A Flow-Based Entropy Characterization of a NATed Network and Its Application on Intrusion Detection

Conference Paper · March 2019

DOI: 10.1109/ICC.2019.8761747

9 authors, including:

Jorge Crichigno
University of South Carolina
80 PUBLICATIONS **484** CITATIONS
SEE PROFILE

Elie Kfoury
University of South Carolina
21 PUBLICATIONS **76** CITATIONS
SEE PROFILE

Elias Bou-Harb
University of Texas at San Antonio
85 PUBLICATIONS **785** CITATIONS
SEE PROFILE

Yasmany Prieto
University of Concepción
12 PUBLICATIONS **10** CITATIONS
SEE PROFILE

A Flow-based Entropy Characterization of a NATed Network and its Application on Intrusion Detection

J. Crichigno[1], E. Kfoury[1], E. Bou-Harb[2], N. Ghani[3], Y. Prieto[4], C. Vega[4], J. Pezoa[4], C. Huang[1], D. Torres[5]
[1]Integrated Information Technology Department, University of South Carolina, Columbia (SC), USA
[2]Cyber Threat Intelligence Laboratory, Florida Atlantic University, Boca Raton (FL), USA
[3]Electrical Engineering Department, University of South Florida, Tampa (FL), USA
[4]Electrical Engineering Department, Universidad de Concepcion, Concepcion, Chile
[5]Department of Mathematics, Northern New Mexico College, Espanola (NM), USA

***Abstract*—This paper presents a flow-based entropy characterization of a small/medium-sized campus network that uses network address translation (NAT). Although most networks follow this configuration, their entropy characterization has not been previously studied. Measurements from a production network show that the entropies of flow elements (external IP address, external port, campus IP address, campus port) and tuples have particular characteristics. Findings include: i) entropies may widely vary in the course of a day. For example, in a typical weekday, the entropies of the campus and external ports may vary from below 0.2 to above 0.8 (in a normalized entropy scale 0-1). A similar observation applies to the entropy of the campus IP address; ii) building a granular entropy characterization of the individual flow elements can help detect anomalies. Data shows that certain attacks produce entropies that deviate from the expected patterns; iii) the entropy of the 3-tuple {external IP, campus IP, campus port} is high and consistent over time, resembling the entropy of a uniform distribution's variable. A deviation from this pattern is an encouraging anomaly indicator; iv) strong negative and positive correlations exist between some entropy time-series of flow elements.**

***Keywords*—Network flows; entropy; network address translation; NetFlow.**

I. Introduction

Today's operators are facing challenges on how to efficiently protect networks. The deployment of 100 Gbps networks is now more frequently observed. At these rates, the use of inline devices such as traditional firewalls and intrusion prevention systems are hindered, in particular when large individual flows -10 Gbps and above- are present [1]. The exponential increase in network traffic adds complexity to techniques that otherwise may be viable, such as payload-based intrusion detection system (IDS). For example, Bro, one of the most reputable payload-based IDSs, has been reported to be excessively CPU intensive on high speed networks [2]. As a result, even when the number of flows is moderate, many packets may be dropped [2].

Approaches that rely on aggregated traffic, such as flow-based techniques, show a better scalability. With flow-based techniques, packets are aggregated into flows by a network device, such as a router or a switch. In this work, a flow is identified by the 5-tuple source and destination IP addresses, source and destination ports[1], and transport protocol. The flow information is collected by the device and then exported for storage and analysis. Thus, the performance impact is minimal and no additional capturing devices are needed [3]-[5].

Entropy has been used in the past to detect anomalies, without requiring payload inspection. Its use is appealing because it provides more information on flow elements (ports, addresses, tuples) than traffic volume analysis. Entropy has also been used for anomaly detection in backbones and large networks. Based on statistical packet distributions, Nychis et al. [6] presented an entropy-based anomaly approach deployed on large networks (ten of thousands of active IP addresses [6]). Other entropy-based approaches have been recently proposed for forensic analysis of DNS tunnels [7] and for classifying darkspace traffic patterns [8].

An additional complexity for anomaly detection is the use network address translation (NAT). Given the scarcity of IPv4 addresses, NAT permits end-users in a network share a single public IPv4 address. While the presence of NAT in the Internet is pervasive [9], the flow-based entropy characterization of NATed networks has not been studied previously.

A. Contribution

This paper presents a flow-based entropy characterization of a campus network. Although, as described above, similar approaches have been used for anomaly detection in different environments (see [6], [10]), this paper presents several findings that have not been previously reported. Specifically:

- The paper presents an entropy characterization of a small/medium-sized campus network. This type of network is located at the edge of the cyber-infrastructure and provides connectivity to a great portion of Internet users. Further, the characterization is based on flow counts.
- The entropies that result from locating a network behind NAT have particular characteristics. E.g., during a typical weekday, the entropies of the external and campus ports can vary from below 0.2 to above 0.8 (in a normalized entropy scale 0-1). A similar observation applies to the entropy of the campus IP address.
- Data shows that certain attacks produce entropies that deviate from the expected patterns.
- The entropy of the 3-tuple {external IP, campus IP, campus port} is high and

[1]In this article, the term *port* refers to transport-layer port.

consistent over time, resembling the entropy of a uniform distribution's variable. A deviation from this pattern is an encouraging anomaly indicator.
- Strong negative and positive correlations exist between some pairs of entropy time-series of flow elements. These relations can help enhance the detection of anomalies.

