Network Flow Management by Probability

Wes Hardaker
USC/Information Sciences Institute *hardaker@isi.edu*

August 19, 2022

Abstract

The continual increase in encrypted Internet traffic has brought network operational management challenges. In this paper, I discuss preliminary results of a project at USC/ISI designed to quickly predict what type of traffic may be contained within encrypted tunnels for the purpose of applying traffic prioritization techniques to low-latency and other critical flows. Our results show that basic traffic analysis can be used with a reasonable level confidence when making decisions about possible actions to take in adjusting prioritization techniques of encrypted flows.

1 Introduction

1.1 Project Background

DARPA is running a project entitled SEARCHLIGHT[1] that *explores novel approaches to analysis and QoS management of an enterprise's distributed applications overlaid on the internet.* USC/ISI's corresponding project within SEARCHLIGHT is APROPOS, where we are studying the feasibility of identifying traffic types within encrypted flows. Within APROPOS, my work has concentrated on creating statistical traffic fingerprinting profiles by analyzing labeled traffic and using these training results to fingerprint unknown flows. In this whitepaper, I outline how these simplistic fingerprints can potentially be used in decision making activities about taking actions on encrypted flows.

1.2 Technique Overview

To design a generic technique that allows for rapid training and evaluation, we accept the requirement that we can begin with known, labeled traffic to analyze. With this labeled traffic, we record the percentage of [encrypted] packet sizes seen for each label. The resulting table of size probabilities is then fed into the runtime classifier, which simply compares the known probabilities against the currently observed probabilities to make a best-guess at what traffic might be traveling through a given flow.

For an example fingerprint, consider the DNS protocol [3, 4] and the likely packet sizes it generates. The histogram size distribution of DNS, unsurprisingly, differs from that of other protocols. Figure 1 shows three histograms: a histogram of multiple protocols on the bottom, a histogram of just the DNS portion of the traffic on the top, and the histogram of all of the traffic but the DNS protocol in the middle. The important takeaway from this graph is that there are obvious features in

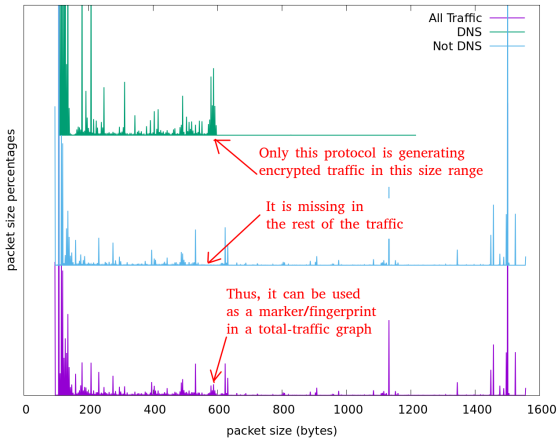
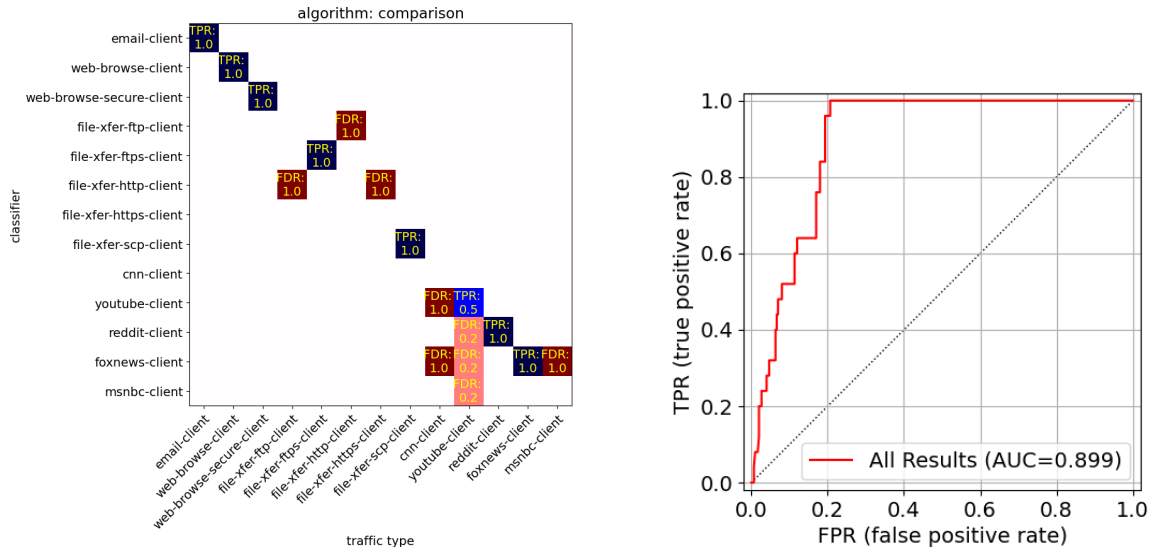


Figure 1: fingerprintability of DNS traffic vs other traffic

[1] https://www.darpa.mil/program/searchlight



(a) True Positive Rate (TPR) vs False Positive Rate (FPR) of the evaluation traffic

(b) ROC curve of resulting classification output

Figure 2: "Test classification results"

the total traffic graph that derive mostly from just the DNS traffic.

