper is described in the following order. We will look over existing researches related to Header signature in Section 2. Suggested Automatic header signature naming system will be explained in Section 3. In Section 4, we will describe results of applying suggested system on real traffic which in order to test its performance. Last, in Section 5, conclusion and further research will be mentioned.

II. RELATED WORK

Current use of Internet is increasing explosively due to emergence of broadcasting and communications, union of various kinds of network, development of many service and application and diversification of user needs. Therefore, a large number of services and applications are created and disappeared ceaselessly. Due to the shortened life-cycles of signatures, existing signatures cannot provide analysis results for perfect network management.

Generation and management of signatures are essential factors regarding analysis of network traffic. Results varies depending on the quality and type of the created signature and it is also a big obstacle in accurate network management. Shapes and attributes of Internet traffic data generated by each services is changed by transforming of server attributes, service type and policy of Internet content provider. Due to these changes, some Internet traffic is impossible to analyze using the old signature. Therefore the network administrator must update the signatures on a regular basis. The signature generation process for signature updates is performed semi-automatically by the advanced workers with specialized skills. The automation of the signature generation process is needed for effective use of advanced human resources.

Recently, many studies are under way to automate the process of signature generation for this reason. In related work

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2015R1D1A3A01018057), Next-Generation Information Computing Development Program through the National Research foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (2010-0020728) and KREONET-Emulab Testbed of Korea Institute of Science and Technology Information.

[1, 2] collects the header information of the traffic from the network and automatically generates the header signature. This study was designed to automate a series of processes for creating and managing the header signature. In conclusion, related work [1,2] propose a system for automatically generating the signature of the application, service which is rapidly changing. Results of this system is the header signature. The header signature is composed of only a number such as IP, PORT, and Protocol. Therefore, the header signature does not provide information as the service name or type explicitly to the network administrator.

In this paper, we propose automatic header signature naming system to address the limitation of the resultant header signature of the existing automatic generation system. A proposed system extract common string pattern from payload of packet in the flow that corresponds to the header signature. And extracted pattern is marked with a name that represents the header signature.

In order to mark a name to the header signature, we must extract a common string pattern from payload of the packets that make up the flow of the header signature. We use Apriori algorithm [3,4] to extract a common string pattern from payload of the packets. Apriori algorithm is one of the association rule mining algorithm technique which is currently widely used. The basic concept is a method to elucidate the association between each data based on the occurrence frequency for the data. Apriori algorithm is simple to implement and its performance is satisfactory. Therefore Apriori algorithm is often used for the pattern analysis. Sequential pattern mining technique uses Apriori algorithm. Sequential pattern mining has been proposed as a technique to find frequently occurring patterns from the sequence of a transaction that occurred with the passage of time. [5].

In this paper, we apply a sequential pattern mining techniques to automatically generate the system existing header signature was given the name in the header signature. Sequential pattern mining extracts a string from the common flow of the packets that make up the header signature. This extracted pattern is a string that can be common strings that satisfy certain support rating. We propose a system that automates a series of processes that generates a header signature and gives a name to the header signature. We applied the proposed system to the environment in which the actual traffic is generated.

III. NAMED HEADER-SIGNATURE BASED TRAFFIC IDENTIFICATION SYSTEM

In this section, we explain the architecture of our Automatic header signature naming system and define the named header-signature. The named header-signature consists of a 3-tuple (IP address, port number, L4 protocol) and its traffic's common payload pattern of an Internet contents that provides a specific application or service. In addition, a weighted value [1, 2] for each header signature is calculated at each maintenance interval.

In Packet-based analysis method case, it causes computation overhead because its traffic data increase dramatically when target network become bigger. For this reason packet-based analysis checks only some packets and it makes difficult to accurate analysis. To supplement this, a flow unit is proposed. A Packet is the smallest unit that is used to send and receive data from the actual Internet network. A flow is a set of packets having the same 5-tuple (source IP, source port, destination IP, destination port, L4 protocol). The flow unit is an effective way to adopt on massive traffic analysis compared to packet-based analysis techniques. But the flow-based analysis does not provide detailed information regarding the behavior of traffic. This is because flow data is generated by using a set of packets in between the end points of two hosts. We need additional analysis on the relationship between flows.