The rest of the paper is organized as follows. Section II reviews related work. Section III describes the experimental setup, methodology, and measurements. Section IV discusses the results, and Section V concludes this paper.

II. Related Work

There is a renew interest in using flow analysis to monitor and secure networks, driven by its substantial reduction in storage and CPU requirements. Hofstede et al. [4] indicate that flow-based analysis leads to data reduction of 1/2000 of the original volume, as packets are not individually processed but aggregated. Hofstede et al. [3] present flow-based detection techniques for SSH and web-application attacks. They also contrast the behavior of flows in production and lab environments. In [5], an application is developed to detect compromised hosts by using a multi-layered security detection. The first layer of detection consists of a flow-based intrusion detection, which pre-selects suspicious traffic. The second layer performs packet-based intrusion detection over the pre-filtered traffic.

Early work in entropy-based anomaly detection applied to backbone networks is presented in [11]. The proposed detection algorithms compute the entropy of IP addresses and ports. The authors showed that changes in the entropy content are indicators of massive network events. The traffic data was collected from a large ISP backbone, namely, the Swiss national research and education network.

Nychis et al. [6] describe the advantages of entropy-based analysis of multiple traffic distributions in conjunction with each other. In particular, their work suggests that for more accurate anomaly detections, IDSs should complement the use of port and IP address distributions with other behavioral features. The different distributions are constructed by counting packets, and the entropy analysis is applied to large data sets containing thousands of active IP addresses from large networks (GEANT [12], Internet2 [13], and others).

Berezinski et al. [10] provide a comparative study of entropy-based approaches for botnet-like malware detection. Their results indicate that, in addition to Shannon's entropy, Renyi's and Tsallis's entropies have very good detection performance. The authors also show that anomaly detection methods based on volume may perform poorly. Nowadays many anomalous network activities such as low-rate distributed denial-of-service (DDoS), stealth scanning, or botnet-like worm propagation and communication do not result in a substantial traffic volume change. Thus, they remain hidden in a traffic volume expressed by the number of packets, bytes, or flows.

Callegari et al. [14] also propose an anomaly detection system which measures the variation in the entropy associated with the network traffic. Similar to [10], different entropy definitions are used.

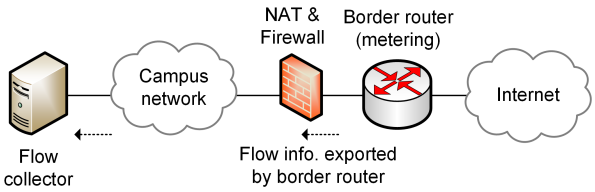

Fig. 1. Topology configuration.

Homem et al. [7] propose an entropy-based technique to detect anomalies perpetrated by encapsulating IP packets carrying malware over the DNS protocol. The proposed method quantifies the information entropy of different network protocols and their DNS tunneled equivalents, and then use such quantities to discriminate normal behavior against anomalies.

The use of NAT to conveniently hide the source of malicious behavior is discussed in [15]. By using only flow information, the authors show that machine learning algorithms may identify devices behind NAT. Tracing such devices is particularly relevant when payload inspection does not provide information because encryption is used at the application layer.

III. Methodology

A. Topology

Fig. 1 shows the flow monitoring architecture used in this work. The border router connects the campus network to the Internet Service Provider (ISP) / Internet. The NAT device translates private IP addresses to a single public IP address (campus IP). The campus network corresponds to the Northern New Mexico College (NNMC). It connects 15 buildings and different departments. There are approximately 250 faculty/staff members, 1,500 students, 20 general-purpose computer laboratories, and faculty and staff offices. Many students access the Internet via WiFi using personal devices.

The analysis presented in this paper is general and can use multiple metering points and flow directions. For simplicity of implementation in the small/medium-sized network considered here, the metering point is the single border router and the traffic direction is inbound (from the Internet to the campus network). Large campus networks with several border routers may require each border router to become a metering point and to monitor inbound and outbound traffic directions. The border router is a Cisco ASR 1000 series [16]. Flow information is ready for export when i) it is inactive for a certain time (i.e., no new packets received for the flow during the last 15 seconds); ii) the flow is long lived (active) and its duration is greater than the active timer (1 minute); and when a TCP flag indicates that the flow is terminated (i.e., FIN, RST flag are received). For each flow, the router exports the source and destination IP addresses, source and destination ports, layer-4 protocol, TCP flags observed during the connection, and connection statistics including number of packets, number

of bytes, bytes per packet, and flow duration. The information is collected by the flow collector, which implements NetFlow protocol version 9 [17]. To avoid confusion, instead of using source and destination terminology, this paper will use external and campus IP addresses, and external and campus ports. Note that from the point of view of the border router, users' flows have the same campus IP address (public IP), because of NAT. The collector organizes flow data in five-minute time slots. Data analysis is conducted for each individual time slot.

