Reviewing Traffic Classification

Silvio Valenti[1,4], Dario Rossi[1], Alberto Dainotti[2,5], Antonio Pescapè[2], Alessandro Finamore[3], and Marco Mellia[3]

[1] Telecom ParisTech, France
first.last@enst.fr
[2] Università di Napoli Federico II, Italy
last@unina.it
[3] Politecnico di Torino, Italy
first.last@polito.it
[4] Google, Inc.
[5] CAIDA, UC San Diego

Abstract. Traffic classification has received increasing attention in the last years. It aims at offering the ability to automatically recognize the application that has generated a given stream of packets from the direct and passive observation of the individual packets, or stream of packets, flowing in the network. This ability is instrumental to a number of activities that are of extreme interest to carriers, Internet service providers and network administrators in general. Indeed, traffic classification is the basic block that is required to enable any traffic management operations, from differentiating traffic pricing and treatment (e.g., policing, shaping, etc.), to security operations (e.g., firewalling, filtering, anomaly detection, etc.).

Up to few years ago, almost any Internet application was using well-known transport layer protocol ports that easily allowed its identification. More recently, the number of applications using random or non-standard ports has dramatically increased (e.g. Skype, BitTorrent, VPNs, etc.). Moreover, often network applications are configured to use well-known protocol ports assigned to other applications (e.g. TCP port 80 originally reserved for Web traffic) attempting to disguise their presence.

For these reasons, and for the importance of correctly classifying traffic flows, novel approaches based respectively on packet inspection, statistical and machine learning techniques, and behavioral methods have been investigated and are becoming standard practice. In this chapter, we discuss the main trend in the field of traffic classification and we describe some of the main proposals of the research community.

We complete this chapter by developing two examples of behavioral classifiers: both use supervised machine learning algorithms for classifications, but each is based on different features to describe the traffic. After presenting them, we compare their performance using a large dataset, showing the benefits and drawback of each approach.

1 Introduction

Traffic classification is the task of associating network traffic with the generating application. Notice that the TCP/IP protocol stack, thanks to a clear repartition between

E. Biersack et al. (Eds.): Data Traffic Monitoring and Analysis, LNCS 7754, pp. 123–147, 2013.

Table 1. Taxonomy of traffic classification techniques

Approach	Properties exploited	Granularity	Timeliness	Comput. Cost
Port-based	Transport-layer port [49, 50, 53]	Fine grained	First Packet	Lightweight
Deep Packet Inspection	Signatures in payload [44, 50, 60]	Fine grained	First payload	Moderate, access to packet payload
Stochastic Packet Inspection	Statistical properties of payload [26, 30, 37]	Fine grained	After a few packets	High, eventual access to payload of many packets
Statistical	Flow-level properties [38, 45, 50, 58]	Coarse grained	After flow termination	Lightweight
	Packet-level properties [8, 15]	Fine grained	After few packets	Lightweight
Behavioral	Host-level properties [35, 36, 67]	Coarse grained	After flow termination	Lightweight
	Endpoint rate [7, 28]	Fine grained	After a few seconds	Lightweight

layers, is completely agnostic with respect to the application protocol or to the data carried inside packets. This layered structure has been one of the main reasons for the success of the Internet; nevertheless, sometimes network operators, though logically at layer-3, would be happy to know to which application packets belong, in order to better manage their network and to provide additional services to their customers. Traffic classification is also instrumental for all security operations, like filtering unwanted traffic, or triggering alarms in case of an anomaly has been detected.

The information provided by traffic classification is extremely valuable, sometimes fundamental, for quite a few networking operations [38, 42, 46, 52]. For instance, a detailed knowledge of the composition of traffic, as well as the identification of trends in application usage, is required by operators for a better *network design and provisioning*. *Quality of service* (QoS) solutions [58], which prioritize and treat traffic differently according to different criteria, need first to divide the traffic in different classes: identifying the application to which packets belong is crucial when assigning them to a class. In the same way, traffic classification enables differentiated class *charging* or Service Level Agreements (SLA) verification. Finally, some national governments expect ISPs to perform *Lawful Interception* [6] of illegal or critical traffic, thus requiring them to know exactly the type of content transmitted over their networks. Traffic classification represents in fact the first step for activities such as *anomaly detection* for the identification of malicious use of network resources, and for security operation in general, like firewalling and filtering of unwanted traffic [53, 56].

If, on the one hand, the applications of traffic classification are plentiful, on the other hand, the challenges classifiers have to face are not to be outdone. First, they must deal with an increasing amount of traffic as well as equally increasing transmission rates: to cope with such speed and volume, researchers are looking for *lightweight algorithms* with as little computational requirements as possible. The task is further exacerbated by developers of network applications doing whatever in their power to hide traffic and to elude control by operators: traffic encryption and encapsulation of data in other