Application Traffic Classification at End Host Level

Seungbae Kim, Hyunchul Kim, Ted “ Taekyoung” Kown, Yanghee Choi
School of Computer Science and Engineering, Seoul National University

{sbkim, hkim, tk, yhchoi}@mmlab.snu.ac.kr

ABSTRACT

Classifying application traffic shows the composition of traffic in terms of number of flows and size. In this work, we classify traffic at end host level to understand characteristics of end host traffic generation. We explain how traffic is composed with number of application categories at end host level with some graphs.

Ⅰ. INTRODUCTION

How many end hosts generate how much application traffic? There has been application traffic classifying researches. [1, 2] The results of these works show how much traffic is generated by each application. However, can we ensure if 40% of the bytes in the trace are web, then 40% of end hosts are responsible for generating web traffic? In this work, we classify traffic at end host level as we assume that an IP address is an end host. Identifying which end host is generating flows using each application is important for understanding end host behavior e.g. the number of end hosts using p2p application, the number of scanning end hosts and victims. It can monitor the tendency of the applications and help to design effective network model. Also, the result of this work can be used as an input of application traffic generation model (e.g. [3]) with the application usage pattern of end host.

Ⅱ. TRAFFIC CLASSIFICATION

We analyze four anonymized payload traces collected at one of four external links located in the KAIST, Korea. The traces have been captured for 3 years annually (Table 1). Therefore, we analyze the annual changes of application traffic generation pattern of end host. For example, we can get the rate of number of end hosts who generate P2P traffic in each year so that we understand tendency of P2P application usage.

Every packet in the traces has 40 Bytes payload. We used the payload-based classifier introduced in [2]. Each application has specific signatures so that classifier categorize into application by inspecting payload contents with the array of signature strings. As a result of the classifier, we get a flow table of traces.

To classify traffic at end host level, we assume that an IP address represents an end host. Based on the assumption, we aggregate flow information with source IP address and destination IP address. The result of source IP aggregation shows the ratio of end hosts which generate application traffic as well as the number of end hosts. The ratio and number of end hosts which receive application flows are shown in the result of destination IP aggregation. The limitation of this work is that an IP address is not an end host in real network because NAT allows many end hosts to use a same IP address and DHCP dynamically assigns IP address to end hosts.

We use two simple filters to separate incoming and outgoing traffic. Since all of four traces are captured in the campus network and most machines are using the same B class network address. We make two different flow tables represent incoming and outgoing flows each. The outgoing flow table has the same prefix of source IP addresses and the incoming flow table has the same prefix of destination IP addresses.

The filtering gives us 8 flow tables. Each trace is separated into two flow tables. The application and IP aggregation processes make complete view of end host level flow information. For analyzing we have four metrics: flows, bytes, srcIP and dstIP.

We use MySQL to maintain big size flow tables and process quickly. It takes long time to make database but it returns results quickly with short queries.

Table 1: Characteristics of analyzed traces

Set	Date	Day	Flows	Bytes	Src. IP	Dst. IP	Payload
KAIST-Ⅰ	2006-09-10	Sun	11M	506G	148K	227K	40 Bytes
KAIST-Ⅱ	2006-09-14	Thu	5M	259G	86K	101K	40 Bytes
KAIST-Ⅲ	2007-01-09	Tue	15M	770G	243K	346K	40 Bytes
KAIST-Ⅳ	2008-03-08	Tue	18M	458G	595K	465K	40 Bytes