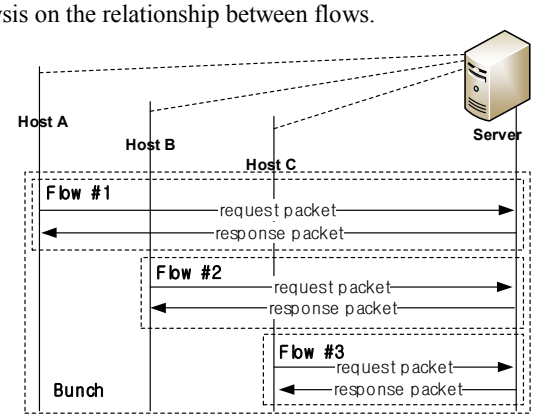


Figure 1. Flows and Bunch between Hosts and Server

Figure 1 shows the bunch. We define a bunch as the relationship between flows for effective signature creation [2]. A bunch is a set of flows with the same 3-tuple (Destination IP address, Destination port number, L4 protocol). All flows in bunch connect to a specific server port. The bunch consist of one server node and a set of flows between counter peers. Named header-signature is generated using a common string extracted from the payload of packets corresponding to the same bunch.

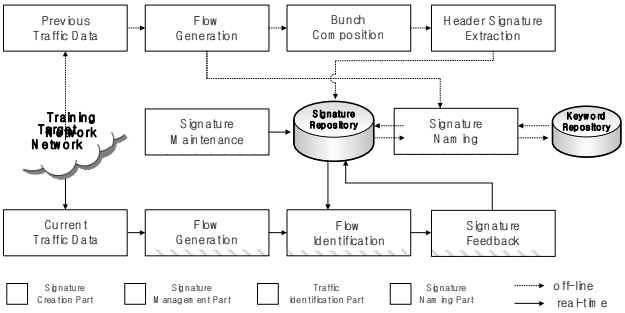


Figure 2. Architecture of the named header signature-based traffic identification system

Figure 2 shows the architecture of the named header signature-based traffic identification system. The system consists of four parts: signature creation part (Flow Generation, Bunch Composition, Signature Extraction), signature naming part (Signature Naming), traffic identification part (Flow Generation, Flow Identification, Signature Feedback), and

signature management part (Signature Maintenance). In signature creation part, previous traffic data which raw packets captured from target network are reconstructed into flows and bunches, and pure header signatures are extracted. Pure header signature is still not naming signature. The reason to create a signature using the previous traffic data is to study the characteristics of the target network. Initial metric values for each extracted signature are set by the statistical values derived from the captured system.

IV. Conclusion and Future Work

In this section, we applied automatic header signature naming system to our campus network in real-time. As a result, we found a header signature that is generated in real-time statistical properties. Also we could find the URI information of actual content providers, which cannot find through IP search such as "whois" or command such as "nslookup". In addition, we have summarized the results of automatically generated named header signature.

A. Experimental Environment

We applied the system proposed in this paper to the campus network environment. The campus network generates traffic in real-time.

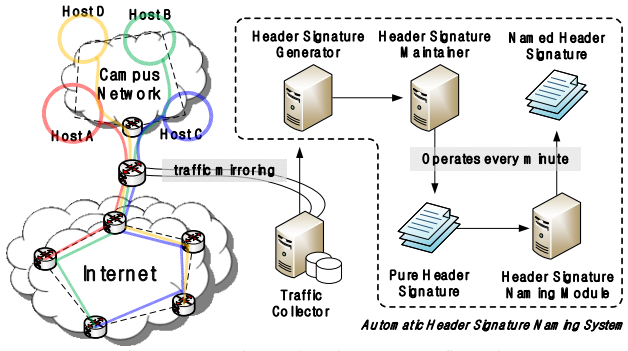
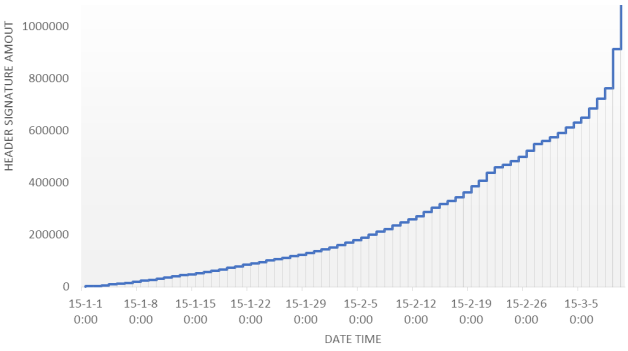


