Computer Networks 76 (2015) 75–89

Contents lists available at ScienceDirect

Computer Networks

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

Independent comparison of popular DPI tools for traffic classification

Tomasz Bujlow [a,*], Valentín Carela-Español [b], Pere Barlet-Ros [b]

[a] Section for Networking and Security, Department of Electronic Systems, Aalborg University, DK-9220 Aalborg East, Denmark
[b] Broadband Communications Research Group, Department of Computer Architecture, Universitat Politécnica de Catalunya, ES-08034 Barcelona, Spain

ARTICLE INFO

Article history:
Received 19 May 2014
Received in revised form 5 September 2014
Accepted 3 November 2014
Available online 10 November 2014

Keywords:
Deep packet inspection
PACE
nDPI
Libprotoident
NBAR
L7-filter

ABSTRACT

Deep Packet Inspection (DPI) is the state-of-the-art technology for traffic classification. According to the conventional wisdom, DPI is the most accurate classification technique. Consequently, most popular products, either commercial or open-source, rely on some sort of DPI for traffic classification. However, the actual performance of DPI is still unclear to the research community, since the lack of public datasets prevent the comparison and reproducibility of their results. This paper presents a comprehensive comparison of 6 well-known DPI tools, which are commonly used in the traffic classification literature. Our study includes 2 commercial products (*PACE* and *NBAR*) and 4 open-source tools (*OpenDPI*, *L7-filter*, *nDPI*, and *Libprotoident*). We studied their performance in various scenarios (including packet and flow truncation) and at different classification levels (application protocol, application and web service). We carefully built a labeled dataset with more than 750 K flows, which contains traffic from popular applications. We used the Volunteer-Based System (VBS), developed at Aalborg University, to guarantee the correct labeling of the dataset. We released this dataset, including full packet payloads, to the research community. We believe this dataset could become a common benchmark for the comparison and validation of network traffic classifiers. Our results present *PACE*, a commercial tool, as the most accurate solution. Surprisingly, we find that some open-source tools, such as *nDPI* and *Libprotoident*, also achieve very high accuracy.

1. Introduction

Network communication became the standard way of exchanging information between applications located on different hosts. The exchanged application-layer data is segmented and encapsulated into IP packets, which are transmitted through the network. Deep Packet Inspection (DPI) tools analyze the content of the packets by searching for specific patterns (i.e., signatures). Thus, DPI became one of the fundamental traffic analysis methods for many tools performing traffic classification, network management, intrusion detection, and network forensics.

DPI-based traffic classification relies on a database of characteristic signatures of protocols (e.g., HTTP, or POP3), applications (e.g., Skype, or BitTorrent), and web services (e.g., Facebook, or YouTube). These signatures must be initially extracted and kept up to date in order to adapt them to the continuous evolution of the applications (e.g., new applications, new versions, new obfuscation techniques). They are later employed, usually in an online phase, to classify the traffic flowing in a network. The pattern matching algorithms for online classification

* Corresponding author. Tel.: +45 5026 2494.
E-mail addresses: tomasz@bujlow.com (T. Bujlow), vcarela@ac.upc.edu (V. Carela-Español), pbarlet@ac.upc.edu (P. Barlet-Ros).
URLs: http://tomasz.bujlow.com (T. Bujlow), http://personals.ac.upc.edu/vcarela/ (V. Carela-Español), http://people.ac.upc.edu/pbarlet/ (P. Barlet-Ros).

http://dx.doi.org/10.1016/j.comnet.2014.11.001
1389-1286/

are computationally complex and usually require expensive hardware. Nevertheless, DPI is commonly considered as the most accurate technique for traffic classification, and most commercial solutions rely on it [1–4].

The scientific literature shows that the existing traffic classification techniques give accurate results. Unfortunately, as pointed out in [5], there are numerous problems with the comparison and validation of these techniques. The quality of the validation process directly depends on the methods used to build and pre-classify the ground-truth dataset, so it is not an easy task to compare how various classifiers perform. The researchers must choose between two approaches of ground-truth establishment: creating their own dataset, or using an already existing dataset created by someone else.

