A FLOW-BASED ENTROPY CHARACTERIZATION OF A NATED NETWORK AND ITS APPLICATION ON INTRUSION DETECTION

Jorge Crichigno
College of Engineering and Computing
University of South Carolina
Columbia, SC

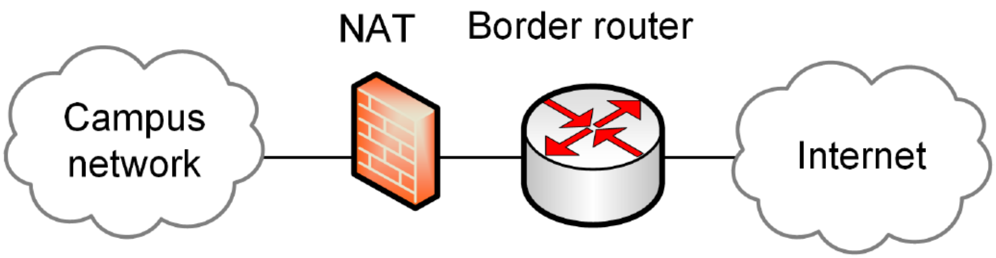
IEEE International Conference on Communications (ICC)
Shanghai, China
May 22, 2019

Agenda

- Motivation for a flow-based entropy characterization
- Overview of campus NATed networks
- Entropy of flow elements
- Results
- Conclusion