While traffic data has been collected for more than a year, the paper presents the analysis of a typical week, from Saturday, March 25, 2017 to Friday, March 31, 2017. The traffic data observed during this week is representative of the campus traffic.

B. Entropy Measures

For each time slot, the entropy of flow elements are computed. Entropy measures the randomness of a data set. The more random the data is, the more entropy it contains. The entropy of a random variable X is

$$H(X) = \sum_{i=1}^{N} p(x_i)\log_2\left(\frac{1}{p(x_i)}\right), \tag{1}$$

where $x_1, x_2, ...x_N$ is the range of values for X, and $p(x_i)$ is the probability that X takes the value x_i [18]. Among all probability distributions, the largest entropy corresponds to that of the uniform distribution: $\log_2(N_0)$. N_0 is the number of distinct x_i values present in a time slot.

This paper computes entropies as described in [6], [10]. However, this work uses flow counts rather than packet counts. Specifically, the entropy of IP addresses and ports is computed as follows. For each external (campus) IP address (port) x_i, the probability $p(x_i)$ is calculated as

$$p(x_i) = \frac{\text{Flows with } x_i \text{ as external (campus) IP addr. (port)}}{\text{Total number of flows}}. \tag{2}$$

The normalization factor is $\log_2(N_0)$, where N_0 is the number of active external (campus) IP addresses (ports) observed during the time slot.

In addition, this paper also considers the entropy of the 3-tuple {external IP, campus IP, campus port}. For a given 3-tuple x_i, the corresponding probability is calculated as

$$p(x_i) = \frac{\text{Flows with } x_i \text{ as 3-tuple}}{\text{Total number of flows}}. \tag{3}$$

The normalization factor is $\log_2(N_0)$, where N_0 is the number of active flows during the time slot.

C. Time-series Correlation and Data Statistics

For each time slot, the five normalized entropies are computed. Let $Y_{i,j}$ denote the normalized entropy of distribution i (e.g., campus IP address) observed in time slot j, and Y_i denote the time-series of normalized entropy values for distribution i. Given the $Y_i s$, the pairwise correlation coefficients between every pair of time-series vectors Y_i and $Y_{i'}$ are computed [6]:

$$r_{i,i'} = \frac{\sum_j Y_{i,j}Y_{i',j} - n\overline{Y_i}\overline{Y_{i'}}}{(n-1)\sigma_{Y_i}\sigma_{Y_{i'}}}, \tag{4}$$

where $\overline{Y_i}$ and $\overline{Y_i'}$ are the sample means of Y_i and $Y_{i'}$, σ_{Y_i} and $\sigma_{Y_{i'}}$ are the sample standard deviations of Y_i and $Y_{i'}$, and n is the number of time slots.

Additional data statistics are also computed: mean, standard deviation, maximum and minimum values. As the measured data sets from weekdays and weekend substantially differ in volume and entropy, the data statistics are separately computed for weekend and weekdays. See [19] for more information on day/night and weekday/weekend patterns in campus network traffic.

D. Time-series Rate of Change

The rate of change of the entropies is also approximated as an indicator of anomalies. A simple approximation of the derivative of Y_i with respect to time is computed as the difference between consecutive time slots j and $j+1$:

$$\Delta Y_{i,j} = Y_{i,j+1} - Y_{i,j}. \tag{5}$$

IV. Results

A. General Description

Fig. 2 shows the total traffic and entropy quantities during a typical week. The red portion of the curve represents the weekend (Saturday March 25, 2017 - Sunday March 26, 2017) and the green portion of the curve represents the weekdays (Monday March 27, 2017 - Friday March 31, 2017). The corresponding statistics for the graphs are listed in Table I. The mean traffic volume on the weekday/weekend is $1,101/168$ Mbytes in a 5-minute time slot. Thus, there is a difference of an order of magnitude between weekday and weekend. The day and night patterns are clear on weekdays, when users (students, faculty, and staff) are on campus. The peak time is slightly after 12:00 noon.

Consider the campus IP's entropy. As a small/medium-sized campus network, the number of public IP addresses is limited (less than 50 active IP addresses. One public IP address is used for NAT (users' flows) and few others are used for externally available servers). When users are on campus during weekdays, the entropy can be as low as ~0.1. The reason of the low entropy is the use of NAT, which maps users' private IP addresses to a single public IP address. During the weekday/weekend, the mean entropy is 0.224/0.345. However, note the large variation. On weekdays, the range $\mu \pm \sigma$ is $0.301 - 0.147$.

The entropy of the campus port diverges from that of the campus IP. During the day, as users connect to the network, the most popular application is browsing. Browsers use ephemeral port numbers, behaving more randomly. At peak hour, the distribution of the campus port approaches a uniform distribution and the entropy approaches ~0.9. The variation is very large during both weekdays and weekend; e.g., on weekdays, the range $\mu \pm \sigma$ is $0.823 - 0.511$.

