Computer Networks 53 (2009) 761–773

Contents lists available at ScienceDirect

Computer Networks

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

A nonlinear, recurrence-based approach to traffic classification

Francesco Palmieri *, Ugo Fiore

Università degli Studi di Napoli Federico II, CSI, Complesso Universitario Monte S. Angelo, Via Cinthia 5, 80126 Napoli, Italy

ARTICLE INFO

Article history:
Available online 29 December 2008

Keywords:
Recurrence plots
Recurrence quantification analysis
Nonlinear analysis
Traffic classification

ABSTRACT

The ability to accurately classify and identify the network traffic associated with different applications is a central issue for many network operation and research topics including Quality of Service enforcement, traffic engineering, security, monitoring and intrusion-detection. However, traditional classification approaches for traffic to higher-level application mapping, such as those based on port or payload analysis, are highly inaccurate for many emerging applications and hence useless in actual networks. This paper presents a recurrence plot-based traffic classification approach based on the analysis of non-stationary "hidden" transition patterns of IP traffic flows. Such nonlinear properties cannot be affected by payload encryption or dynamic port change and hence cannot be easily masqueraded. In performing a quantitative assessment of the above transition patterns, we used recurrence quantification analysis, a nonlinear technique widely used in many fields of science to discover the time correlations and the hidden dynamics of statistical time series. Our model proved to be effective for providing a deterministic interpretation of recurrence patterns derived by complex protocol dynamics in end-to-end traffic flows, and hence for developing qualitative and quantitative observations that can be reliably used in traffic classification.

1. Introduction

The accurate classification of traffic flows traversing a specific network perimeter and the reliable determination of the applications associated with each flow's endpoints is an essential task for network security and traffic engineering, in order to protect network resources and enforce institutional policies – i.e. limiting bandwidth for sharing of music files or gaming or detecting intrusions, malicious software, or simply new network-hungry applications which may impact the future provisioning of communication resources. The most common identification technique based on the inspection of "known port numbers" is no longer accurate because many applications no longer use fixed, predictable port numbers. The Internet Assigned Numbers Authority (IANA) assigns the well-known ports from 0 to 1023 and registers port numbers in the range from 1024 to 49151. But many applications have no IANA assigned or registered ports and only utilize 'well known' default ports. Often these ports overlap with IANA ports and an unambiguous identification is no longer possible. Furthermore some applications (e.g. passive FTP or video/voice communication) use dynamic ports unknowable in advance, and some others (e.g. systems for peer-to-peer information sharing) use a combination of dynamic port numbers, masquerading techniques and encryption to bypass firewalls and avoid detection. A more reliable technique used in many current industry products involves stateful reconstruction of session and application information from packet content. Although this technique avoids reliance on fixed port numbers, it requires the inspection of the payload of every packet and hence raises privacy issues and imposes significant processing complexity and load on the traffic identification device. It must be kept up-to-date with extensive knowledge of application and protocols, which can continuously evolve and be poorly documented, and must be powerful enough to perform concurrent analysis of a potentially large number of flows.

* Corresponding author. Tel.: +39 0812537054; fax: +39 081 676628.
E-mail addresses: fpalmier@unina.it (F. Palmieri), ufiore@unina.it (U. Fiore).

