LAYSEE: Learn-As-You-SEE Traffic Classifier

Alok Tongaonkar, Marios Iliofotou, Ram Keralapura
Narus Inc., Sunnyvale, USA
Email: {alok, miliofotou, rkeralapura}@narus.com

***Abstract*—The ability to classify all traffic that traverses a network is a critical aspect of network management. Signature based traffic classifiers are widely used to provide that capability. The state of the art classifiers rely on static, manual, and tedious approach of protocol reverse engineering to obtain signatures. However, the explosion of never-seen-before applications on the internet has resulted in a drastic reduction in the effectiveness of such systems. To overcome these limitations, we have developed a novel system, called Learn-As-You-SEE (LAYSEE), that aims to provide dynamic, automated, and exhaustive application identification. Our system automatically extracts signatures from network traffic by leveraging the benefits of packet content signature inference techniques and sophisticated behavioral-based analysis. These signatures are used for classifying subsequent traffic.**

I. Introduction

Total visibility and understanding of all protocols, applications, and services on the Internet is one of the fundamental necessities for network management and security applications. Traditionally, traffic classifiers have played a crucial role in providing this traffic visibility to network operators. However, recent years have seen tremendous changes in the way we use and interact with the Internet. More and more applications have become obfuscated, making the traditional means for identification less useful. Today, applications use random ports or standard ports of other applications; hence using port-based approach to detect applications leads to inaccuracies. Even the use of payload signatures does not give operators a true picture of what is going on in their network. There are 3 reasons for this: (i) Manually creating payload signatures for millions of applications is infeasible (and non-scalable), (ii) Cannot extract signatures for encrypted applications, and (iii) Almost infinite number of applications can ride over HTTP (referred to as wrapped applications). Today, a single HTTP connection can carry anything from simple text, to file transfers, to audio and video. Many undesirable applications also use HTTP connections as wrappers to conceal their activities. Identifying all these applications as "web" traffic is neither accurate nor helpful to a network/security operator. The actual application that is originating the traffic in a network must be identified. More over, with the advent of Web 2.0, sophisticated encryption techniques, and the ease of developing new applications (for example, today, more than 1 Million applications exist on Android, iOS, and Facebook combined), the amount of "unknown/misclassified" traffic is bound to skyrocket. As a case in point, according to the Lobster project, the amount of Internet traffic reported as "unknown/misclassified" by state-of-the-art technologies increased from 37% in 2004 to 69% in 2008, implying that specialized technologies in traffic classification are unable to cope with new applications.

To overcome these limitations, we developed a system, called Learn-As-You-SEE (LAYSEE), that takes protocol detection to the next level by providing the ability to move from a static, manual, and tedious approach of protocol reverse engineering to dynamic, automated, and exhaustive protocol identification. The main benefits of this framework are: (1) It can extract signatures for (known and unknown) protocols in a completely automated way using both payload-content and statistical-inference techniques; (2) It can classify incoming flows with high accuracy, and very low false-positives and false-negatives; (3) The system is robust to routing asymmetry (i.e., the input to our system can capture only one direction of a bidirectional communication). In other words, all the signatures and classification methodology utilizes only one direction of a bidirectional flow; (4) It can accomplish all of the above irrespective of whether the protocols are disguised (e.g., DNS running on port 5434, etc.) and/or encrypted (e.g., HTTPS, SSH, SSL, etc.).

We have developed LAYSEE as a fully functional prototype. Some of the key features of LAYSEE are:

- **DISTRIBUTED ARCHITECTURE**: LAYSEE incorporates an architecture designed to be scalable and modular from the ground-up. The platform provides all the necessary functionality to move the data in the system from one module to another in the distributed LAYSEE system, and also the ability to plug-in new analytical engines with almost no additional effort. The platform uses industry-standard design paradigms including REST interfaces and XML DOM.
- **ADVANCED SIGNATURE GENERATION ALGORITHMS**: We have designed, developed and incorporated sophisticated algorithms to automatically extract signatures for a protocol based on packet content [1] and flow features [2] (finite state machines when packet payload is available and behavioral/statistical signatures when only packet headers are available). These algorithms include methods for grouping flows together, either by DNS names [3] or flow-features [4] or ports, and subsequently eliminating noise and extracting a signature from these flows;
- **SIGNATURE CLEANSING ALGORITHMS**: LAYSEE includes automated signature consolidation algorithms that can resolve conflicts, eliminate duplicates, and ensure signature co-existence.
- **OPTIMIZED CLASSIFIER ALGORITHMS**: In or-

978-1-4799-0056-5/13/$31.00 ©2013 IEEE

der to use the signatures for real-time classification of incoming flows, in LAYSEE, we have developed novel classification algorithms (both packet content based and flow feature based) that can use the new signatures that have been generated and classify all flows (traffic rates up to 1Gbps).

- **SIMPLE WEB INTERFACE**: Simple browser-based UI is provided to monitor the output of the LAYSEE system in terms of classified/unknown flows, and the discovered signatures.
- **INTEGRATED WORK BENCH**: LAYSEE provides a work bench tool, that is integrated with the UI, to modify the automatically extracted signatures. This tool allows a user to further customize the extracted signatures for her particular needs.

II. Demo System

LAYSEE consists of various modules, such as Signature Generation Module and Classifier Module, which run in their own individual JVMs. These JVMs can be distributed over multiple machines or can be hosted on a single machine. For the purpose of the demo, we will run all modules in single Virtual Machine that is configured to run Red Hat Enterprise Linux 5.3. The UI for LAYSEE allows a user to login into the system in one of the following two roles (i) Admin (ii) Supervisor. Only an Admin can configure how the system is deployed, i.e., which module runs on which host. All other functionalities are available to both Admin and Supervisor users.

The system does not have any signatures at the start. We will provide pcaps extracted from real-world traces to our system. Initially, all the flows will be unknown (as there are no signatures). Over time, LAYSEE will learn new signatures for the applications present in the traces and start classifying traffic using these signatures. We will show that the system quickly reaches a point where unknown traffic is only a small fraction of the total traffic. We will show the payload content based and behavioral signatures that are automatically extracted by our system from the traffic in the network. The users can edit and submit these signatures into the system.

References

[1] A. Tongaonkar, R. Keralapura, and A. Nucci, "Sane: Self adapting network monitoring," in *Technical Report*, 2012.

[2] G. Xie, M. Iliofotou, R. Keralapura, A. Nucci, and M. Faloutsos, "Subflow: Towards practical flow-level traffic classification," in *IEEE INFOCOM (mini-conference)*, 2012.

[3] I. Bermudez, M. Mellia, M. Munafo, R. Keralapura, and A. Nucci, "Dns to the rescue: Discerning content and services in a tangled web," in *Internet Measurement Conference*, 2012.

[4] L. Grimaudo, M. Mellia, E. Baralis, and R. Keralapura, "Select: Self-learning classifier for internet traffic," in *Technical Report*, 2012.