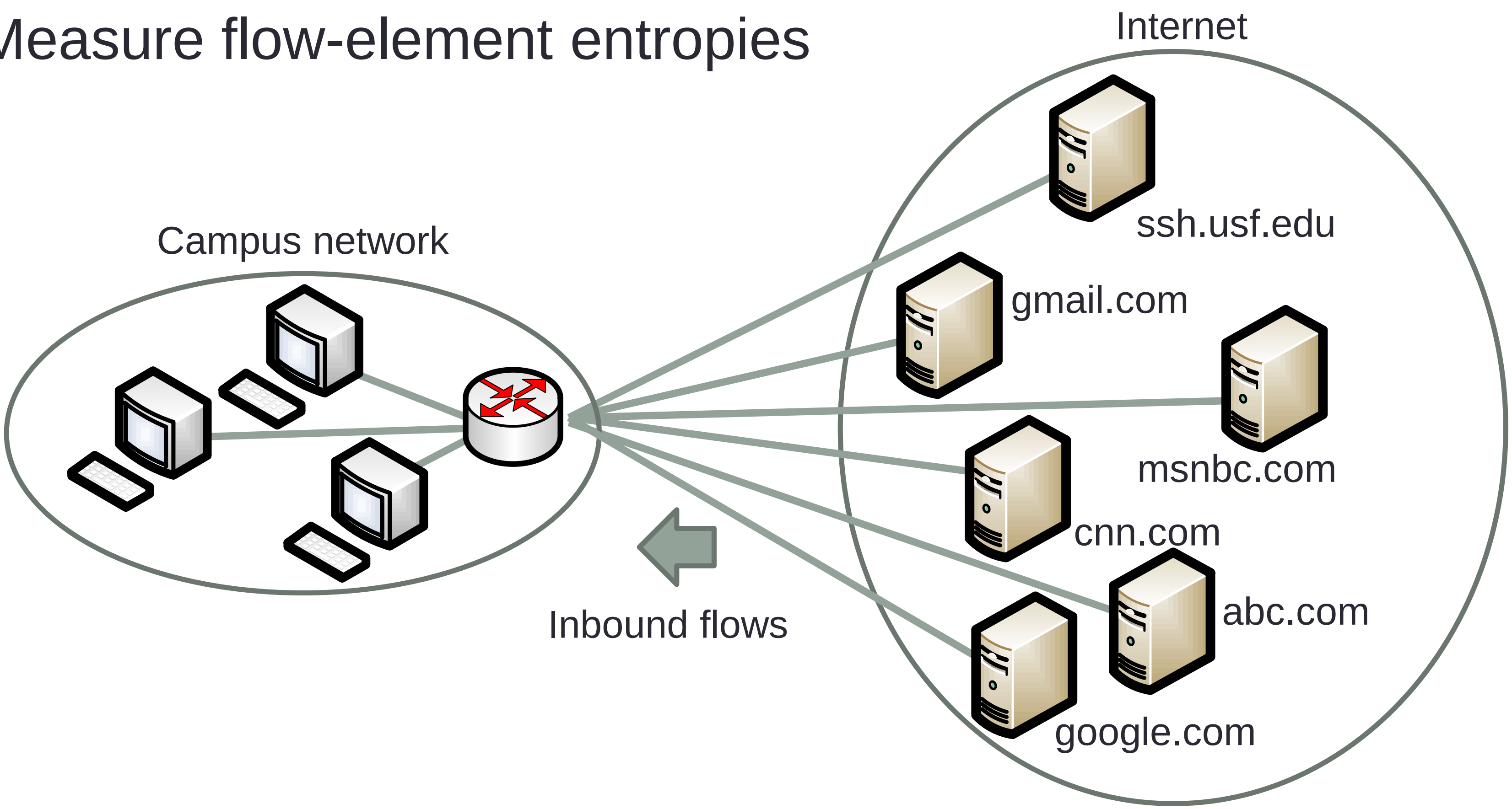
Motivation

- Most networks use Network Address Translation (NAT)
- Although NAT has been used since early 2000s, traffic behind NAT has not been characterized using entropy
- One approach for flow characterization is to measure the *randomness* or *uncertainty* of elements of a flow
- E.g., entropy of IP addresses, ports, and combinations
- Goal: characterize normal traffic behavior by using flow information only (no payload inspection) and entropy



Methodology

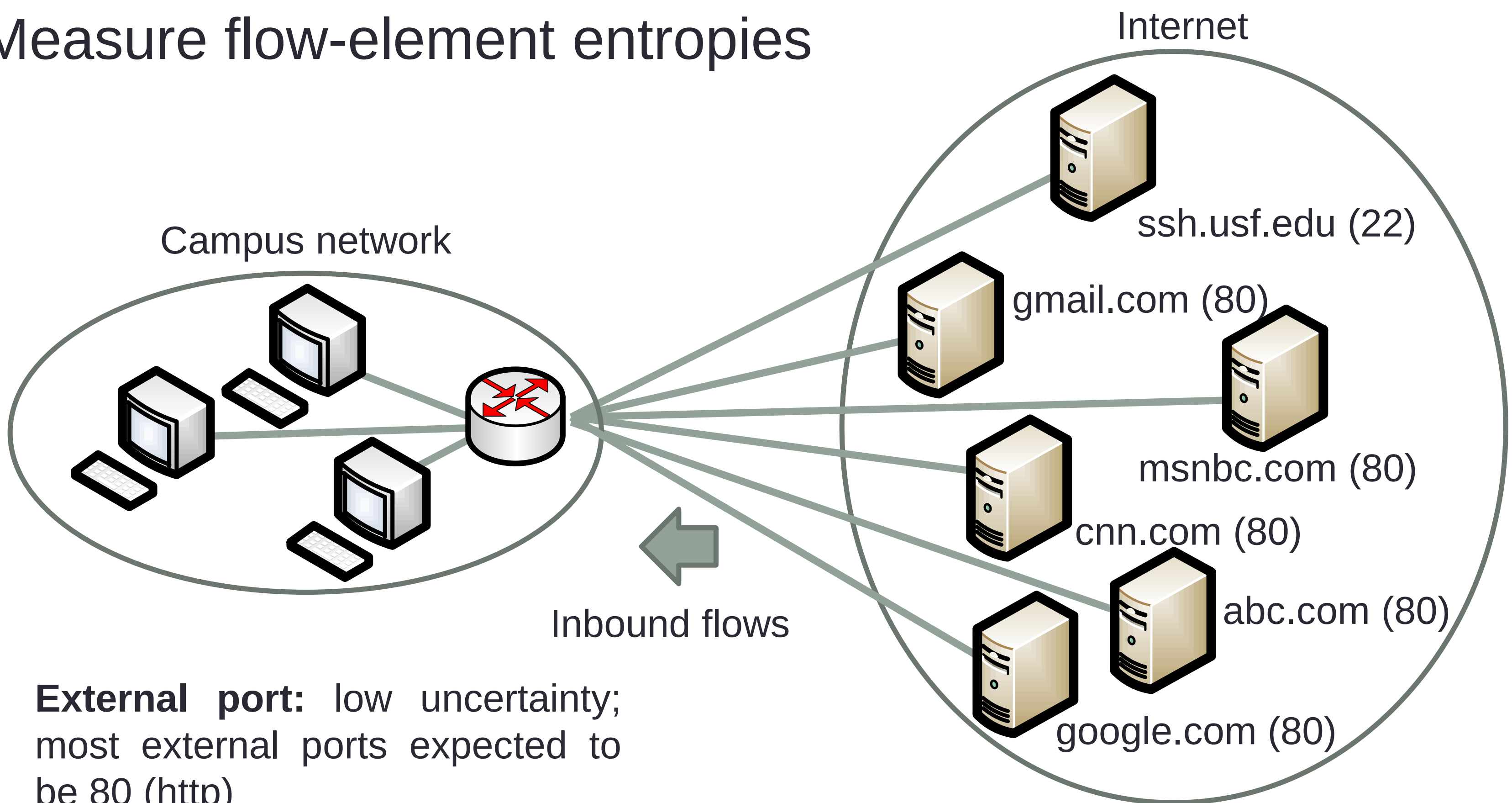
- A flow is uniquely identified by the external IP, campus IP, external port[1], campus port, protocol
- Measure flow-element entropies



