Normal and Anomalous Traffic Flow Pattern Analysis for Organizational Networks

Safia Rahmat, Quamar Niyaz, Ahmad Y Javaid, and Weiqing Sun

College Of Engineering
The University of Toledo, Toledo, OH-43606, USA
{safia.rahmat, quamar.niyaz, ahmad.javaid, weiqing.sun}@utoledo.edu

***Abstract*—Traffic monitoring and analysis has become necessary to understand the nature of information flowing within an organization. This is particularly important due to the recent trend of increase in the percentage of anomalous traffic in overall organizational traffic composition. In this work, we attempt to determine the typical characteristics seen in various organizational network traffic. We use simple flow analysis methods on different datasets which include normal and anomalous traffic. Results from such an analysis can play a vital role in problems ranging from feature selection for machine learning based models to help tune the rules of an intrusion detection system (IDS). Based on the analysis of number of flows, packet size, number of packets per flow, flow duration, and protocol composition present in each dataset, we present our findings in this work.**

***Keywords*—Flow analysis, Network traffic, Normal traffic, Anomalous traffic, Monitoring**

I. Introduction

There has been a tremendous growth in the Internet traffic in recent years. It is expected that IP traffic will reach 2.3 zettabyte (10^{21} Bytes) by 2020 and the annual growth rate will be 22% [1]. Such an increase is leading to a constant bandwidth battle between various Internet applications and simultaneously contributing to increased number of security threats. Therefore, Internet traffic analysis is necessary for network security applications and efficient operation. It helps in identifying the unexpected traffic and indicates any misuse or abuse over the Internet by deducing information from patterns obtained during the analysis. To differentiate the anomalous traffic from normal traffic, it is important to know how the normal flow characteristics differ from that of anomalous traffic.

Most of the Internet traffic analysis methods require usage of machine learning techniques. Flow analysis is one of the simpler methods that can be used to determine patterns followed by the packets belonging to individual flows. Internet flow, in most general terms, is defined as a unidirectional sequence of packets which share *IP protocol, source IP address, destination IP address, source port*, and *destination port* for TCP and UDP based traffic. For ICMP based traffic, ICMP *message type* and *code* are used instead of source port and destination port. Patterns observed in traffic flows for packet size, number of packets, and duration can help in categorizing the traffic as normal or anomalous. These observed patterns can be used along with the signature-based IDSs such as Snort [2] for rules installation/update to detect anomaly or generate alarms.

Through this paper, we attempt to determine patterns obtained after flow analysis of a few datasets including three normal and one attack traffic datasets. We analyze flows for a regular network in contrast to backbone or core network. Many of the previous works discussed in Section II used core or backbone network traffic for the flow analysis. We consider all the packets from traffic traces for flow analysis instead of sampling flows.

The organization of this paper is as follows. Section II discusses a few related works for Internet traffic characterization. In Section III, we discuss the datasets used to determine patterns in traffic flows. Section IV discusses the flow analysis and results obtained. It describes various patterns observed for the anomalous and normal datasets. In Section V, we conclude our work with possible future extensions.

II. Related Work

In this section, we discuss some of the existing works. In [3], Jae-Sung Park et. al. characterized the Internet traffic collected for a high-speed Internet enabled network in Korea for two days. The flow analysis was presented for normal traffic and was based on the size and duration of flows. It showed that 45% of the packets in the traffic had a size of 40 bytes and had a duration of a few milliseconds. TCP-based flows were predominant in this work. To determine the importance of direct observation for better understanding of DDoS attacks, Mao et. al. provided an analysis of DDoS attacks using four datasets obtained from three different sources [4]. The three data sources included a commercial and a custom anomaly detection system, and backscatter dataset. After studying the attack duration distribution, they observed that overall duration for all the datasets were similar. Analysis of packet count distribution showed that long-lived attacks had more attack traffic and contributed to lower attack rate.

Anomaly detection through subspace method has been proposed by Lakhina et. al. [5]. Flow data was collected from Abilene, a major academic network for two period: four weeks in April 2013, and three weeks in December 2013. Number of flows, packets, and bytes from the sampled flows for origin and destination (OD) flows were collected in this work. This enabled examination of three different representations of sampled flows. These representations were then used for anomaly detection using subspace method. In [6], Myung-Sup Kim et. al. presented a flow-based network attack detection. To generate the dataset, they launched various attacks using available tools in a

testbed environment. Unlike the flow analysis for both normal and anomalous traffic, their method focused on DDoS attack detection. Single detection function was used to detect several attack traffic with similar patterns.

