Automating Root-Cause Analysis of Network Anomalies using Frequent Itemset Mining

Ignasi Paredes-Oliva
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
iparedes@ac.upc.edu

Xenofontas Dimitropoulos
ETH Zurich
Zurich, Switzerland
fontas@tik.ee.ethz.ch

Maurizio Molina
DANTE
Cambridge, United Kingdom
maurizio.molina@dante.net

Pere Barlet-Ros
Universitat Politècnica de Catalunya (UPC)
Barcelona, Spain
pbarlet@ac.upc.edu

Daniela Brauckhoff
ETH Zurich
Zurich, Switzerland
brauckhoff@tik.ee.ethz.ch

ABSTRACT

Finding the root-cause of a network security anomaly is essential for network operators. In our recent work [1, 5], we introduced a generic technique that uses frequent itemset mining to automatically extract and summarize the traffic flows causing an anomaly. Our evaluation using two different anomaly detectors (including a commercial one) showed that our approach works surprisingly well extracting the anomalous flows in most studied cases using sampled and unsampled NetFlow traces from two networks. In this demonstration, we will showcase an open-source anomaly-extraction system based on our technique, which we integrated with a commercial anomaly detector and use in the NOC of the GÉANT network since late 2009. We will report a number of detected security anomalies and will illustrate how an operator can use our system to automatically extract and summarize anomalous flows.

Categories and Subject Descriptors: C.2.6 [Computer - Communication Networks]: Internetworking

General Terms: Design, Experimentation, Measurement, Security, Verification.

Keywords: Anomaly extraction, anomaly validation, association rules.

1. INTRODUCTION

In our recent work [1, 2], we studied the problem of precisely identifying all the traffic flows associated with an anomaly among a large set of candidate flows, during a time interval where a detector has triggered an alarm and has given some initial, but possibly incomplete, meta-data about the anomaly. We call finding these flows the anomalous flow extraction problem or simply *anomaly extraction*. At the high-level, anomaly extraction reflects the goal of gaining more information about an anomaly alarm: without complete and compactly presented meta-data, root-cause analysis, mitigation, and prevention of future similar anomalies are very hard tasks for network and security engineers.

SIGCOMM'10, August 30–September 3, 2010, New Delhi, India.
ACM 978-1-4503-0201-2/10/08.

The steps we follow for anomaly extractions are as follows: 1) a detector raises an alarm for a time interval and identifies related meta-data, such as affected IP addresses or port numbers: this provides a set of candidate anomalous flows; 2) we model a flow as an itemset and use frequent itemset mining to extract from the large set of candidate flows the top-k itemsets with the highest support. Our assumption and intuition for applying frequent itemset mining is that anomalies often result in many flows with similar characteristics, *e.g.,* common IP addresses or ports, since they have a common root-cause, like a network failure or a scripted DoS.

We implemented our anomaly extraction technique using the Apriori frequent itemset algorithm and applied it on a histogram-based anomaly detector [3] using the Kullback-Leibler (KL) distance to detect anomalies. Our results using labeled unsampled NetFlow traces from the medium-size backbone network of SWITCH showed that our approach effectively extracted the anomalous flows in all 31 analyzed cases and it triggered very few false-positive itemsets, which can be trivially filtered out by an administrator.

In our follow-up work [5], we further evaluated and improved Apriori on the GÉANT Europe-wide backbone network, this time using 1/100 sampled NetFlow traces and a commercial anomaly detection system (NetReflex by Guavus) that since fall 2009 is deployed in the network. NetReflex is based on a well-known anomaly detector [4] using Principal Component Analysis. This second evaluation leveraged DANTE's experience in manual anomaly investigation (more than one thousand of anomalies were checked previously to this work, during a benchmarking of anomaly detection tools [6]). We observed that if an anomaly is not characterized by a significant volume of flows, Apriori cannot extract it. For instance, this occurs in the case of point to point UDP floods (involving a small number of flows but a large number of packets), which happen frequently in the GÉANT network. For this reason, we extended Apriori to also compute the support of an itemset in terms of packets in addition to flows. We added to Apriori as well the capability of automatically self-adjusting some of its configuration parameters to properly select meaningful itemsets depending on the anomaly being analyzed.

To ease the use of the extended Apriori, we implemented

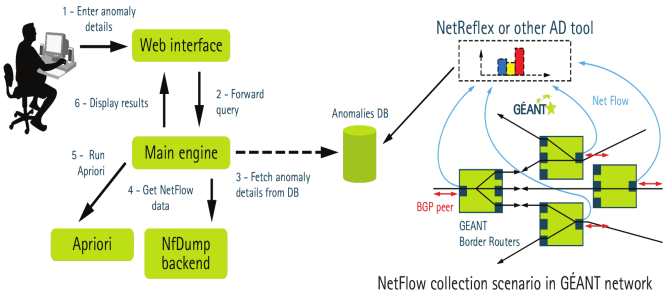

Figure 1: Anomaly extraction system architecture.

an anomaly extraction system whose architecture is depicted in Figure 1. The operator, through a GUI, can compute the frequent itemsets associated with an alarm, investigate the flows of any returned itemset, and tune the extraction parameters if needed. The GUI integrates with a back-end that stores flow records and that is based on the popular open-source tool NfDump. In addition, our system reads from a database information about an alarm (*e.g.,* the time interval and the affected traffic features) and thus can be integrated with any anomaly detection system that provides these data. Our implementation is open-source and will be made publicly available.

