Computer Networks 58 (2014) 127–140

Contents lists available at ScienceDirect

Computer Networks

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

A measurement-based study on the correlations of inter-domain Internet application flows ☆

Xiaofei Wu [a,*], Xin Wang [b], Ke Yu [a], Frank Y. Li [c]

[a] School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[b] Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, New York, USA
[c] Department of Information and Communication Technology, University of Agder (UiA), Grimstad, Norway

ARTICLE INFO

Article history:
Received 7 April 2013
Received in revised form 6 August 2013
Accepted 4 September 2013
Available online 20 September 2013

Keywords:
Internet traffic monitoring
Application flow
Assortativity
Complex network

ABSTRACT

Internet traffic characterization has a profound impact on network engineering and traffic identification. Existing studies are often carried out on a per-flow basis, focusing on the properties of individual flows. In this paper, we study the interaction of Internet traffic flows and network features from a complex network perspective, focusing on six types of applications: P2P file sharing, P2P stream, HTTP, instant messaging, online games and abnormal traffic. With large-volume traffic flow records collected through proprietary line-speed hardware-based monitors, we construct flow graphs of these different application types. Based on the flow graphs, we calculate the correlation coefficients on various properties for individual or multiple applications. Our studies on associativity among degree and strength of individual hosts and connected nodes reveal distinct correlative behavior of different types of applications. Especially, the correlations of P2P applications are observed to be much stronger than those of the other applications. We also investigate the correlations between different types of applications, and observe that HTTP has remarkably different correlations from those of the two P2P applications due to the fact that multiple application types rely on HTTP. Finally, we study the dynamics of correlations for a period of 24 h and reveal a few interesting trends. We believe that our work which focuses on the assortativities of Internet applications provides insightful understanding on Internet traffic classification of up-to-date applications and will be helpful for Internet traffic classification and engineering.

1. Introduction

Study on the characteristics of Internet traffic is important for understanding the activities and behavior of the Internet, and it is also essential for Internet service providers (ISPs) to better manage the operation of the Internet [1]. Moreover, understanding the characteristics of applications and traffic flows is helpful for many network operations, including planning and provisioning of the network infrastructure, traffic engineering and performance optimization, guaranteeing of service quality and protection of network from fraud. Indeed, the characteristics of Internet traffic have been studied extensively, and properties such as heavy-tail distribution, self-similarity and fractal behavior have long been understood [2].

Many recent studies on Internet traffic are carried out at the flow level. An Internet flow is formed with a series of packets exchanged between two hosts, identified by the well-known five tuples: source IP address, destination IP address, source port, destination port, and protocol type. Usually, a flow tracks the information exchanged for a complete Internet interaction, and the study of flow characteristics reveals how the Internet is accessed. Therefore, investigating Internet flows, both at the

☆ BHASKARANJ1 17-SEP-2013 12:04:51 Open Access Yes in EES, but no funding form. OA and CC undone.

* Corresponding author. Tel.: +86 10 62283147; fax: +86 10 62283295.
E-mail addresses: wuxf@bupt.edu.cn (X. Wu), xwang@ece.sunysb.edu (X. Wang), yuke@bupt.edu.cn (K. Yu), frank.li@uia.no (F.Y. Li).

1389-1286/$ - see front matter © 2014 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.comnet.2013.09.001

aggregate level and at the individual level, can provide insight on the features of Internet traffic. Flow records are widely used to examine the characteristics of Internet traffic and applications in the literatures, e.g., the studies for inter-domain traffic [3], peer-to-peer (P2P) applications [4], security reasons [5] and entertainment [6].

