Internet Archeology: Estimating Individual Application Trends in Incomplete Historic Traffic Traces

Sebastian Zander[1], Nigel Williams[1], Grenville Armitage[1]

Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology, Melbourne, Australia
{szander,niwilliams,garmitage}@swin.edu.au

Abstract. Public traffic traces are often obfuscated for privacy reasons, leaving network historians with only port numbers from which to identify past application traffic trends. However, it is misleading to make assumptions simply based on default port numbers for many applications. Traffic classification based on machine learning could provide a solution. By training a classifier using representative traffic samples, we can differentiate between distinct, but possibly similar, applications in previously anonymised trace files. Using popular peer-to-peer and online game applications as examples, we show that their traffic flows can be separated after-the-fact without using port numbers or packet payload. We also address how to obtain negative training examples, propose an approach that works with any existing machine-learning algorithm, and present a preliminary evaluation based on real traffic data.

1 Extended Abstract

This extended abstract outlines the methodology and several key findings of [1]. Public traffic traces are often obfuscated for privacy reasons, leaving network historians with only port numbers from which to identify past application traffic trends. With an ever-increasing number of network applications, extensive use of network address translation (NAT) and dynamic port allocation, port-based identification is fast becoming ineffectual. The most prominent examples are peer-to-peer file sharing (p2p) applications, of which a significant amount of traffic is found on non-default ports. Traffic classification based on machine learning (ML) could provide a solution. ML algorithms can be trained on data that describes each application by packet-payload independent characteristics (packet length, inter-arrival time distributions), and as such are ideal for use with anonymised traces. By training a classifier using representative traffic samples, we can classify (and differentiate between) distinct, though possibly characteristically similar, applications of interest.

While previous work has focussed on classifying flows into application types, such as interactive and non-interactive, we attempt to classify flows into specific applications. We use a hand-classified dataset of representative traffic for the

[1] This paper has been made possible in part by a grant from the Cisco University Research Program Fund at Community Foundation Silicon Valley.

applications of interest (positive examples) and two historic anonymised traces. First, we evaluate if different applications can be separated purely by packet level and flow level statistics. For the historic traces we construct classes based on default ports. This approach is justified by previous analysis, which suggests that although a significant amount of flows use non-default ports the majority still use default ports. We find that different applications can be separated with more than 90% accuracy.

When training a classifier on the hand-classified dataset and testing on the historic datasets we find accuracy in the range of 60-80%. It seems that the representativeness of our hand-classified dataset is limited by a lack of size. When generating the training dataset from the historic trace classes or from combining hand-classified and historic data accuracies are much higher, ranging between 80-90%. This suggests that applications can be separated, but any dataset representing an application of interest must cover a large range of characteristics.

As any given historic trace will contain an arbitrary mixture of applications, when building a classifier the applications not of interest must also be represented in training data (negative examples class). Since an effective solution for building an accurate classifier in the absence of negative examples has yet to be proven, we propose an approach that obtains negative examples from the historic traces. An advantage of this is that our classifier is tuned to the actual problem space. We first build and test a classifier trained on classes for the applications of interest and all port numbers of the historic trace with a significant amount of flows. We then determine a 'crossover ratio' between the applications of interest and traffic on all other ports. Ports showing a high crossover with a given application are aggregated into the positive training examples. Ports that do not have significant crossover are included into a single negative examples class. Examination of the resulting aggregated port number distributions for our applications found them not only broadly consistent across the different traces but also with prior empirical results. We then use the positive and negative training examples to build a classifier and estimate application usage within the entire historic trace files. We find significant amounts of p2p and game traffic on non-default ports, with each application showing a similar pattern of non-default port usage across the different traces.

We have outlined our technique for estimating traffic trends in historic anonymised traffic traces without using port numbers or payload information. Our initial results show that specific applications can be separated with high accuracy, and that a classifier trained on representative application data can accurately classify traffic flows in historic traces and detect a significant amount of traffic that could not be detected purely based on port numbers. We also proposed an approach to overcome the problem of missing negative examples. Although this approach shows very promising results it has a number of limitations and further research is required.

2 References

1. S. Zander, N. Williams, G. Armitage, "Internet Archeology: Estimating Individual Application Trends in Incomplete Historic Traffic Traces", Technical Report, http://caia.swin.edu.au/urp/dstc/dstc-papers.html, Centre for Advanced Internet Architectures, Swinburne University of Technology, Melbourne, Australia, Feb 2006.