The extended Apriori and its GUI were tested on the GÉANT backbone network, where since fall 2009 NetReflex regularly detects security related anomalies and provides the initial meta-data that Apriori uses as input. We used the GUI to analyze 40 alarms flagged by NetReflex on Sampled NetFlow data from GÉANT. The anomaly extraction process effectively identified useful itemsets associated with a security incident in 94% of the cases. For the remaining 6% of the alarms we were not able to extract meaningful flows, which could be due to a stealthy anomaly not captured by our extraction technique or due to a false positive-alarm. In addition, for 28% of the cases with useful itemsets, the algorithm evidenced additional flows not provided by the anomaly detector. We believe that this capability of finding more flows related to an anomaly has general applicability.

2. DEMONSTRATION OUTLINE

Our demonstration will have two parts. First, we will use NetReflex to show detected anomalies on the basis of volume and IP features entropy variations [4]. It is thus suited to detect both anomalies resulting in large amounts of packets, bytes or flows (such as DoS and DDoS, both TCP and UDP based) as well as low volume anomalies but resulting in big shifts in the concentration or distribution of the number of contacted hosts and/or contacted port numbers, which happens in network and port scans. It provides fine-grained meta-data often at the level of individual IPs and port numbers for which the entropy variation algorithm detected a minimum concentration. Such fine-grained meta-data, as we describe in [2], can miss part of an anomaly or may include a large number of false-positive flows in the cases of popular port numbers or IP addresses.

The second part of the demo will use our anomaly extraction system to further analyze the anomalies detected by NetReflex and to extract the responsible flows. The GUI starts from the meta-data provided by the anomaly detection tool to select flows with a large support in terms of flows or packets and tries all possible combinations of their union [2]. In several cases (26% in our evaluation) our system finds flows related to the anomaly that the anomaly detector missed. These are particularly interesting cases. For example, the following meta-data were signaled and labeled as a port scan by NetReflex:

Table 1: List of itemsets found by our system for a particular port scan detected by NetReflex.

srcIP	dstIP	srcPort	dstPort	#flows
X.191.64.165	Y.13.137.129	55548	*	312.59K
X.191.64.165	Y.13.137.129	55548	*	270.74K
*	Y.13.137.129	3072	80	37.19K
*	Y.13.137.129	1024	80	37.28K

srcIP	dstIP	srcPort	dstPort
X.191.64.165	Y.13.137.129	55548	*

When analyzing the same anomaly using our system, the itemsets in Table 1 were found. The 1st was precisely the itemset responsible of the anomaly already flagged by NetReflex. The 2nd was another host doing a similar port scan on the same target, while the 3rd and 4th were two simultaneous DDoS on port 80 against the same target. By inspecting the raw flows with our system we observed that the DDoS was a TCP SYN flood and that it happened a few minutes after the scan. During the demo we will show other interesting anomalies like this one that will demonstrate the usefulness of this kind of information for a security engineer. The demo will be based on both live and historical data from the 18 points-of-presence of the GÉANT network.

3. ACKNOWLEDGMENTS

This work has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 216585 (INTERSECTION Project). It also has been supported by two STSM grants (no. 5202 and no. 5663) from TMA COST Action IC0703 and the Spanish Ministry of Science and Innovation under the COPERNic project (TEC2009-13252).

4. REFERENCES

[1] D. Brauckhoff et al. Anomaly extraction in backbone networks using association rules. In *Proc. of IMC*, 2009.

[2] D. Brauckhoff et al. Anomaly extraction in backbone networks using association rules. TIK-Report 309, ETH Zurich, 2009.

[3] A. Kind, M. P. Stoecklin, and X. Dimitropoulos. Histogram-based traffic anomaly detection. *IEEE TNSM*, 6(2), 2009.

[4] A. Lakhina, M. Crovella, and C. Diot. Mining anomalies using traffic feature distributions. In *Proc. of SIGCOMM*, 2005.

[5] I. Paredes-Oliva, P. Barlet-Ros, and M. Molina. Automatic validation and evidence collection of security related network anomalies. In *Proc. of PAM (Poster Session)*, 2010.

[6] W. Routly. A quantitative cross-comparative analysis of tools for anomaly detection. In *TF-CSIRT/FIRST Technical Seminar*, 2009.