Chapter 9

A Scalable Approach to Network Traffic Classification for Computer Network Defense using Parallel Neural Network Classifier Architectures

Bereket M. Hambebo
Florida Institute of Technology, USA

Marco Carvalho
Florida Institute of Technology, USA

Fredric M. Ham
Florida Institute of Technology, USA

ABSTRACT

The ability to recognize network traffics plays an important role in securing modern computer network infrastructures. In this chapter, we propose a machine learning approach that is based on statistical features of communication flow between two end-points. The statistical features are then used to develop and test a Parallel Neural Network Classifier Architecture (PNNCA), which is trained to recognize specific HTTP session patterns in a controlled environment, and then used to classify general traffic. The classifier's performance and scalability measures have been compared with other neural network based approaches. The classifier's correct classification rate (CCR) is calculated to be 96%.

INTRODUCTION

Online traffic classification is an important capability for modern computer network defense infrastructures. The ability to recognize, at runtime, specific network traffic as belonging to a particular application or activity allows automated defense systems and analysis to better access risk and better contextualize other observed events, or alarms in the system. In the majority of cases, automated monitoring systems for computer defense rely on well-defined traffic features and signatures to track and identify specific communications and application activity. Such signatures may include,

DOI: 10.4018/978-1-4666-3942-3.ch009

for example, the source or destination network ports in use, or specific sequences of network commands and keywords.

While conventional methods for traffic classification in cyber defense are still useful as a first indicator, more advanced attacks are likely to disguise their activities, avoiding easily identifiable features that can be detected by an automated system. An advanced adversary may rely, for example, on end-point redirection through non-standard port numbers, or on encrypted tunnels to an internal machine in the victim's network to hide well-known payload signatures. The simple techniques may greatly impair the capacity of standard monitoring tools to properly classify the traffic.

The problem is especially relevant in the context of critical infrastructure protection systems, where port numbers, protocols and basic traffic signatures are very specific and well defined. In such contexts, there is a natural tendency to look for these known features and fail to recognize that they can be, sometimes easily, manipulated by a sophisticated adversary.

In order to mitigate this problem, an alternative approach to traffic classification may take into account second order statistical properties of the communications. For example, statistical properties of communication flow between end-points may be a good indicator or the context, or specific applications involved. In most cases, even if using non-standard port numbers or encrypted payloads, there are required steps and timings in the protocol that have to be respected to ensure a successful transaction. In general, this is especially true for the critical infrastructure protection setting, where protocol timing is often critical in networked control systems. If such statistical patterns of communications can be efficiently learned and used for online classification, they could provide a powerful support capability for advance network defense systems and security analysis.

In the context of the work, traffic classification consists of the ability to identify a type, or a class of network traffic between applications, based only on its network properties, that is, without any pre-conceived knowledge about the source and destination applications, or their host operating systems. The underlying assumption is that network traffic can be observed at any point between source and destination and an assessment about its class can be made based on its properties.

The specific classes of interest are application dependent and will drive the required set of features, and often the proposed classification strategy. For example, previous research efforts have sought to classify the end-point applications (e.g. email, web-browsers, etc.) based on their network traffic, while others, have focused on the identification of specific protocols and its variants, or on the detection of traffic anomalies for intrusion detection and network defense.

To illustrate our approach, we focus on the specific problem of classifying TCP/IP session patterns between source and destination. The goal is to create a tool that recognizes specific HTTP session patterns so they can be compared with known, or expected profiles for given services or applications. The motivation for this work comes from a cyber security application for supervisory control and data acquisition (SCADA) systems.

One intuitive approach to enable access to SCADA systems is through the deployment of protected web services and web interfaces (Zecevic, 1998). This capability enables and greatly facilitates remote access to the systems, but it also tends to create an opportunity for cyber attacks and compromises.

We propose that a general traffic classifier that can be trained to recognize access to these specific web servers and services, independent of the source, destination and network delays will help protect these networks by identifying unexpected and unauthorized sessions that could be traced to compromised proxies, or fake interface sites used for capturing passwords or other use information.

For that purpose, our approach is focused on the classification of specific web-sessions, using a pre-