The first approach requires the researcher to build and label the dataset by himself. Both of these tasks are challenging. Building the dataset involves the selection of the applications to be included in the dataset. Data labeling is usually carried out by DPI tools given their presumably high accuracy. However, the actual accuracy of DPI-based tools is still not clear, as the previous works that tried to compare the performance of different DPI tools showed that the ground-truth used to validate the proposals was obtained through port-based classifiers, other DPI-based tools, or methodologies of unknown reliability [6–10]. Thus, the results of the comparison are arguable. In addition, most commercial tools are black boxes that claim high accuracy based on their own studies, which cannot be validated, because they were performed using private datasets.

The second approach to the proper validation of traffic classification techniques is the use of publicly available datasets. This way, it is easier to compare the results obtained from different tools. Examples are the datasets published by the Cooperative Association for Internet Data Analysis (CAIDA) [11] or the Internet Measurement Data Catalog [12]. Although the datasets are pre-classified, the actual reliability of this labeling is unknown, which is a very important factor for testing traffic classifiers. Most of the traces have no payload or just the first bytes of each packet. The MAWI repository [13] contains various packet traces, including daily 15-min traces made at an trans-Pacific line (150 Mbit/s link). The bandwidth of this link has changed through the years. The traces contain the first 96 bytes of the payload and the traffic is usually asymmetric. Another useful data source is the Community Resource for Archiving Wireless Data At Dartmouth (CRAWDAD) [14], which stores wireless trace data from many contributing locations. Another interesting project is The Waikato Internet Traffic Storage (WITS) [15], which aims to collect and document all the Internet traces that the WAND Network Research Group from the University of Waikato has in their possession. Some of the traces can be freely downloaded and they contain traffic traces from various areas and of different types (as DSL residential traffic, university campus traffic, etc.). Most of the traces do not have payload (i.e., it is zeroed) or truncated. All these datasets although useful for many network evaluations are of limited interest for DPI validation given that no correct labeling can be performed on them.

Szabó et al. [16] introduced a method for validation of classification algorithms, which is independent of other classification methods, deterministic, and allows to automatize testing of large data sets. The authors developed a Windows XP driver based on the Network Driver Interface Specification (NDIS) library. Another approach to obtain the ground-truth was taken in [17]. The authors created a tool, which collects the data from the network and labels the flows with the real application names (e.g., *Thunderbird*) and application protocol names (e.g., *SMTP*). This tool is similar to our Volunteer-Based System (VBS), which was used in our work for the ground-truth generation and is further described in Section 3.1.1. Yet another way of establishing the ground-truth was shown in [18], which describes a system developed to accelerate the manual verification process. The authors proposed Ground Truth Verification System (GTVS) based on the DPI signatures derived from the databases available in the Internet, including L7-filter. GTVS, however, does not collect the application names from the operating systems, so the established truth cannot be completely verified.

In a former conference paper [19], we tried to face the problem of ground-truth reliability. We described various methods of obtaining the ground-truth for testing various traffic classification tools. As part of the evaluation, we tested several DPI-based traffic classifiers and assessed if they can be used for obtaining reliable ground-truth. This paper is not only an extension but a more broad and comprehensive validation of DPI-based tools. The focus of this paper is different, as we directly compare the selected DPI tools regarding their per-class accuracy. Reference datasets used in this paper as well as the previous one are generated by the same methodology further explained in Section 3. However, both datasets are significantly different. The dataset used in this paper is significantly larger than the one used in the conference paper. Furthermore, the new dataset also contains labeled non-HTTP flows belonging to various web services, which is a unique feature. The methodology of testing the classifiers is also different. In our previous paper, we tested the accuracy of the classifiers on a single level. In the current paper, we evaluate different levels (i.e., application, web service, etc.), so the new evaluation method is more detailed and complete.

This paper compares and validates six well-known DPI-based tools used for network traffic classification. In order to allow the validation of our work, we publish the reliable labeled dataset used to perform our study. Two main aspects have been carefully addressed when building this dataset: the reliability of the labeling and the representativeness of the data. We used the VBS tool [20] to guarantee the correctness of the labeling process. This tool, described in Section 3, is able to label the flows with the name of the process that creates them. This allowed us to create a reliable ground-truth that can be used as a reference benchmark for the research community to compare other proposals. The selection of applications for our dataset was made based on well-known indexes of the most commonly used Internet applications and web services. In order to allow the publication of the dataset and avoid any privacy issues, we created the traffic by running a large