So far, most of the existing studies on flows were based on the observations of single flows, where millions of flow records were examined one by one and the properties of flows were retrieved. The properties studied include the number of packets and bytes in a flow, flow duration, distributions of packet size and interval, etc. Based on the information retrieved from flows, such as destination port and protocol type, flows can be classified into different applications and the statistics of applications are obtained from the observation of a large number of flows. While this method is able to reveal the characteristics of an application flow, it does not fully utilize the flow records for more thorough understanding of the features of Internet traffic. In fact, flows do not exist independently but correlate with each other via network elements, e.g., hosts and connections. Therefore, traffic is also observed to present complex network characteristics in Internet [7]. For instance, in P2P file sharing, flows between two hosts do not only carry file contents of the two hosts, but also include information of other participants. Therefore, flows sharing the same content are related and properties such as flow duration are affected by the composition of concurrent flows. The Internet flows form dynamic overlays on top of the physical and logical network structures, and these dynamic overlays manifest the vivid prospect of Internet interactions, therefore are very important for us to better comprehend the nature of Internet applications. Instead of constraining our studies to individual flow traffic, in this paper, we investigate the important properties of the overlaying flow infrastructure from a complex network point of view based on aggregated and correlated flows.

The main objective of this study is to investigate the degree and strength correlations within and across the Internet applications for inter-domain traffic. The concept of complex networks and graph [8] is adopted in this study. We construct flow graphs from the Internet traffic flow records we collected from operational networks and investigate their properties. In our previous paper [9], we have investigated the important characteristics of the flow graphs such as the distributions of node degree and the strength of different applications. In this paper, we further examine the flow graphs and analyze the correlations between different properties of network elements, both within one type of application and across different types of applications. More specifically, we study the mixing pattern and correlations, which are the important properties of complex networks [10]. To the best of our knowledge, this is the first effort to study the mixing pattern of different Internet applications based on a large number of inter-domain flow records and new traffic types such as online game and abnormal traffic. Besides the pattern mixing of individual application, we also study the assortativities between different applications, for example, between HTTP and P2P streams. The flow records were collected by a proprietary line-speed hardware-based monitor with a capturer and a classifier to track the traffic of a 10 Gbps trunk link between an access network and the backbone. The records consist of detailed traffic information with application classifications. Taking the advantage of this application classification ability, we are able to analyze the properties of multiple types of applications. In addition to the widely studied application types, i.e., HTTP and P2P file sharing (P2PF), we also analyze four other types of applications: instant messaging (IM), online games (OG), P2P streams (P2PS), and abnormal traffic (AT). The AT traffic includes virus and attacks detected. The mixing patterns and assortativities of specific Internet applications revealed by our work would help better predict traffic trends, which will in turn serve as a guideline for better Internet service provisioning and traffic engineering.

The main contributions of this study are as follows:

1. We examine flow records for six different types of traffic, namely P2PF, HTTP, IM, OG, P2PS and AT. We exhibit the profiles of nodes, connections and traffic volume of these applications, as well as their variations of traffic parameters within a 24-h period. The analysis is performed based on a huge volume of data captured in a two-day period by a proprietary hardware monitor located between an access network and the backbone network. From the traffic analysis, we find that many hosts run more than one application simultaneously, and on an average each host runs 1.43 application.
2. We construct graphs from the flow records and examine the assortativities among node degree and strength of a specific application. More specifically, we calculate four types of correlation coefficients for a node and six types of coefficients for a connected node pair, and identify different assortative behavior for different type of applications.
3. We investigate the correlations across different types of applications by examining correlations between the properties of nodes participating in multiple applications. Based on the calculation of four types of correlation coefficients related to degree and strength, we conclude that the correlations between different pair of applications are very different.
4. We present the correlation coefficients in a 24-h period to explore the time variation of application relationship, from which we observe different trends of different types of applications.

The rest of the paper is organized as follows. After a brief review of the related work in Section 2, our data sources and the flow graph construction procedures are introduced in Section 3. In Section 4, The methodology of correlation analysis is expatiated. Then our results are presented and analyzed in Section 5. Finally, the paper is concluded in Section 6.

2. Related work

Internet traffic has been studied extensively to reveal its characteristics. In general, data traffic is well known to be self-similar and fractal [2,11]. The traffic of a residential