Support Vector Machine Detection of Peer-to-Peer Traffic in High-Performance Routers with Packet Sampling

Francisco J. González-Castaño¹, Pedro S. Rodríguez-Hernández¹, Rafael P. Martínez-Álvarez¹, and Andrés Gómez-Tato²

¹ Departamento de Ingeniería Telemática, Universidad de Vigo, Spain
ETSI Telecomunicación, Campus, 36310 Vigo, Spain
² CESGA, Spain
{javier,pedro,rmartinez}@det.uvigo.es, agomez@cesga.es

Abstract. In this paper, we explore the possibilities of support vector machines to identify peer-to-peer (p2p) traffic in high-performance routers with packet sampling. Commercial networks limit user access bandwidth -either physically or logically-. However, in research networks there are no individual bandwidth restrictions, since this would interfere with research tasks. User behavior in research networks has changed radically with the advent of p2p multimedia file transfers: many users take advantage of the huge bandwidth (e.g. compared to domestic DSL access) to exchange movies and the like. This behavior may have a deep impact on research network utilization. Consequently, in the framework of the MOLDEIP project, we have proposed to apply support vector machine detection to identify those activities in high-performance research network routers. Due to their high port rates, those routers cannot extract the headers of all the packets that traverse them, but only a sample. The results in this paper suggest that support vector machine detection of p2p traffic in high-performance routers with packet sampling is highly successful and outperforms recent approaches like [1].

1 Introduction

In this paper, we apply support vector machines [234] to detect peer-to-peer (p2p) traffic in high-performance routers with packet sampling. Commercial networks limit user access bandwidth -either physically or logically-. However, in research networks there are no individual bandwidth restrictions, since this would interfere with research tasks. This is the case of the Galician Network of Science & Technology (RECETGA), which comprises seven campuses and many other research institutions in NW Spain, with over 100,000 users (http://www.cesga.es/en/defaultE.html).

User behavior in research networks has changed radically with the advent of p2p multimedia file transfers: many users take advantage of the huge bandwidth (e.g. compared with domestic DSL access) to exchange movies and the like. This
behavior may have a deep impact on research network utilization nowadays. We have identified similar concerns in other research networks (http://security.uchicago.edu/peer-to-peer/no_fileshare.shtml).

The increasing usage of p2p software in the last three years has raised the need of p2p detection tools. There are some related initiatives. Among them, in [5], the authors describe a methodology to identify p2p flows at the transport layer, based on connection patterns of p2p flows, regardless of packet payloads. The authors point that these patterns are more difficult to conceal than explicit flow-conveyed information. In [6], the authors propose to detect p2p traffic by identifying protocol-dependent signatures (key strings) in TCP packet payloads. This is unfeasible in high-performance routers that sample packet headers. However, it may be valid for low-to-medium performance routers. In [1], the authors present a Bayesian classifier of protocols. Although the mean classification accuracy (across all application types) is quite good, the results for p2p traffic are quite poor (an accuracy of 56%).

The MOLDEIP p2p activity detection tool we have developed (i) is independent from router performance (ii) is transparent to the users and (iii) works with sampled packet header [4].

MOLDEIP does not consider individual flows but the average activity of individual IP addresses. Consequently, there is no short-term technological dependence (first goal). Recent network activity is taken from netflow export files [7], and thus we do not consider packet payloads (unlike the approach in [6]). As in [5], the system is transparent to network users because it does not scan their machines (second goal). It works with packet headers, and it does not check them all, but only a sample (third goal).

Off-line analysis is feasible because p2p traffic is a nuisance, but it does not disable the network. Thus, a 24-hour margin to take corrective actions is acceptable.

MOLDEIP relies on a support vector machine (SVM). As far as we know, this is an original approach. Reference [8] proposes SVMs to detect anomalous traffic. However, it focuses on intrusion attacks instead of p2p traffic identification. The results in this paper suggest that support vector machine detection of p2p traffic in high-performance routers with packet sampling is highly successful and outperforms recent approaches like [1].

Specifically, we consider the problem of constructing SVM classifiers based on a given classification of \( m \) training vectors (points) in the \( n \)-dimensional space \( \mathbb{R}^n \), represented by the \( m \times n \) matrix \( D \), given the membership of each IP point \( D_i, i = 1, \ldots, m \) in one of two classes—“innocent” or “guilty”. Each point \( D_i \in \mathbb{R}^n \) is a vector representing an IP address within RECETGA’s ranges, whose components indicate particular aspects of the behavior of that IP address.

For this problem, we follow the linear programming model in [9]:

\[
\begin{align*}
\min_{w, z, y} & \quad e' y + e' z \\
\text{such that} & \quad C(Dw - e\gamma) + \frac{1}{\alpha} y \geq e \\
& \quad -z \leq w \leq z, \quad y \geq 0, \quad z \geq 0,
\end{align*}
\]

\(^1\) MOLDEIP project, Xunta de Galicia grant PGIDIT03TIC00101CT.