Toward the Accurate Identification of Network Applications

Andrew W. Moore[1,⋆] and Konstantina Papagiannaki[2]

[1] University of Cambridge
andrew.moore@cl.cam.ac.uk
[2] Intel Research, Cambridge
dina.papagiannaki@intel.com

Abstract. Well-known port numbers can no longer be used to reliably identify network applications. There is a variety of new Internet applications that either do not use well-known port numbers or use other protocols, such as HTTP, as wrappers in order to go through firewalls without being blocked. One consequence of this is that a simple inspection of the port numbers used by flows may lead to the inaccurate classification of network traffic. In this work, we look at these inaccuracies in detail. Using a full payload packet trace collected from an Internet site we attempt to identify the types of errors that may result from port-based classification and quantify them for the specific trace under study. To address this question we devise a classification methodology that relies on the full packet payload. We describe the building blocks of this methodology and elaborate on the complications that arise in that context. A classification technique approaching 100% accuracy proves to be a labor-intensive process that needs to test flow-characteristics against multiple classification criteria in order to gain sufficient confidence in the nature of the causal application. Nevertheless, the benefits gained from a content-based classification approach are evident. We are capable of accurately classifying what would be otherwise classified as unknown as well as identifying traffic flows that could otherwise be classified incorrectly. Our work opens up multiple research issues that we intend to address in future work.

1 Introduction

Network traffic monitoring has attracted a lot of interest in the recent past. One of the main operations performed within such a context has to do with the identification of the different applications utilising a network's resources. Such information proves invaluable for network administrators and network designers. Only knowledge about the traffic mix carried by an IP network can allow efficient design and provisioning. Network operators can identify the requirements of

[⋆] Andrew Moore thanks the Intel Corporation for its generous support of his research fellowship.

C. Dovrolis (Ed.): PAM 2005, LNCS 3431, pp. 41–54, 2005.

different users from the underlying infrastructure and provision appropriately. In addition, they can track the growth of different user populations and design the network to accommodate the diverse needs. Lastly, accurate identification of network applications can shed light on the emerging applications as well as possible mis-use of network resources.

The state of the art in the identification of network applications through traffic monitoring relies on the use of well known ports: an analysis of the headers of packets is used to identify traffic associated with a particular port and thus of a particular application [1, 2, 3]. It is well known that such a process is likely to lead to inaccurate estimates of the amount of traffic carried by different applications given that specific protocols, such as HTTP, are frequently used to relay other types of traffic, e.g., the NeoTeris VLAN over HTTP product. In addition, emerging services typically avoid the use of well known ports, e.g., some peer-to-peer applications. This paper describes a method to address the accurate identification of network applications in the presence of packet payload information[1]. We illustrate the benefits of our method by comparing a characterisation of the same period of network traffic using ports-alone and our content-based method.

This comparison allows us to highlight how differences between port and content-based classification may arise. Having established the benefits of the proposed methodology, we proceed to evaluate the requirements of our scheme in terms of complexity and amount of data that needs to be accessed. We demonstrate the trade-offs that need to be addressed between the complexity of the different classification mechanisms employed by our technique and the resulting classification accuracy. The presented methodology is not automated and may require human intervention. Consequently, in future work we intend to study its requirements in terms of a real-time implementation.

The remainder of the paper is structured as follows. In Section 2 we present the data used throughout this work. In Section 3 we describe our content-based classification technique. Its application is shown in Section 4. The obtained results are contrasted against the outcome of a port-based classification scheme. In Section 5 we describe our future work.

2 Collected Data

This work presents an application-level approach to characterising network traffic. We illustrate the benefits of our technique using data collected by the high-performance network monitor described in [5].

The site we examined hosts several Biology-related facilities, collectively referred to as a *Genome Campus*. There are three institutions on-site that employ

[1] Packet payload for the identification of network applications is also used in [4]. Nonetheless, no specific details are provided by [4] on the implementation of the system thus making comparison infeasible. No further literature was found by the authors regarding that work.