Additionally, Thompson et. al. presented a high-performance and low-cost monitoring system for traffic analysis [7]. The system was deployed on OC-3 trunks within Internet MCI's backbone and NSF-sponsored vBNS to analyse wide-area Internet traffic and characterize it in terms of packet sizes, flow duration, volume, protocol, and application. It did not categorize traffic into normal or anomalous, however, presented a general analysis such as domestic and international link traffic patterns along with traffic composition.

It is clear that different techniques have been used to classify the Internet traffic to understand its characteristics. One of our previous works dealt with feature extraction of network traffic for the use in a distributed intrusion detection system [8]. This work helped us understand the dynamics of an organizational network and motivated us for the presented work. The method proposed in this paper allows us to determine general patterns seen in anomalous and normal traffic flow.

III. Traffic Datasets Description

We used four datasets to analyse flow patterns for normal and anomalous Internet traffic: i) CAIDA Attack ii) UNIBS iii) ISCX, and iv) Home Network. The CAIDA dataset is the traffic trace of a 2007 DDoS attack that lasted roughly 40 minutes, and was used for anomalous traffic pattern analysis. The 21 GB packet header trace contains one hour of traffic [9]. The trace was split into 5 minute traces to address the analysis tool limitation. UNIBS dataset is publicly available through University of Brescia (UNIBS) [10]. It contains Internet traffic captured on the edge router of the campus network on three consecutive days using `tcpdump` [11]. The data captured was 27 GB in size and generated by 22 workstations. After anonymizing and removing the payloads from packets, reduced size is 2.7 GB.

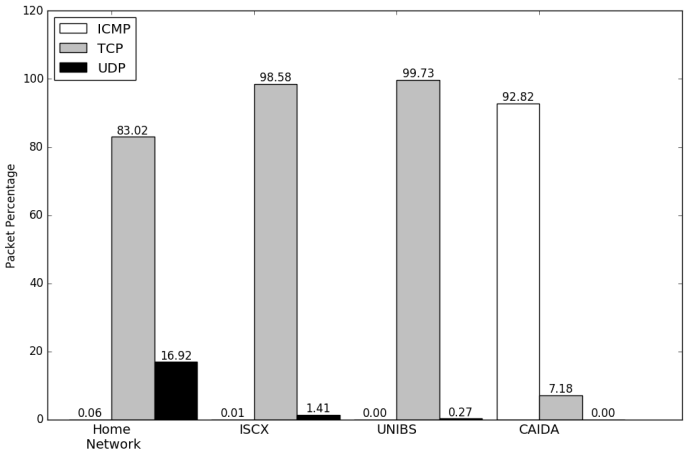


Fig. 1: TCP, UDP, and ICMP fraction in normal and attack datasets

ISCX dataset was collected in a testbed environment using a systematic approach to model real-world traffic [12]. It used profiles containing abstract representation of features or events depicting real-world behaviours that help in normal and attack traffic generation close to the real-world. It contains traffic of seven days which include normal traffic along with attack traffic of various types subsequently generated during that period. We used only the normal traffic from this dataset for our analysis as it was found that attack traffic in this dataset is targeted at the application layer and resembles normal traffic from network layer perspective. We captured the home network data from a home wireless network (HWN) that comprised 12 active nodes. We used this dataset for normal traffic analysis. The dataset contains traffic collected for 72 hours using `tcpdump` and port mirroring on a Linux system. The traffic traces contain flows for web browsing, video streaming, and on-line gaming.