This leads to the obvious question: *can we use this technique to quickly fingerprint whether a given protocol, given a training model, exists in a larger set of traffic with some level of confidence?* And if so, *can we use this techinque to perform network management operations, such as QoS adjustments, with just a rough estimate of the encapsulated traffic contents?* Finally, *can this be done in near real-time while processing live traffic crossing an interface?*

2 Testing fingerprinting of real traffic

We implemented this plan to study the accuracy and effectiveness of the proposed technique. We gathered a collection of known traffic types traveling through an IPsec ESP [2] tunnel, and used the captures to both train and test a packet size histogram comparison algorithm. The traffic captures used come from both simulated traffic and user generated traffic. The simulated traffic tests file transfer traffic over various protocols, server-to-server email traffic, and general web browsing over both HTTP and HTTPS [1]. The user generated traffic was created by recording a user's traffic as it was directed through an IPsec tunnel while the user viewed specific web pages (*CNN, YouTube, FoxNews, Reddit, and MSNBC*).

Figure 2a shows the results of the evaluation portion of the experiment, with the detected true positive rates shown on the diagonal axis (in blue) and the false positive rates shown outside the diagonal (in red). From these results we see a number of positive success cases, indicating that a classifier build from our training data successfully identified the traffic of interest. Other classifiers (like the foxnews-client classifier), however, falsely identified other traffic incorrectly (the foxnews-client classifier incorrectly identified CNN, MSNBC and some YouTube client as foxnews). Figure 2b shows the results of a ROC curve measurement to evaluate the success of sweeping each classifier's confidence levels over the test results.

3 Applicability to Network Management

The results of the test cases from §2 in shows that *network management decisions could be made to unknown traffic when a ~90% accuracy would be considered sufficient to take an action.* Certainly, care needs to be taken with respect to what actions are safe to perform. For example, deciding to block traffic based on a

Device	Median	Mean	StdDev
Hue	8	7.81	69.96
Nest #1	6799	6799	0
Nest #2	6588	6588	0
Refrigerator	2	10.13	38.03
Pi #1	8	8	0.21
Pi #2	8	8	0.21

Table 1: Average number of packets seen in IoT TCP sessions

probability is potentially much more dangerous than simply upgrading a flow's service level (nobody minds an upgraded service). Each potential action taken based must consider if its appropriate and safe given the evaluation of a true positive vs a false positive possibility (or their inverses).

3.1 Consideration of Traffic Identifiers

Another important consideration is how to identify traffic of interest. The common 5-tuple identifier (*protocol, source address, source port, destination address, and destination port*) are commonly used for identifying a particular traffic flow. 3-tuple identifiers are also frequently used to identify general host-to-host aggregated connections. For IPsec connections, however, a combination of *protocol, source address, destination address, and SPI value* will be required instead.

One issue with using 5-tuple identifiers we discovered in our project is that many network connections are extremely short lived. As an example, Table 1 shows the average number of packets seen in TCP flows eminating from Internet of Thing (IoT) devices of various types. Using statistical inferences to flows that only contain 8 packets (or less) will be insufficient to both study and to take action on. Thus, the use of 3-tuple identifiers may be more effective in these cases.

3.2 Related Work

Although many other techniques have shown similar results in the past, often with more complex (and likely successful) algorithms, our technique and implementation is designed to minimize the mathematical operations to achieve operating at faster interface speeds. We do not yet have speed comparison results available yet, however.

3.3 Acknowledgments

This work is supported by DARPA through W911NF19C0058. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

References

[1] M. Belshe, R. Peon, and M. Thomson (Ed.). Hypertext Transfer Protocol Version 2 (HTTP/2). RFC 7540 (Proposed Standard), May 2015.

[2] D. Eastlake 3rd. Cryptographic Algorithm Implementation Requirements for Encapsulating Security Payload (ESP) and Authentication Header (AH). RFC 4305 (Proposed Standard), December 2005. Obsoleted by RFC 4835.

[3] P.V. Mockapetris. Domain names - concepts and facilities. RFC 1034 (Internet Standard), November 1987. Updated by RFCs 1101, 1183, 1348, 1876, 1982, 2065, 2181, 2308, 2535, 4033, 4034, 4035, 4343, 4035, 4592, 5936, 8020.

[4] P.V. Mockapetris. Domain names - implementation and specification. RFC 1035 (Internet Standard), November 1987. Updated by RFCs 1101, 1183, 1348, 1876, 1982, 1995, 1996, 2065, 2136, 2181, 2137, 2308, 2535, 2673, 2845, 3425, 3658, 4033, 4034, 4035, 4343, 5936, 5966, 6604, 7766.