Figure 3. Experimental environment configuration

Figure 3 shows the experimental environment configuration with the automatic header signature naming system in our campus network environment. The traffic data generated in real-time is collected by traffic collector. Header signature generator makes header signature using the traffic data sent by traffic collector. The header signature generated by header signature generator is sent to header signature maintainer. In order to update the header signature, the header signature maintainer calculates weighted value of each header signature.

Finally, the header signature maintainer generates the pure header signature. And next step is header signature naming. The header signature naming module generates the named header signature using pure header signature and the bunch traffic data that matches the pure header signature. In order to extract name candidate of HS, we use Apriori algorithm to bunch's traffic data. Apriori algorithm extracts candidate basis on support value of the pattern in payload. The automatic header signature naming system operates every minute and the named header signature is updated.

B. Experimental Result



Graph 1. Header signature Generation amount

The graph 1 shows the number of generated header signature until 10 March 2015, from 1 January 2015. The number of header signature while steadily increased until the end of February and March were increasing rapidly. In March, the new semester started. Therefore the number of host and amount of traffic in the campus network increased rapidly. From 1 January to 10 March, The number of the header signatures that is maintained with the passage of time is about 1,100,000.

Figure 4 shows the result of candidate generated by automatic header signature naming system that is name of header signature. This is the result generated by using the actual traffic data generated in our campus network. In order to header signature naming, we define the criteria for selecting candidates of header signature. Condition of the candidates is selected as the name of the header signature is as follows.

In Figure 4 shows the candidate of header signature's name. In case, selected name is "Host: www.gs.washington.edu" which has the highest support value and includes string "Host:". Therefore name of header signature (3-tuple information - IP address: 128.208.9.184, Port number: 80, Protocol: TCP) is www.gs.washington.edu.

```
SequenceID : 0  Supporter : 1   ( 24 ) SourceSet : { (152:0) } tcp 163.152.223.228 59018 -> 128.208.
Content : id : 0 Len : 16 Protocol :  HTTP  (content:"GET /favicon.ico"; offset:0; depth:0; within:0;
SequenceID : 1  Supporter : 24  ( 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 ) 
Content : id : 0 Len : 27 Protocol :  HTTP  (content:"Host: www.gs.washington.edu"; offset:0; depth:
SequenceID : 2  Supporter : 23  ( 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 ) Sou
Content : id : 0 Len : 54 Protocol :  HTTP  (content:"GET /academics/courses/akey/56008/lecture/lectu
SequenceID : 3  Supporter : 0   ( ) SourceSet : { (83:0) } tcp 128.208.9.184 80 -> 163.152.223.228 5
```

Figure 4. Candidate of header signature naming process

Table 1 shows the selected name from the names of the candidates in figure 4.

Table 1. Result of header signature naming process

Figure 3	IP address	Port	protocol
3-tuple	128.208.9.184	80	6
Signature name	www.gs.washington.edu		

Table 2 and 3 show the results of IP search with "whois" and "nslookup" command on linux. We can know the owner of IP address through the result of "Whois" IP search and "nslookup" command. However, the result is the information that cannot become name of the header signature. Because the result is not the name of the service or application that generated traffic.