Consider the entropy of the external IP address. The entropy is much larger than that of the campus IP. This is a reflection of users connecting to a large number of websites. However, note that the distribution is far from uniform, as entropy is well

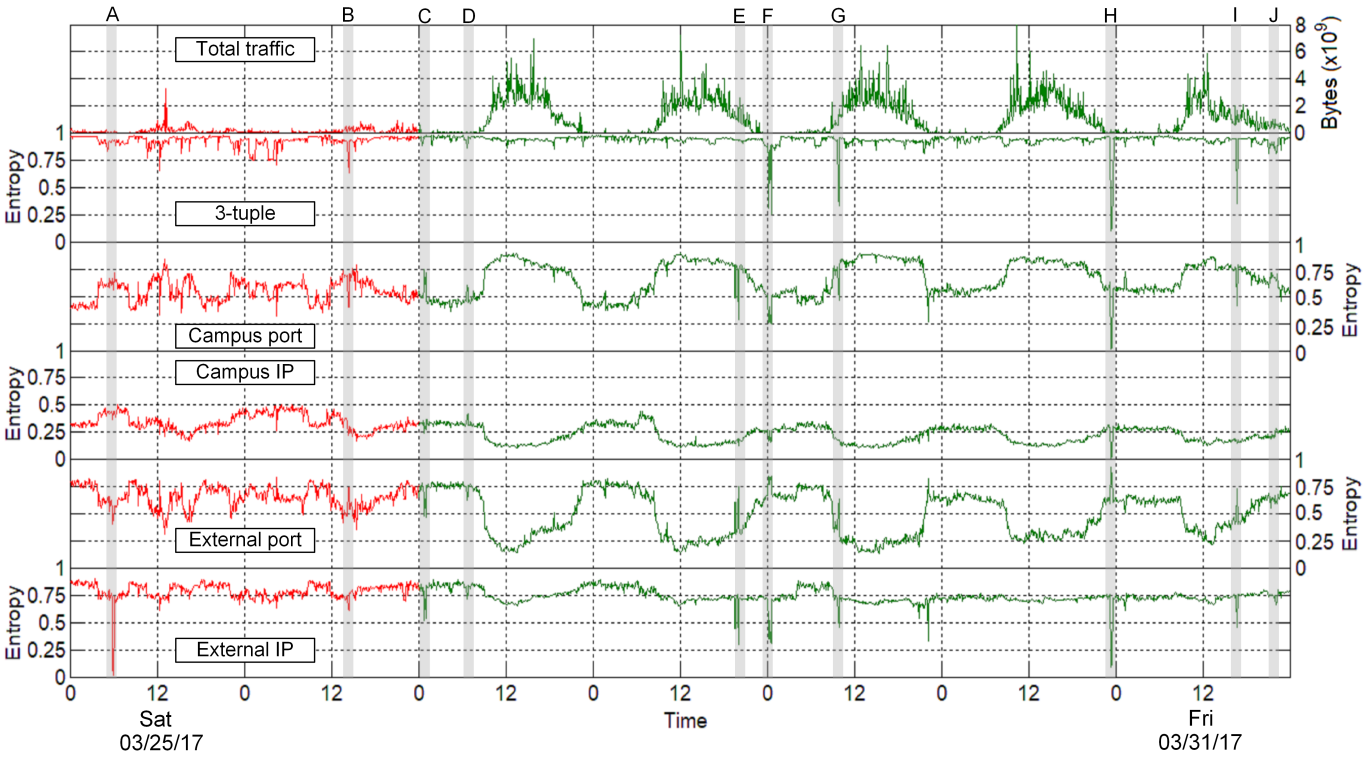

Fig. 2. Entropy time-series for the small/medium-sized network studied in this paper. Anomalies are labeled with letters A through J.

TABLE I. STATISTICAL INFORMATION FOR THE WEEK DATA OF FIG. 2.

Feature	Mean (μ)	Std (σ)	Max	Min	$\mu+\sigma$	$\mu-\sigma$
Total traffic weekday / weekend ($x10^6$ bytes)	1,101 / 168	1,254 / 224	8,875 / 3,262	5.5 / 4.1	2,355 / 392	-153 / -56
3-tuple entropy weekday/weekend (bits)	0.938 / 0.933	0.061 / 0.049	0.98 / 0.979	0.104 / 0.632	0.999 / 0.982	0.877 / 0.884
Campus IP entropy weekday/weekend (bits)	0.224 / 0.345	0.077 / 0.0769	0.44 / 0.503	0.011 / 0.172	0.301 / 0.429	0.147 / 0.268
Campus port entropy weekday/weekend (bits)	0.667 / 0.548	0.156 / 0.100	0.893 / 0.839	0.015 / 0.319	0.823 / 0.648	0.511 / 0.448
External IP entropy weekday/weekend (bits)	0.739 / 0.791	0.070 / 0.071	0.898 / 0.902	0.085 / 0.016	0.809 / 0.862	0.669 / 0.72
External port entropy weekday/weekend (bits)	0.486 / 0.656	0.204 / 0.098	0.926 / 0.836	0.138 / 0.309	0.69 / 0.754	0.282 / 0.558

below 1. This indicates that there are users connecting to the same sites/IP addresses (e.g., popular sites include YouTube, Google, Facebook). However, note that while the entropy variation is much lower than that of other flow elements discussed above, the range $\mu \pm \sigma$ is $0.809 - 0.669$, still significant.