IV. Flow Analysis and Discussion

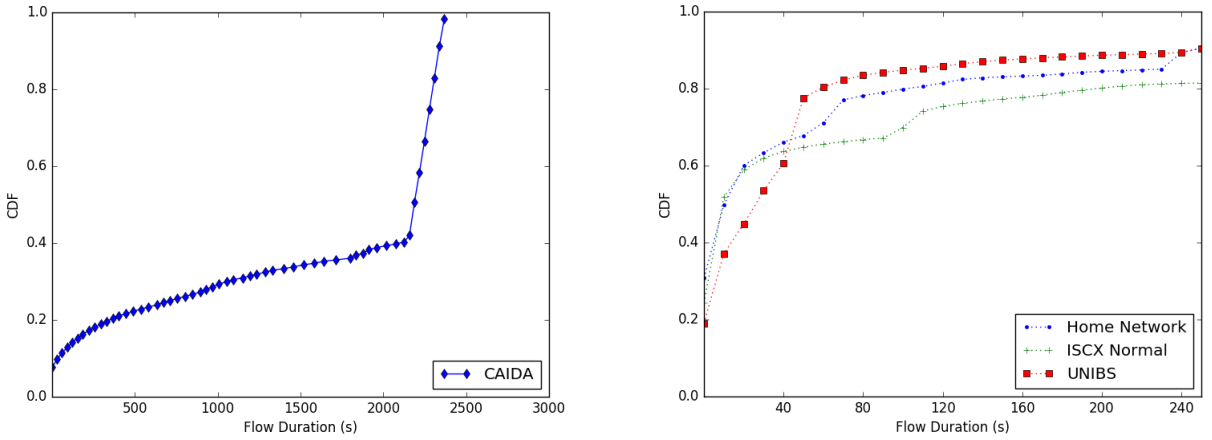
In this section, we discuss flow analysis results for normal and attack traffic datasets discussed above. First, we analyze the fraction of TCP, UDP, and ICMP based flows in each dataset. We consider flow duration, the number of packets in a flow, the number of flows in a fixed time window, and packet size for the analysis. We used ISCX, UNIBS, and HWN datasets for normal traffic analysis. For attack traffic analysis, we used the CAIDA dataset. We performed all the traffic analysis using Scapy [13], a Python-based packet manipulating, decoding and analysis tool.

Figure 1 shows the percentage of TCP, UDP, and ICMP flows in normal and attack datasets. We observe that a major portion of the traffic is contributed by TCP based flows in normal datasets. In HWN dataset, we find UDP based flows contribute 16.92% of the traffic, however, for the other two normal datasets, their contribution is insignificant. ICMP based flows are negligible in all three normal datasets since they are used as diagnostic protocol to check the availability of hosts and services in networks. In CAIDA dataset, attackers used ICMP based flows to launch DDoS, therefore their contribution is above 90%.

Figure 2 shows the cumulative distribution function (CDF) plot for flow duration of attack and normal datasets. We observe that flows in the attack dataset have longer flow durations. More than 20% of the flows have flow duration greater than 500 seconds as shown in Figure 2a. This is in contrast to what we expect of attacks having smaller flow durations due to IP spoofing. For the normal datasets, we find that 80% of UNIBS and HWN flows have duration less than 60s and 100s as shown in Figure 2b. However, ISCX dataset has a longer flow duration compared to UNIBS and HWN for 80% of its flows.

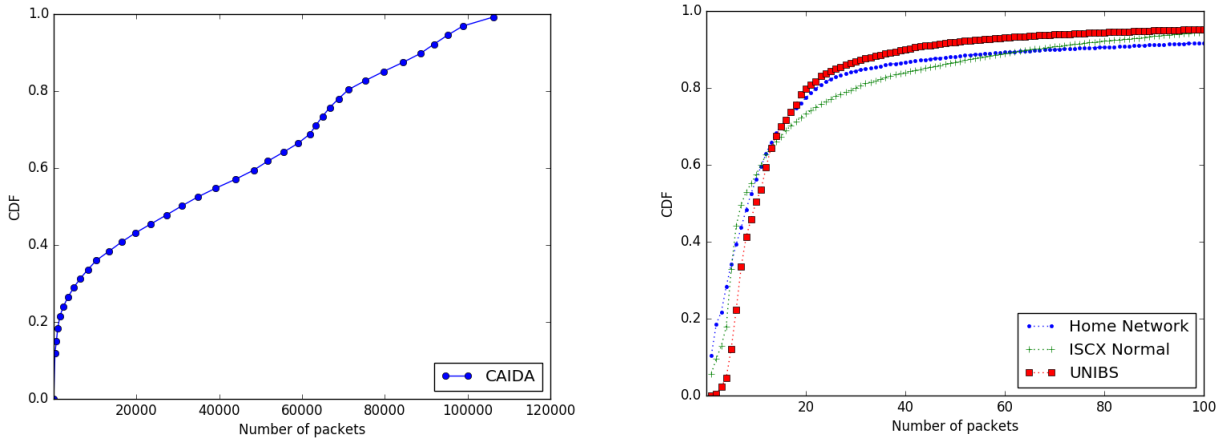
In Figure 3, we show CDF for the number of packets in a flow. In CAIDA attack dataset, more than 40% flows have at least 20000 packets or above as indicated in Figure 3a. Whereas, for normal datasets, we observe that 80% flows have less than 20 packets as shown in Figure 3b. This shows a contrast behavior in terms of number of packets in an attack flow when compared to the normal flows.

