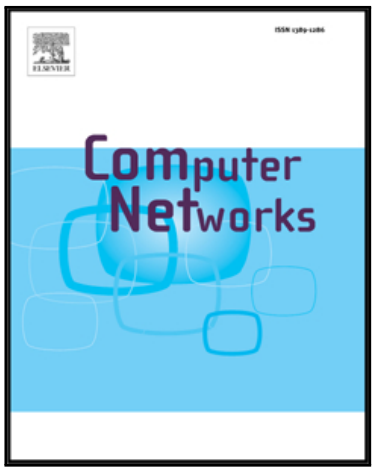

Internet Traffic Classification Based on Expanding Vector of Flow

Lei Ding, Jun Liu, Tao Qin, Haifei Li

PII: S1389-1286(17)30366-3
DOI: 10.1016/j.comnet.2017.09.015
Reference: COMPNW 6314

To appear in: *Computer Networks*

Received date: 22 March 2017
Revised date: 24 September 2017
Accepted date: 27 September 2017

Please cite this article as: Lei Ding, Jun Liu, Tao Qin, Haifei Li, Internet Traffic Classification Based on Expanding Vector of Flow, *Computer Networks* (2017), doi: 10.1016/j.comnet.2017.09.015

Internet Traffic Classification Based on Expanding Vector of Flow☆

Lei Ding[a], Jun Liu[a], Tao Qin[a], Haifei Li[b]

[a]MOEKLINNS Lab, Department of Computer Science and Technology, Xi'an Jiaotong University, 710049, China.
[b]Department of Computer Science, Union University, Jackson, TN 38305 USA.

Abstract

To reduce the number of packets used in categorizing flows, we propose a new traffic classification method by investigating the relationships between flows instead of considering them individually. Based on the flow identities, we introduce seven types of relationships for a flow and a further Expanding Vector (EV) by searching relevant flows in a particular time window. The proposed Traffic Classification method based on Expanding Vector (TCEV) does not require an inspection of the detailed flow properties, and thus, it can be conducted with a linear complexity of the flow number. The experiments performed on real-world traffic data verify that our method (1) attains as high a performance as the representative methods, while significantly reducing the number of processed packets; (2) is robust against packet loss and the absence of flow direction; and (3) is capable of reaching higher accuracy in the recognition of TCP mice flows.

Keywords: Traffic Classification, Flow Relationship, Packet Loss, Mice Flow.

1. Introduction

1.1. Background

Traffic classification, which associates packets or flows to the generating application, is an essential task in Internet management. It helps in the allocation, control, and monitoring of the usage of resources in networks[1], and plays an important role within a quality-of-service (QoS) enabled network and in the field of network security.

Methods of traffic classification are categorized using three techniques: port-based, payload-based, and behavior-based. In the early stage of Internet development, port numbers used for classifying applications are either assigned by or registered to the Internet Assigned Numbers Authority (IANA)[2]. However, as regards the continuous increase in registered numbers, as well as the use of dynamic ports by many P2P applications and the emergence of tunneling protocol techniques, port-based methods have become less convincing. Moreover, 30%-70% of traffic flows cannot be identified[3].

Payload-based methods, also known as deep packet inspection (DPI), investigate the transferring content in packets and are capable of a high recognition accuracy. However, they also lead to low efficiency, frequent updating of matching rules, and privacy problems[1, 4]. Furthermore, an increasing number of applications encrypt their payloads, thereby making the methods unavailable.

At present, methods based on traffic behavior are widely used, and obtain effective performances with various machine learning techniques, such as Naïve Bayes, k-NN, Support Vector Machine (SVM), and Decision Tree[4, 5, 6]. In these methods, flow is often represented by features extracted from its packets, like the average packet size and the average inter-arrival time. Although the current behavior-based methods have many advantages, some weaknesses remain, especially for real-time classification. First, the extraction of many features must check the network and transport layer headers of every packet in the flow, a process that involves large computational and memory overheads. Second, certain features like maximal packet size are a posteriori, because they are calculated until the flow terminates. Third, it remains a challenge in mice flow classification because these flows possess too few packets to collect representative features, thereby decreasing the total accuracy of the classification[4].

In some works, researchers use the similarity between flows (in IP addresses and port numbers) to assist the behavior-based methods. For example, flows at the same destination side (i.e., those with the same IP address, port number, and transport layer protocol) are supposed to be generated by the same application[7]. Although this heuristic helps improve the classification performance, the methods are still based on flow features mentioned above and suffer from the shortcomings of typical behavior-based methods.

1.2. Overview of the Proposed Method

To reduce the number of processed packets required for classification, we introduce the concept of flow relationship. If

☆The research was supported in part by National Key Research and Development Program of China (2016YFB1000903), National Natural Science Foundation of China (61672419, 61532004, 61532015), Innovative Research Group of the National Natural Science Foundation of China(61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Natural Science Foundation of Shaanxi Province (2016JM6040) and the Project of China Knowledge Centre for Engineering Science and Technology.

Email addresses: lding@sei.xjtu.edu.cn (Lei Ding), liukeen@mail.xjtu.edu.cn (Jun Liu), tqin@sei.xjtu.edu.cn (Tao Qin), hli@uu.edu (Haifei Li)