The external port's entropy shows the largest variation among all distributions. During weekdays, at peak times, the entropy decreases to the lowest value, even below 0.2 some days (Monday, Tuesday, and Wednesday). The users' main application is browsing, thus they connect to few well-known ports (i.e., port 80, 443) which decreases the entropy. On the other hand, the distribution changes in opposite direction around midnight, when the entropy increases to the largest value, ~0.8. Note the large variation, in particular for weekdays. The range $\mu \pm \sigma$ is $0.69 - 0.282$ on weekdays.

The entropy of the 3-tuple {external IP, campus IP, campus port} is the most consistent over time with a distribution that resembles a uniform distribution. Most flows generated by users have a unique 3-tuple. Thus, the entropy is high during both weekdays and weekend, with mean values 0.938 and 0.933 respectively. Note also the low variation. In a network without anomalies, the number of flows having the same 3-tuple would be close to zero. Thus, a deviation from this situation may indicate anomalies.

B. Event Use Cases

Entropy values should be interpreted in a day/time context. For example, a value of 0.25 for the entropy of the external port is not abnormal for a weekday at noon; however, it does represent an anomaly if that value is seen at 6 AM. Along these lines, events A through J represent anomalies. Most of them deviate from the normal values that should be observed at a given day/time.

Before describing few event examples, consider Fig. 3(a). External and internal ports are shown in orange and green respectively. Under normal circumstances, most flows are unicast connections with unique 2-tuple {external IP, external port}, unique campus port (if necessary,

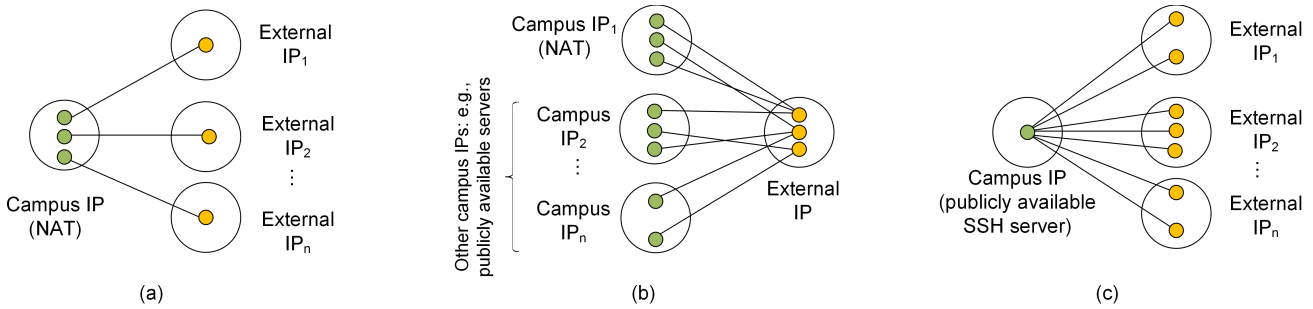

Fig. 3. (a) Typical flow pattern. External port is shown in orange, and campus port is shown in green. Flows have unique 2-tuple {external IP, external port}, unique campus port (if necessary, the NAT device performs port address translation), and common campus IP (NAT public IP address). (b) Flow pattern showing a unique external IP address generating multiple flows to multiple campus IP addresses. (c) Flow pattern where each of the multiple external IP addresses generates multiple flows to a campus IP address, single campus port (SSH).

the NAT device performs port address translation), and common campus IP (NAT public IP address shared by campus users).

1) Event A: This is a SYN flood event from a single perpetrator, thus the external IP's entropy is ~0. While not as pronounced, the external port's entropy also decreases, because the perpetrator uses a relatively small number of ports in relation to the number of flows being generated (875 different ports were observed). Most external ports are used to attempt opening up to ~1,000 connections each. Campus IP's and port's entropies do not change, because the attack targeted the entire set of campus IP addresses of the target institution (note that the campus NAT address is one of the IP addresses of the target. Other addresses are allocated to few servers providing specific services, including email and web services) (~1,000 attempted connections per campus IP address and ~230 attempted connections per campus port). The 3-tuple's entropy does not indicate anomalies because the volume of any aggregate traffic changes proportionally to the total traffic. On average, there are ~1,080 flows from the single external IP to each campus IP. Fig. 3(b) shows a simplified illustration of the flow pattern observed during event *A*.

2) Event B: This event does not correspond to an attack but to DNS activity. Approximately ~50 Amazon servers generate flows addressed to the local DNS server on campus. Each Amazon server generates between ~100 and ~300 flows to the local DNS server. Flows from a single Amazon server have identical 3-tuple {external IP, campus IP, campus port}. Thus, the 3-tuple's entropy is the best event indicator. The campus port's entropy decreases because of the increase in port 53 (DNS) activity. Under normal conditions, the number of DNS flows in a 5-minute window is typically below 1,000. During event *B*, the number of DNS flows increased to more than 6,000. The total traffic in bytes is not an indicator for this event, as the increase in DNS traffic is minimal when compared to the total traffic observed during the time slot.