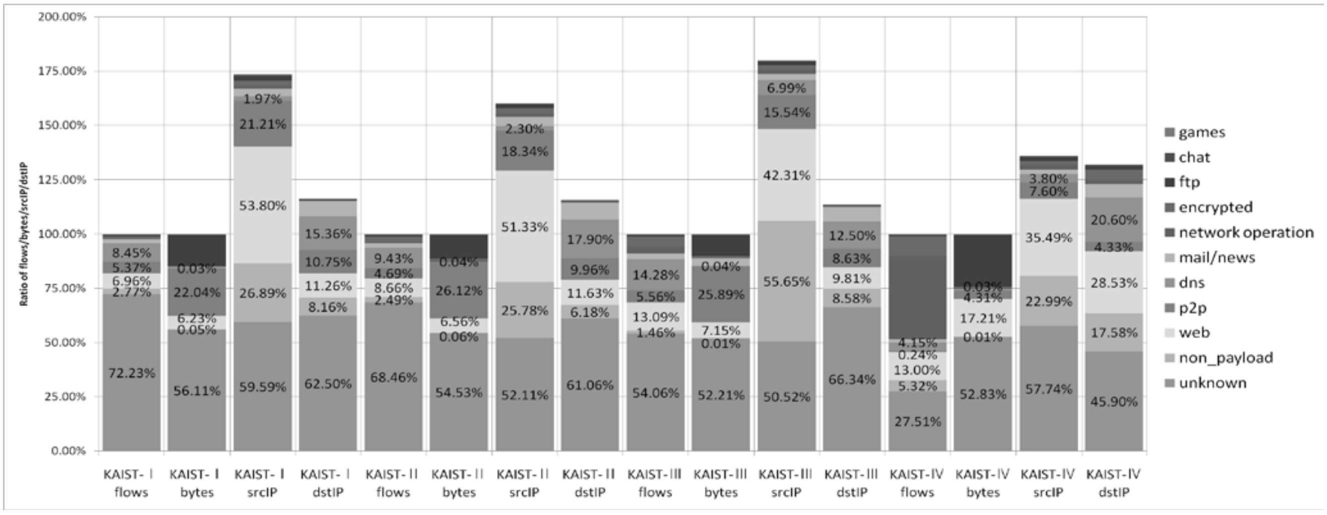

Figure 1: Ratio of four metrics with source IP filtering

Ⅲ. RESULTS

Figure 1 is the result of source filtering traces. There are four traces and each trace has four different metrics. All of source IP addresses have same network prefix, therefore it represents characteristics outgoing flows. Because of space limitation, other figures are not included in this paper.

The sum of flow and bytes ratio is always 100% but sum of IP addresses is over 100% because some end hosts generate or receive more than two application traffic. The rates of four metrics are significantly different as the graph shows. In [3], they proposed the traffic generation model and made an assumption that rate of bytes is same as rate of end hosts. The results of our work give the answer of the assumption. For example, 35~55% of end hosts generate web traffic but it occupies only 7~13% number of flows and 6~17% of total trace size. The ratio of source IP address is also different from the ratio of destination IP address. Some application has more source IP address such as web traffic: 35~55% source end host IP addresses but 10~30% destination end host IP addresses. The number of P2P end hosts is noticeable because the number of source end hosts is twice than the destination end hosts. Since the measurement point is edge and small numbers of end hosts have popular data so that the number of destination is relatively smaller than source end hosts.

The ratio of unknown traffic is about 50% in all metrics with the source filtering. We checked all IP addresses and port numbers in the unknown traffic. Half of the unknown flows are generated by Planet-Lab machines in the KAIST. The result of destination filtering shows that only about 6% end hosts are received unknown traffic. Other unknown traffic is not classified with our method (e.g. new application) or attack traffic. To confirm how much attack traffic contributes all of flows, we used a heuristic algorithm [4]. The number of scanners make many small size packet to many end hosts. We observed that about 70~95% of end hosts are received scanning flows.

Ⅳ. LIMITATIONS AND FUTURE WORK

In this paper, we aggregate flows at end host level based on the assumption that an IP address represents one end host. We cannot distinguish invisible end host. We need to collect trace from end host side, not in network, so that end host information is gathered with the permission from users. With the results of our work, we can design an application traffic generation model or use for input of [3].

Ⅴ. CONCLUSION

In this work, we classified and analyzed four KAIST traces and found some interested results. The ratio of flows, bytes, source end hosts and destination end hosts are totally different from each other and some patterns are observed in the results. Analyzing other traces and making traffic generation model will be our future work.

ACKNOWLEDGMENTS

This work was supported by NAP of Korea Research Council of Fundamental Science & Technology.

REFERENCES

[1]T. Karagiannis, K.P Apagiannaki, M.F Aloutsos, “ BLINC: Multilevel Traffic Classification in the Dark,” in Proc. of ACM SIGCOMM, August 2005.

[2]H. Kim, K.C. Claffy, M. Fomenkova, D. Barman, M. Faloutsos, K.Y. Lee, “ Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices.” In: ACM CoNEXT, Madrid, Spain December 2008.

[3]K. Vishwanath, A. Vahdat. Realistic and responsive network traffic generation. In Proceedings of ACM SIGCOMM, 2006.

[4]M. Allman, V. Paxson, and J. Terrell, “ A brief history of scanning,” in Internet Measurement Conference, October 2007.