Computer Networks 55 (2011) 1158–1167

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Quantifying the accuracy of the ground truth associated with Internet traffic traces

Maurizio Dusi, Francesco Gringoli, Luca Salgarelli *

Università degli Studi di Brescia, Italy

ARTICLE INFO

Article history:
Received 9 February 2010
Received in revised form 21 July 2010
Accepted 9 November 2010
Available online 23 November 2010
Responsible Editor: I.F. Akyildiz

Keywords:
Internet
Measurement
Traffic
Characterization

ABSTRACT

Ground truth information for Internet traffic traces is often derived by means of port analysis and payload inspection (Deep Packet Inspection – DPI). In this paper we analyze the errors that DPI and port analysis commit when assigning protocol labels to traffic traces. We compare the ground truth provided by these approaches with that derived by `gt`, a tool that we developed, which provides error-free ground truth at the application level by construction. Experimental results demonstrate that, depending on the protocols composing a trace, ground truth information from port analysis and DPI can be incorrect for up to 91% and 26% of the labeled bytes, respectively.

1. Introduction

An increasing number of research activities in the field of Internet traffic measurement and analysis rely on the availability of network traces associated with ground truth data, i.e., information about the protocol and the application behind each flow. The research community commonly adopts two methods to derive ground truth, based on the analysis of port numbers at the transport layer and of application payloads, the latter through Deep Packet Inspection (DPI) techniques.

However, known protocols can work on ports different than those assigned by IANA, for example in order to circumvent security restrictions. Furthermore, emerging protocols such as those used by Peer-to-Peer (P2P) and Streaming applications do not even use standard ports. Finally, even DPI mechanisms often fail with encrypted or obfuscated traffic: this is the case with protocols protected by TLS or with applications such as Skype.

In this paper we quantify the error that classical, port-based and DPI-based approaches commit in establishing ground truth related to application-layer protocols. We evaluate these approaches on traffic traces that we collected, for which ground truth is made available by means of `gt` [1], a tool that we developed and that, by construction, generates accurate ground truth at the application level.

We show that classical approaches can lead to high true-positive rates labeling the traffic produced by clear text protocols such as HTTP and SMTP, yet they sometimes fail to produce a label for some of the traffic produced by those same protocols, and end up mis-labeling much of the P2P, Streaming and Voice over IP traffic. For example, we show that in our traces DPI mislabels up to 6% of the bytes produced by email-related protocols, and up to 60% of those produced by P2P applications. Finally, we use our findings to evaluate the errors one would commit if they were to use transport ports as ground truth in one of the most recent publicly-available, anonymized traces such as the one available at [2].

The rest of the paper is organized as follows. In Section 2 we report on related work. In Section 3 we describe the methodology we used to evaluate the accuracy of the DPI

* Corresponding author. Tel.: +39 030 371 5847; fax: +39 030 380 014.
E-mail addresses: luca.salgarelli@ing.unibs.it, first.last@ing.unibs.it (L. Salgarelli).

1389-1286/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2010.11.006

and port-based techniques. In Section 4 we describe the traffic traces we considered, while in Section 5 we quantify the errors those techniques commit on our dataset. Section 6 puts such errors in perspective considering the impact they might have on Internet traffic analysis works that use one of the latest publicly available, anonymized traces. Finally, Section 7 concludes the paper.

2. Related work

DPI alone is one of the most popular techniques to establish *protocol* ground truth. This mechanism is used in many works, such as [3–6], if the traces contain at least a portion of the payload: examples of tools implementing DPI are Bro [23] and *l7filter* [19].

Port-based mechanisms are also used, and can be the only option when working with publicly available traces where payloads have been entirely stripped [7,8]. Karagiannis et al. [24] showed that transport-layer information can be exploited to profile end-host systems, and proposed graphs to capture flow information and inter-flow interdependencies that help the provisioning and the anomaly detection in a network environment.

However, port-based classification cannot detect traffic that is mis-using well-known ports: some works [9] showed that this approach is inaccurate to identify network applications, given that there is a variety of applications that either do not use well-known port numbers or use other protocols, such as HTTP, as wrappers to elude security policies.

To the best of our knowledge, port analysis and payload inspection, with the exception of the four publications mentioned in the following, are the only mechanisms currently used to associate ground truth with network flows. No studies that we are aware of have yet fully quantified the accuracy of those methods, even though several works [6,10] support the idea that those techniques cannot serve such purpose anymore. In [25] the authors exploit the Bro system's Dynamic Protocol Detection (DPD), a payload-based technique, to quantify the accuracy achieved by a port-based approach on residential broadband traffic traces. Given a well-known port on which the protocol P is supposed to run, they compute the fraction of bytes that belongs to *P* according to the DPD classifier out of the volume of bytes on that port. Given a protocol *P*, they also compute the fraction of bytes that runs on P's default port. According to [25], our results support the idea that port-based classification works well for some classes of traffic and protocols, such as Web and Mail, and cannot be considered as reliable in detecting P2P and VoIP traffic. Unlike [25], we provide a way that allows us to quantify the accuracy of payload-based techniques starting from accurate ground truth at the application level as provided by `gt`, and we report the results of this study in term of flow and byte accuracy. We compare the accuracy of three techniques used to provide ground truth to traffic traces, i.e., port-based, payload-based and a combination of both, on traces captured on two campus environments, and we finally report on the accuracy of the DPI approach with increasing values of payload size. Our results show that big fraction of traffic are erroneously classified even by DPI techniques (up to 60% of the volume of P2P bytes).

An alternative method for deriving accurate ground truth is proposed in [11]. The mechanism records part of the name of the application that generated each IP packet and embeds this information into the packet itself inside a Router Alert IP option. The purpose of the tool is similar to that of `gt`, i.e., both tools allow to collect a ground truth information which is 100% accurate (the name of the application that generated a given flow is error free). However, in [11] such information is used to verify the performance of a statistical classification mechanism, whereas in this paper we are interested in using accurate ground truth to quantify the errors produced by DPI and port analysis.

In [12], the authors recently presented a new platform, the Ground Truth Verication System (GTVS), that uses a combination of heuristics at different levels (host, flow, packet) to improve the quality of ground truth associated with packet traces. While the ground truth provided by GTVS is compared in their paper with that obtained by DPI alone, in our work we further extend the analysis to port-based mechanisms. Furthermore, the GTVS platform does not (yet) include application labels guaranteed to be accurate, such as those provided by `gt` and used in this paper.

In [13], payloads are analyzed to establish ground truth beyond the *application* level and to determine what a network flow is actually embedding, e.g., what kind of attachment inside an email message, or what kind of object in a HTTP GET response. Though this work is the first to check the ultimate components inside a network flow, it is more focused on "object" ground truth than on the application protocol.

Finally, we introduced `gt` in [1]. Besides describing the architecture of the software toolset, we also analyzed how `gt` could be paired with an application-layer signature-matching approach to improve the accuracy of DPI. However, we did not perform the extensive analysis against both DPI and port analysis that we report in this paper.

3. Mechanisms to associate protocol ground truth with network flows

In this section we briefly report on the three methods generally used to associate protocol ground truth with traffic flows: the analysis of port numbers, payload inspection and a combination of those techniques. We then describe the method we adopted to establish protocol ground truth when the traces are collected with the help of the `gt` software toolset.

3.1. The port-based approach

Inferring the application protocol from port numbers at the transport layer is a fast and simple method, easy to implement. In this case the destination port number towards which the flow is exchanging traffic is cross-referenced with the IANA table [14] to detect the corresponding application which is supposed to run on that port. Currently, this method is the only one available for