3) Event G: This event occurred at 9:50 AM on Wednesday 03/29/17 and was a dictionary/brute-force attack to an SSH server (campus port 22). The total traffic does not reflect an anomaly, because SSH brute-force login attempts do not produce much volume (the traffic volume in bytes from the perpetrator was below 0.3% of the total traffic volume). However, the anomaly is captured by a drop in the 3-tuple entropy from ~0.92 to ~0.33. While a smaller change is also observed in the campus port's entropy from ~0.75 to ~0.5, this change occurs in an opposite direction to the natural entropy change for that day and time of the week (i.e., the normal behavior of the campus port's entropy during a weekday should show a steady increase until approximately noon). Similarly, although the natural tendency during this time slot is the decrease in external port's entropy, the anomaly sharply reverses this trend by increasing the entropy from ~0.35 to ~0.6. The increase in the external port's entropy occurs because the perpetrator's device opens several connections using ephemeral ports. A reader can carefully note that an external port's entropy value of ~0.6 is a valid value for a different time window, but not for the time slot of event *G*. The external IP's entropy also captures the anomaly with a decrease in entropy from ~0.75 to ~0.5.

4) Event H: From few external IP addresses, the perpetrators of event *H* opened multiple connections (using different external ports) to attempt to gain access to a single IP / port on campus. Event *H* is similar to an SSH dictionary / brute-force attack, but perpetrated by several devices (e.g., botnet). The external IP's entropy decreases as the number of flows from the perpetrators increases. Fig. 3(c) illustrates this attack.

Other events labeled as C, D, E, F, I, J show similarities to those described above. Anomalies can be detected by the rapid change in one or more entropy measures.

C. Entropy Time-series Correlation

Table II shows the correlation between the entropy time-series.

1) Total traffic: During weekdays, the total traffic is negatively correlated to the entropies of the campus IP (-0.8) and external port (-0.81). Traffic increases as a result of users accessing mostly web applications; thus the campus IP's entropy decreases because users use the same campus IP address (NAT public IP). The external port's entropy also decreases because most traffic uses http/https. The total traffic

TABLE II. CORRELATION OF ENTROPY TIME-SERIES.

	Campus IP	Campus port	External IP	External port	Total traffic
Weekday					
3-tuple	0.23	0.1	0.6	-0.02	-0.05
Campus IP		-0.85	0.6	0.89	-0.8
Campus port			-0.37	-0.98	0.78
External IP				0.45	-0.36
External port					-0.81
Weekend					
3-tuple	-0.23	-0.12	0.56	0.06	-0.03
Campus IP		0.15	-0.38	0.06	-0.38
Campus port			-0.48	-0.93	0.31
External IP				0.48	-0.05
External port					-0.39

is directly correlated to the campus port's entropy (0.78), because as users on campus access the web, their respective browsers open ephemeral ports that collectively resemble a uniform distribution. On weekend, there is a low or no correlation between the total traffic and other time-series.

2) Campus IP: On weekdays, the entropies of the campus IP and campus port are negatively correlated (-0.85). As users use the network, the campus IP's entropy decreases because users use the same campus IP address (NAT public IP). On the other hand, the campus port's entropy increases because users' browsers open ephemeral ports. The entropies of the campus IP and external port show a direct correlation (0.89): the more traffic is generated by users, the more NATed flows exist, and the lower the campus IP's entropy is. As most traffic is http/https, the external port's entropy also decreases. On weekend, there is a low or no correlation between campus IP and other time-series.

3) Campus port: On weekdays, the strongest negative correlation is between the entropies of the campus port and external port (-0.98). This relation is produced by the use of a large number of browser's ephemeral ports (each user's browser likely uses a different port number) to connect to few external ports (i.e., http/https). On weekend, there is a low or no correlation between campus port and most distributions, with the exception of external port.

4) External IP: The entropies of the external IP and external port are correlated on weekdays (0.45) and on weekend (0.48). Note that the external IP has high entropy (mean is 0.739) when compared to other flow elements. This reflects the variety of external IP addresses users connect to.

5) External port: As mentioned above, the entropies of the external port and campus IP are strongly correlated (0.89). On the other hand, the entropies of the external port and campus port are negatively correlated on weekdays (-0.98) and on weekend (-0.93).

6) 3-tuple {external IP, campus IP, campus port}: The entropy of the 3-tuple shows correlation with that of the external IP on weekdays (0.6) and on weekend (0.56). Low or no correlation is noted between the 3-tuple and other time-series.

D. Time-series Rate of Change

Fig. 4 shows the rate of change of the total traffic and entropy time-series, computed according to Eq. (5). Corresponding values are provided in Table III. The first observation here is the large rate changes in the total traffic, in particular during weekdays. Thus, traffic rate changes may not always be accurate indicators of anomalies, as they occur naturally in this small/medium-sized network. In contrast, changes in the entropy time-series from one time-slot to another (in bits / time unit) are small. The mean values for entropy changes are approximately zero.