Figure 4 shows the number of flows for each destination host in a particular time window for attack and normal datasets. We consider 60s as the time window. We find a significant difference between the attack and normal datasets for the number of flows. We find that more than 20% of the traffic in CAIDA dataset have



(a) (b)

Fig. 2: Flow Duration for (a) Attack Dataset and (b) Normal Datasets



(a) (b)

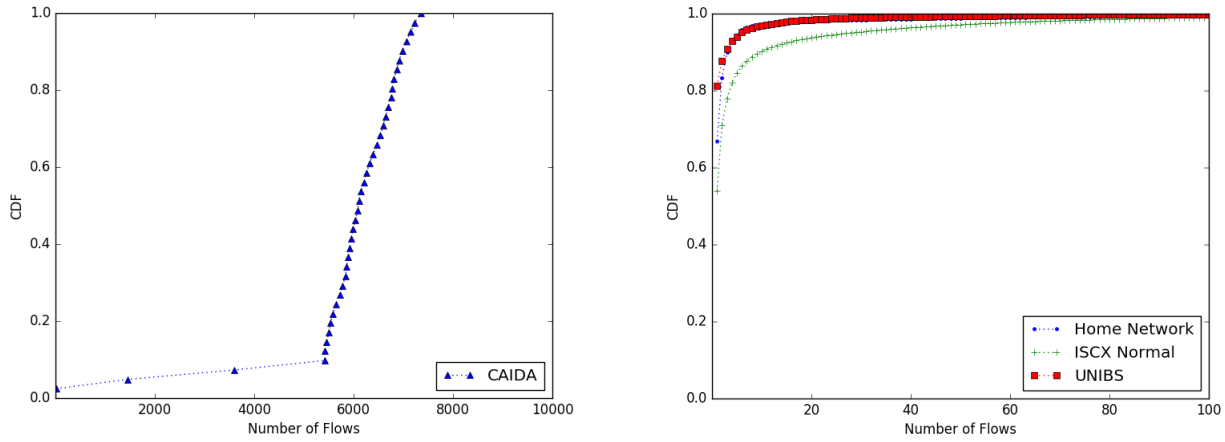
Fig. 3: Packet Distribution for (a) Attack Dataset and (b) Normal Datasets

6000 or more flows. Observing the plot for normal datasets, we found that 90% of the traffic of UNIBS and HWN datasets have 2-3 flows whereas ISCX normal dataset has around 10 flows for 90% of its traffic. This is in accordance to what we expected.

We analyze the packet size of flows for attack and normal datasets as well. We find that packet size for most of the flows in CAIDA dataset is smaller compared to normal datasets. Figure 5a shows more than 99% of the flows have packet size of around 60 bytes. In fact, most of the flows of the CAIDA dataset have a constant packet size of 60 bytes. For normal datasets, we observe that more than 30% of flows have packets of size 100 bytes or above shown in Figure 5b. The packet size is varying and includes both the IP header and payload.