1389-1286/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2008.12.015

This approach can be difficult when dealing with proprietary protocols or encrypted traffic. The strong limitations of port-based and payload-based analysis motivate the success of new traffic classification paradigms based on the study of some flow properties that are more difficult to masquerade, such as transport layer statistics. These classification techniques rely on the fact that different applications typically have distinct behavior patterns when communicating on a network. For instance, a large file transfer using FTP would have a longer connection duration and larger average packet size than an instant messaging client sending short occasional messages to other clients. Similarly, some peer-to-peer (P2P) applications such as Bit Torrent can be distinguished from FTP data transfers because these P2P connections typically are persistent and bidirectional; FTP data transfer connections are non-persistent (a separate connection is opened for each data transfer operation) and send data only unidirectionally. Transport layer statistics such as the total number of packets sent, the ratio of the bytes sent in each direction, the duration of the connection, and the average size of the packets characterize these behaviors. The statistical flow characteristics that can be considered (eventually together) include the total number of packets, mean packet size, mean payload size excluding headers, number of bytes transferred (in each direction and combined), and mean inter-arrival time of packets. Note that analysis and comparison of the distribution of packet sizes in both directions may not always be possible, due to asymmetric routing. Accordingly, we propose a novel traffic classification scheme, particularly suitable for IP networks, based on nonlinear statistical analysis and, more precisely, on the evaluation of the non-stationary transition patterns in end-to-end traffic flow time series. That is, in the IP network environment, traffic flow features and hence the characteristics of probability distributions of their IP-layer packets change dynamically in the time domain [1,2]. It follows that, since power laws apply to changes in traffic density, the traffic flow statistical characteristics change with the phase transition patterns and that their fractal-like behaviors can be affected by the packet density and its time-variation trend [3]. In addition, a self-organizing model can be considered in order to assess the non-stationary time-variation patterns of end-to-end traffic flows [4]. The dynamic transitional patterns, that are the fractal-related characteristics of the involved traffic can be used to precisely describe some aggregated flow properties and hence can be considered as an interesting way to discriminate their characteristics and hence classify the individual flows. To study the evolution dynamics and specific non-stationary features of the traffic flowing between a couple of hosts, we used recurrence plots (RP), which are a very simple and effective tool for visualizing the variation patterns of such dynamical systems [5,6]. In addition, to obtain quantitative information associated with each generated RP, we applied recurrence quantification analysis (RQA) [7,8] that has been used for measuring the degree of non-stationarity in the above patterns. The strength of our approach is twofold. First, we base our classification strategy on the nonlinear characteristics of the traffic flows that are not affected by payload encryption or dynamic port change and hence cannot be easily masqueraded. Second, nonlinear approaches are now considered as a prominent alternative to linear time series analysis of traffic data. This is because, linear methods cannot account for all the irregular phenomena observed in the network traffic flowing end-to-end between two hosts and, if a nonlinear process underlies the involved time series, the use of linear rules to describe it is conceptually false and can lead to largely erroneous results [9]. The latter line of thought is not bidirectional, as modeling the irregularities of a traffic series by nonlinear methods does not indicate the existence of nonlinearities either in time series or in the underlying process generating them. As such, the ability to identify the nature of the series is essential in improving the understanding of the process involved and in providing an, as accurate as possible, approximation of complex traffic data structures such internet traffic flows. Our approach also differentiates from the majority of the statistic-based classification schemes by its peculiar theoretical perspective: we chose a method (recurrence analysis) that does not make any specific assumption on the mathematical structure of data, does not rely on assumptions of stationarity and does not need to consider the studied traffic data as the output of a linear dynamic system. We demonstrated the possibility of a pure operational use of concepts and techniques derived by complex systems dynamics for developing deterministic qualitative and quantitative observations that can be reliably used in traffic classification. To the best of our knowledge, this is a first attempt at classifying traffic by leveraging only upon the nonlinear properties of network dynamics.

2. Related work

The idea of using statistical properties to classify traffic flows, or at least to model their behavior, is not new. Some pioneering works by Paxson et al. on Internet traffic characterization [10,11] focus on the relationship between the observed statistical flow properties and the associated application/protocols. However, such works show that analytical models based on random variables such as packet length, inter-arrival times and flow-duration can be suitable to express the behavior of only a few protocols and do not make any attempt to classify flows according to application layer protocol features. Other works that have evidenced the relationship between the class of traffic and its observed statistical properties include those due to Dewes et al. [12], making effective use of the packet-size profile of particular applications, and to Claffy [13] that observes that DNS traffic is easily identifiable using the joint-distribution of flow-duration and the number of packets transferred. A trained approach for class discrimination has been proposed with a supervised machine learning technique by Moore et al. [14]. Although based on full and deterministic payload analysis, Moore et al. [15] also try to identify classes of traffic, instead of focusing on the classification of specific application layer protocols. The above work has been extended by Auld et al. [16] who proposed a supervised machine learning approach based on a Bayesian trained neural network. In contrast, McGregor