Peer-to-Peer IP Traffic Classification Using Decision Tree and IP Layer Attributes

Bijan Raahemi, University of Ottawa, Canada
Ahmad Hayajneh, University of Ottawa, Canada
Peter Rabinovitch, Alcatel-Lucent Research and Innovation, Canada

ABSTRACT

We present a new approach using data-mining technique and, in particular, decision tree to classify peer-to-peer (P2P) traffic in IP networks. We captured the Internet traffic at a main gateway router, performed preprocessing on the data, selected the most significant attributes, and prepared a training-data set to which the decision-tree algorithm was applied. We built several models using a combination of various attribute sets for different ratios of P2P to non-P2P traffic in the training data. We observed that the accuracy of the model increases significantly when we include the attributes “Src IP addr” and “Dst IP addr” in building the model. By detecting communities of peers, we achieved classification accuracy of higher than 98%. Consequently, we recommend that: (a) the classification must be done within the authority of the Internet service providers (ISP) in order to detect communities of peers, and (b) the decision tree needs to be frequently trained to ensure the fairness and correctness of the classification algorithm. Our approach is based only on information in the IP layer, eliminating the privacy issues associated with deep-packet inspection.

Keywords: data mining; decision tree; IP traffic classification; peer-to-peer traffic

INTRODUCTION

Peer-to-peer (P2P) is a type of Internet application that allows a group of Internet users to communicate with each other, directly access and download files from the peers’ machine, and share computing resources (building a distributed computing environment). BitTorrent, KaZaA, LimeWire, and eDonkey are among the most commonly used P2P applications.

Some recent studies show that more than 70% of broadband traffic is peer-to-peer (Azzouna & Guillemin, 2004; Kamai & Kimura, 2003). P2P traffic and its
characteristics have changed the original assumptions under which the data networks were designed. P2P traffic is more symmetric (contrary to the assumption on which asymmetric digital services line (ADSL) was designed); and P2P traffic is less “bursty,” which makes it difficult to take advantage of statistical multiplexing (under which the original data networks were designed). Also, P2P traffic lasts longer than typical Web surfing or e-mail traffic, and packet lengths are mostly large. This keeps the queues in intermediate switches and routers more utilized and consumes more bandwidth and processing resources in the network devices. Finally, P2P traffic is less local and more spread among different autonomous systems spanning the globe.

P2P applications utilize significant bandwidth and network resources, resulting in network congestion, affecting the availability, reliability, and quality of services and consequently impairing customer satisfaction. While allocating equipment for such significant network usage, telecom carriers and service providers do not gain proportional profits from the services they offer through their infrastructure. As such, telecommunication equipment vendors and ISPs are interested in efficient solutions to classify P2P traffic for further control.

Telecom carriers and equipment providers are interested in efficient solutions to classify and filter P2P traffic for further control and regulation. Inline with the carrier’s business goals, the control policy could be blocking the P2P traffic or billing the customer accordingly. Another policy could be to let the traffic flow in the network as long as it does not congest the network and, when resources (e.g., bandwidth, processing power, etc.) are limited, then block or limit the P2P traffic generated by the end users.

In this article, we present the results of our study where we captured Internet traffic at a main gateway router, performed preprocessing on the captured data, selected the most significant attributes, and prepared a training-data set to which the decision-tree algorithm was then applied. Our approach is based only on information in the IP layer, eliminating the privacy issues associated with deep-packet inspection.

The rest of the article is organized as follows: in “Related Work,” we review the related works and techniques already proposed for classification of P2P traffic, highlighting their features and weaknesses. In “Classification Using Decision Tree And The IP Layer Attributes,” we explain the process by which we collected, labeled the traffic data, and built the models and present the simulation results of various models, analyze the simulation results, discuss our observations, and present actionable recommendations. Finally, in “Conclusion,” we present our conclusions and discuss potential future directions for this work.

RELATED WORK

In recent years, classification of P2P traffic has gained much attention in both academia and industrial-research communities, and various solutions have been developed for P2P traffic classification. A popular approach is the port-based analysis where tools such as cflowd (Cisco Inc., 2006) and Netflow (Crovella & Krishnamurthy, 2006) are configured to read the service port numbers in the TCP/UDP packet headers and compare them with the known (default) port numbers of the P2P applications (see Appendix A for default port numbers of most popular P2P applications]. The packets