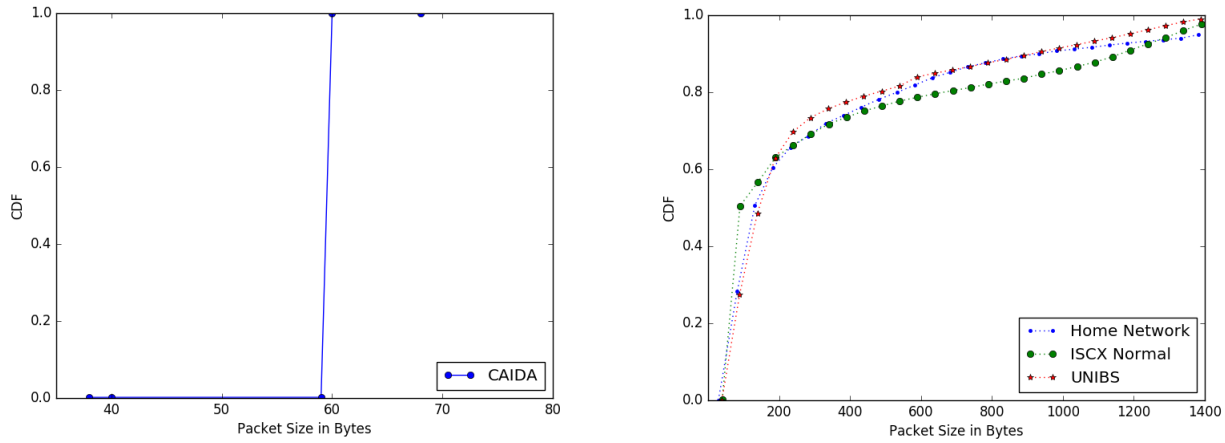
V. Conclusion

Through this work, we found patterns that distinguish normal and anomalous traffic using four datasets. We observed that normal traffic is mostly dominated by TCP-based flows. Most of the flows have small flow duration, less number of packets, and varying packet sizes. However, the attack traffic has large number of flows targeting hosts in a particular time window. The packet size is also constant for most of the flows. These flow characteristics can help us distinguish the datasets in a simple way without using complex computation intensive algorithms. As a future work, one of the primary limitations of this work, i.e., using only a single attack dataset (CAIDA) for analysis, could be addressed and different attack datasets could be used to concretely establish the difference in pattern of anomalous flow compared to normal traffic flow.



(a) (b)

Fig. 4: Number of Flows for (a) Attack Dataset and (b) Normal Datasets



(a) (b)

Fig. 5: Packet Size for (a) Attack Dataset and (b) Normal Datasets

REFERENCES

[1] Cisco VNI. http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html Accessed 24 Feb. 2017.

[2] Snort. https://www.snort.org/ Accessed 24 Feb. 2017.

[3] J. S. Park, J. Y. Lee, and S. B. Lee, "Internet Traffic Measurement and Analysis in a High Speed Network Environment: Workload and Flow Characteristics," *Journal of Communications and Networks*, vol. 2, pp. 287–296, Sept 2000.

[4] Z. M. Mao, V. Sekar, O. Spatscheck, J. van der Merwe, and R. Vasudevan, "Analyzing Large DDoS Attacks Using Multiple Data Sources," in *Proceedings of the 2006 SIGCOMM Workshop on Large-scale Attack Defense*, LSAD '06, pp. 161–168, ACM, 2006.

[5] A. Lakhina, M. Crovella, and C. Diot, "Characterization of Network-wide Anomalies in Traffic Flows," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, IMC '04, (New York, NY, USA), pp. 201–206, ACM, 2004.

[6] M.-S. Kim, H.-J. Kong, S.-C. Hong, S.-H. Chung, and J. W. Hong, "A Flow-based Method for Abnormal Network Traffic Detection," in *2004 IEEE/IFIP Network Operations and Management Symposium (IEEE Cat. No.04CH37507)*, vol. 1, pp. 599–612 Vol.1, April 2004.

[7] K. Thompson, G. J. Miller, and R. Wilder, "Wide-area Internet Traffic Patterns and Characteristics," *IEEE network*, vol. 11, no. 6, pp. 10–23, 1997.

[8] A. M. Karimi, Q. Niyaz, W. Sun, A. Y. Javaid, and V. K. Devabhaktuni, "Distributed Network Traffic Feature Extraction for a Real-time IDS," in *2016 IEEE International Conference on Electro Information Technology (EIT)*, pp. 0522–0526, May 2016.

[9] The CAIDA UCSD DDoS Attack 2007 Dataset. https://www.caida.org/data/passive/ddos-20070804_dataset.xml Accessed 24 Feb. 2017.

[10] M. Dusi, F. Gringoli, and L. Salgarelli, "Quantifying the Accuracy of the Ground Truth Associated with Internet Traffic Traces," *Computer Networks*, vol. 55, no. 5, pp. 1158 – 1167, 2011.

[11] Tcpdump. http://www.tcpdump.org/ Accessed 24 Feb. 2017.

[12] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward Developing a Systematic Approach to Generate Benchmark Datasets for Intrusion Detection," *Computers & Security*, vol. 31, no. 3, pp. 357–374, 2012.

[13] Scapy. http://www.secdev.org/projects/scapy/ Accessed 24 Feb. 2017.