V. CONCLUSION

This paper presents a flow-based entropy characterization of a small/medium-sized campus network that uses NAT. Measurements from a production network show that in a typical weekday, the entropies of the external and campus ports may widely vary from below 0.2 to above 0.8 (in a normalized entropy scale 0-1). Similarly, the entropy of the campus IP address may vary from 0.1 to 0.4. Despite the wide range of values, findings indicate that building a granular (small time slots) entropy characterization of flow elements facilitates anomaly detection. Data shows that certain attacks produce entropies that deviate from the expected patterns. Data also shows that the entropy of the 3-tuple {external IP, campus IP, campus port} is high and consistent over time, resembling the entropy of a uniform distribution's variable. A deviation from this pattern is an encouraging anomaly indicator.

The total traffic and the flow element entropies in a NATed network are correlated. Strong negative correlation is observed between i) campus IP's entropy and total traffic, ii) external port's entropy and campus port's entropy, iii) external port's entropy and total traffic, and iv) campus IP's entropy and campus port's entropy. On the other hand, strong positive correlation is observed between i) campus IP's entropy and external port's entropy, and ii) campus port's entropy and total traffic. Future work includes the development of anomaly detection algorithms that exploit the entropy characterization of flow elements and tuples, and their relations.

ACKNOWLEDGMENT

This work was supported by the U.S. National Science Foundation, awards 1723323, 1829698, and 1755179.

The authors acknowledge M. Husák for his suggestions.

REFERENCES

[1] "Processing of Single Stream Large Session (Elephant Flow) by the Firepower Services," Cisco Systems White Paper, Jan. 2017. [Online]. Available: https://www.cisco.com/c/en/us/support/docs/security/firepower-management-center/200420-Processing-of-Single-Stream-Large-Sessio.pdf

[2] A. Gonzalez, J. Leigh, S. Peisert, B. Tierney, A. Lee, J. Schopf, "Monitoring Big Data Transfers Over International Research Network Connections," in *IEEE International Congress on Big Data,*, Jun. 2017.

[3] R. Hofstede, A. Pras, A. Sperotto, G. Dreo, "Flow-based Compromise Detection: Lessons Learned," *IEEE Security and Privacy*, vol. 16, issue 1, Feb. 2018.

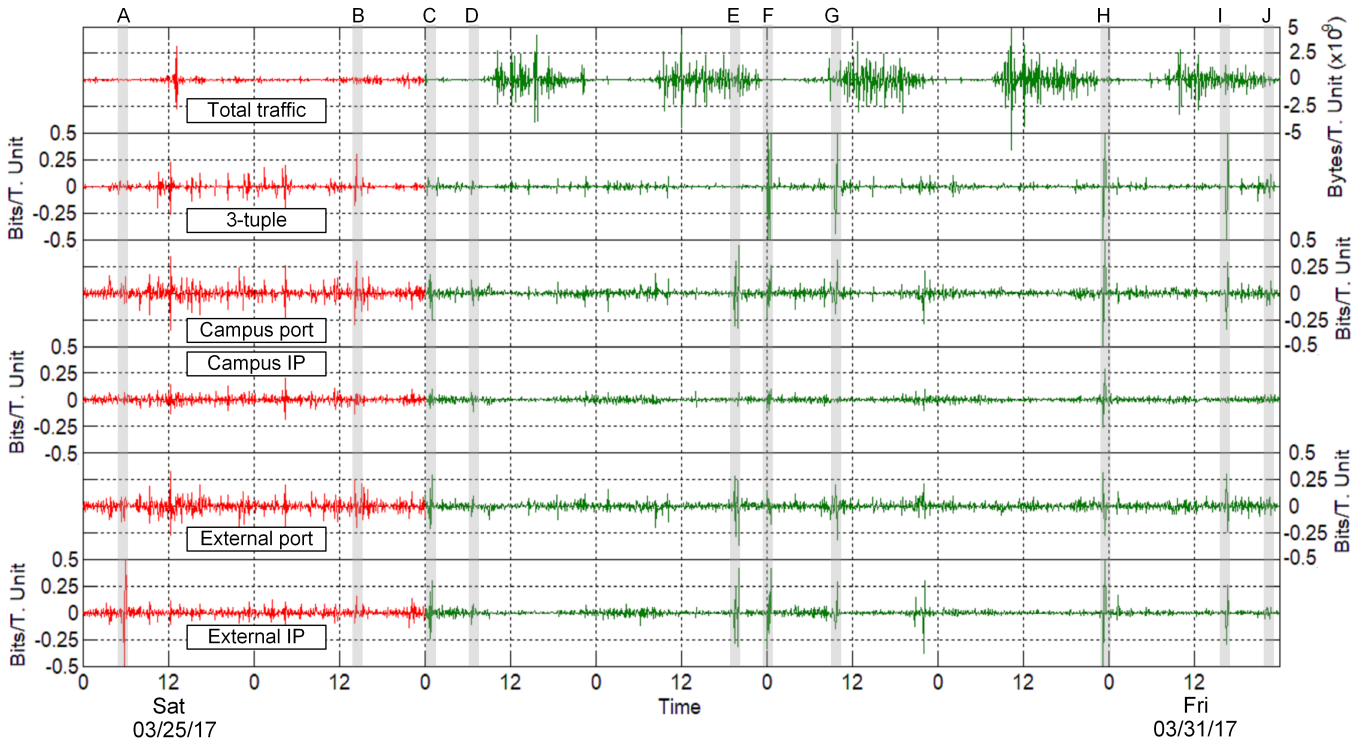

Fig. 4. Rate of change for the time-series shown in Fig. 2. Values are computed using Eq. (5).