Table 2. result of IP address - ***211.58.221.101*** search with "whois"

NetRange	216.58.192.0 – 216.58.223.255
CIDR	216.58.192.0/19
NetName	GOOGLE
NetHandle	NET-216-58-192-0-1
Parent	NET216 (NET-216-0-0-0-0)
NetType	Direct Allocation
OriginAS	AS15169
Organization	Google Inc. (GOGL)
RegDate	2012-01-27
Updated	2012-01-27
Ref	http://whois.arin.net/rest/net/NET-216-58-192-0-1

Table 3. result of "nslookup command on linux

nslookup ***216.58.221.101***
Server : 168.126.63.1
Address : 168.126.63.1#53
Non-authoritative answer :
101.221.558.216.in-addr.arpa name – hkg07s01-in-f5.le100.net.
Autoritative answers can be found from:
221.58.216.in-addr.arpa nameserver – ns3.google.com
221.58.216.in-addr.arpa nameserver – ns2.google.com
221.58.216.in-addr.arpa nameserver – ns1.google.com
221.58.216.in-addr.arpa nameserver – ns4.google.com
ns1.google.com internet address = 216.239.32.10
ns2.google.com internet address = 216.239.34.10
ns3.google.com internet address = 216.239.36.10
ns4.google.com internet address = 216.239.38.10

Table 4. Named header signature of the major application or services

Name	IP address	Port
accounts.google.com	216.58.221.141	443
	216.58.221.109	443
mail.google.com	***216.58.221.101***	443
clients1.google.com	216.58.221.110	443
www.youtube.com	216.58.221.142	80
cyad.nate.com	203.226.255.9	80
cyad1.nate.com	203.226.255.11	80
i1.shop.nate.com	180.70.134.13	80
m.pann.nate.com	117.53.122.26	80
pann.nate.com	117.53.122.26	80
urs.microsoft.com	111.221.26.253	443
	191.234.231.250	443
crl.microsoft.com	121.254.188.139	80
	121.254.188.138	80
microsoft.com	192.229.145.200	443
fe2.update.microsoft.com	191.232.80.62	443
dmd.metaservices.microsoft.com	65.55.54.42	80
facebook.com	31.13.82.1	443
static.nid.naver.com	202.179.179.108	443
	125.209.226.239	443

As a result of the automatic header signature naming system proposed in this paper, Table 4 shows a summary of the named header signature list major applications and services. We could find the URI information of actual content providers, which cannot find through IP search such as "whois" or command such as "nslookup".

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an automatic header signature naming system in order to overcome the limitations of the existing header signature generated by header signature generation system. To prove the feasibility of the proposed systems, we applied the system to the campus network environment. And we have summarized the named header signature list of major applications or services in our campus network. As a result, we could find the URI information of actual content providers, which cannot find through IP search such as "whois" or command such as "nslookup". In conclusion, we can get the trend information of the target network in a short period of time by using the system proposed in this paper.

We expect to provide a more intuitive information to the network manager by using header signature naming system. Our further research will focus on the construction of precise naming system which can extract the function of header signature and we will construct traffic monitoring system based on more improved automatic header signature naming system.

REFERENCES

[1] Sung-Ho Yoon, Myung-Sup Kim, "An Efficient Method to Maintain the Header Signature for Internet Traffic Identification," Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2013, Hiroshima, Japan, Sep. 25-27, 2013.

[2] Sung-Ho Yoon, Jun-Sang Park, and Myung-Sup Kim, "Signature Maintenance for Internet Application Traffic Identification using Header Signatures," Proc. of the 4th IEEE/IFIP International Workshop of the Management of the Future Internet (ManFI 2012), Hwaii, USA, Apr. 16, 2012.

[3] Reshamwala, Alpa, and Sunita Mahajan. "Improving Efficiency of Apriori Algorithms for Sequential Pattern Mining." Bonfring International Journal of Data Mining 4.1 (2014): 01-06.

[4] Chavan, Mukul B., and Mrs Sarita Patil. "Survey on Web Log Analysis from Central Database System using E-Web Miner Algorithm." (2014).

[5] Han, Jiawei, et al. "FreeSpan: frequent pattern-projected sequential pattern mining." Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2000.