New Methods in Payload Based Network Traffic Classification

Béla Hullár, Sándor Laki, András György, and Gábor Vattay

The services provided over the Internet have gone through an enormous evolution in the last decade. Numerous new services and applications have emerged (eg. VoIP, IPTV, file sharing, etc). The appearance of these new services and the growing number of network attacks have brought along the development of new traffic analyzer methods. Network operators would like to identify and control the traffic that travels through their networks. This knowledge may lead to more efficient resource allocation strategies and improved service quality.

In the past different services used their well known TCP or UDP port numbers defined by the Internet Assigned Numbers Authority (IANA). Nowadays most of the applications use dynamic port numbers or well-known trusted ports (such as HTTP or SMTP ports). The reason of this behavior is that applications have to by-pass firewalls and routers, while others try to hide their presence. State-of-the-art, widely used traffic analyzer applications are mostly based on the, so called, deep packet inspection (DPI) approach that aims at identifying typical protocol-patterns in the messages of different network applications. Since these methods cannot handle encrypted traffic, recently several machine learning based traffic classification methods have been developed that do not consider the content of the packets [1]. Furthermore, although the majority of network traffic is still unencrypted and the DPI solutions perform well in practice, their disadvantages are well-known: these methods require significant computing resources and the integration of new applications requires expert knowledge about the application's protocol. Statistical payload inspection can give a proper solution to these issues [2, 3].

This paper examines how different data compression models can be used for packet payload based protocol identification. Our results show that compression-based modeling can provide an effective solution for traffic classification, similarly to their performance in other domains of clustering [4]. We found that the majority of the protocols are identifiable from the first several bytes of the application data. To demonstrate our results, different real network traces generated by different network applications were used.

Acknowledgements

The authors thank the partial support of the National Office for Research and Technology (NAP 2005/ KCKHA005) project.

References

[1] Thuy T.T. Nguyen, Grenville Armitage: A Survey of Techniques for Internet Traffic Classification using Machine Learning, *IEE Communications Surveys and Tutorials, vol. 10 no. 4 pp. 56-76, 2008.*

[2] A. Finamore, M. Mellia, M. Meo, and D. Rossi: KISS: Stochastic Packet Inspection,*In Traffic Measurement and Analysis (TMA), Springer- Verlag LNCS 5537, May 2009.*

[3] Justin Ma, Kirill Levchenko, Christian Kreibich, Stefan Savage, Geoffrey M. Voelke: Unexpected means of protocol inference, *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement, October 25-27, 2006, Rio de Janeriro, Brazil*

[4] R. Cilibrasi, P.M.B. Vitanyi: Clustering by compression,*IEEE Trans. Information Theory, 51:4(2005), 1523-1545.*