TABLE III. STATISTICAL INFORMATION, RATE OF CHANGE OF ENTROPY TIME-SERIES.

Feature	Mean (μ)	Std (σ)	Max	Min
Total traffic weekday / weekend ($x10^6$ bytes/time unit)	0.2 / 0.18	774 / 256	5,753 / 3,162	-6,500 / -2,723
3-tuple entropy change weekday/weekend (bits/time unit)	~0 / ~0	0.057 / 0.038	0.83 / 0.3	-0.77 / -0.26
Campus IP entropy change weekday/weekend (bits/time unit)	~0 / ~0	0.021 / 0.03	0.28 / 0.2	-0.25 / -0.19
Campus port entropy change weekday/weekend (bits/time unit)	~0 / ~0	0.045 / 0.058	0.54 / 0.34	-0.53 / -0.35
External IP entropy change weekday/weekend (bits/time unit)	~0 / ~0	0.04 / 0.05	0.61 / 0.49	-0.58 / -0.62
External port entropy change weekday/weekend (bits/time unit)	~0 / ~0	0.04 / 0.056	0.32 / 0.32	-0.37 / -0.27

[4] R. Hofstede, P. Čeleda, B. Trammell, I. Drago, R. Sadre, A. Sperotto, A. Pras, "Flow Monitoring Explained: From Packet Capture to Data Analysis with NetFlow and IPFIX," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 4, 2014.

[5] M. Golling, R. Koch, R. Hofstede, "Towards Multi-layered Intrusion Detection in High-Speed Networks," *International Conference On Cyber Conflict*, Jun. 2014.

[6] G. Nychis, V. Sekar, D. Andersen, H. Kim, H. Zhang, "An Empirical Evaluation of Entropy-based Traffic Anomaly Detection," *ACM SIGCOMM Conference on Internet Measurement*, Oct. 2008.

[7] I. Homem, P. Papapetrou, S. Dosis, "Entropy-based Prediction of Network Protocols in the Forensic Analysis of DNS Tunnels," arXiv:1709.06363 [cs.CR], Sep. 2017.

[8] T. Zseby, N. Brownlee, A. King, K. Claffy, "Nightlights: Entropy-Based Metrics for Classifying Darkspace Traffic Patterns," *International Conference on Passive and Active Network Measurement*, Mar. 2014.

[9] I. Livadariu, K. Benson, A. Elmokashfi, A. Dainotti, A. Dhamdhere, "Inferring Carrier-Grade NAT Deployment in the Wild," in *IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2018.

[10] P. Berezinski, B. Jasiul, M. Szpyrka, "An Entropy-Based Network Anomaly Detection Method," *Entropy Journal*, vol. 17, no. 4, Apr. 2015.

[11] A. Wagner, B. Plattner, "Entropy Based Worm and Anomaly Detection in Fast IP Networks," *IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprise*, Jun. 2005.

[12] F. Farina, P. Szegedi, J. Sobieski, "GEANT World Testbed Facility: Federated and Distributed Testbeds as a Service Facility of GEANT," in *International Tele-traffic Congress*, Sep. 2014.

[13] Internet2. [Online]. Available: https://www.internet2.edu/

[14] C. Callegari, S. Giordano, M. Pagano, "Entropy-based Network Anomaly Detection," *IEEE International Conference on Computing, Networking and Communications (ICNC)*, Jan. 2017.

[15] Y. Gokcen, V. Foroushani, A. N. Zincir-Heywood, "Can We Identify NAT Behavior by Analyzing Traffic Flows?," in *IEEE Security and Privacy Workshops (SPW)*, May 2014.

[16] "Cisco ASR 1000 Series Aggregation Services Routers" Cisco Systems Data Sheet, Feb. 2018. [Online]. Available: https://www.cisco.com/c/en/us/products/collateral/routers/asr-1000-series-aggregation-services-routers/datasheet-c78-731632.pdf

[17] B. Claise, "Cisco Systems NetFlow Services Export Version 9," Internet Request for Comments, RFC Editor, RFC 3954, Oct. 2004. [Online]. Available: https://www.ietf.org/rfc/rfc3954.txt

[18] T. Cover, J. Thomas, "Elements of Information Theory," Wiley, 2nd Edition, 2006.

[19] P. Velan, J. Medková, T. Jirsík and P. Čeleda, "Network traffic characterisation using flow-based statistics," *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, Apr. 2016.