1. Port refers to transport-layer port

Methodology

- A flow is uniquely identified by the external IP, campus IP, external port[1], campus port, protocol
- Measure flow-element entropies

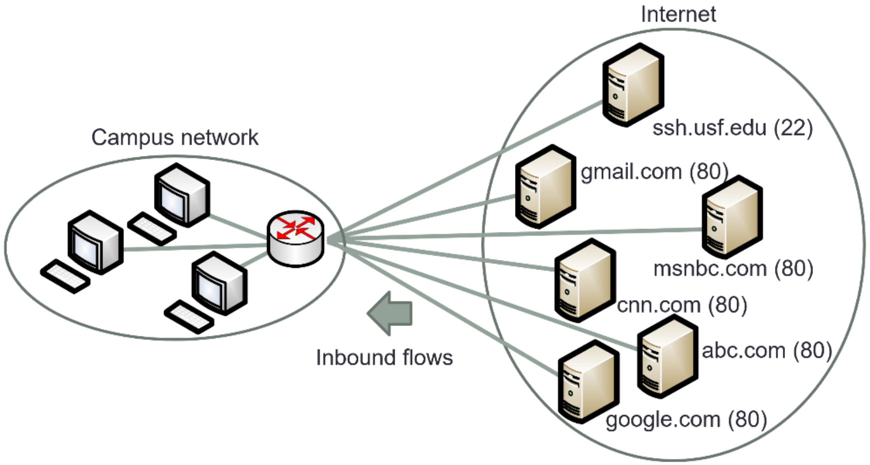


External port: low uncertainty; most external ports expected to be 80 (http)

Methodology

- Entropy provides a measure of randomness or uncertainty
- For a variable X, entropy of X = $\sum_{x \in X} p_x \log_2 \left(\frac{1}{p_x} \right)$
- For the previous port example, let *X* be the variable indicating the external port

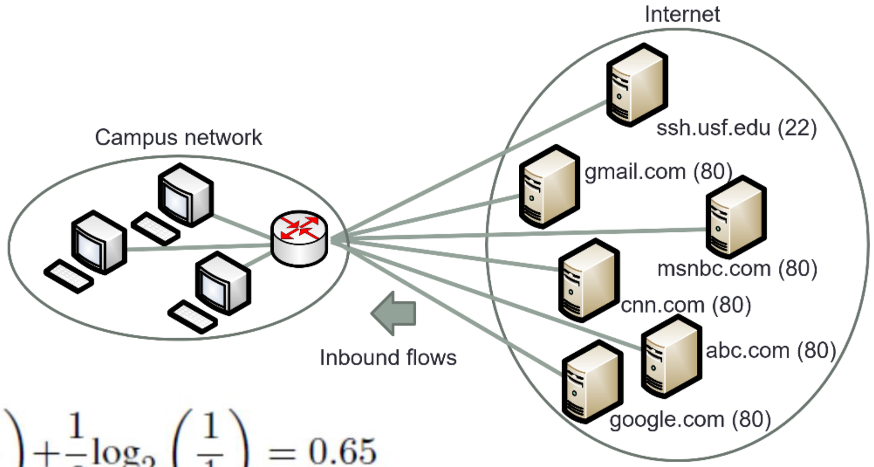
$$X = \begin{cases} 80 \text{ with probability } p_1 = \frac{5}{6} \\ 22 \text{ with probability } p_2 = \frac{1}{6} \end{cases}$$



Methodology

- Entropy provides a measure of randomness or uncertainty
- For a variable X, entropy of X = $\sum_{x \in X} p_x \log_2\left(\frac{1}{p_x}\right)$
- For the previous port example, let *X* be the variable indicating the external port

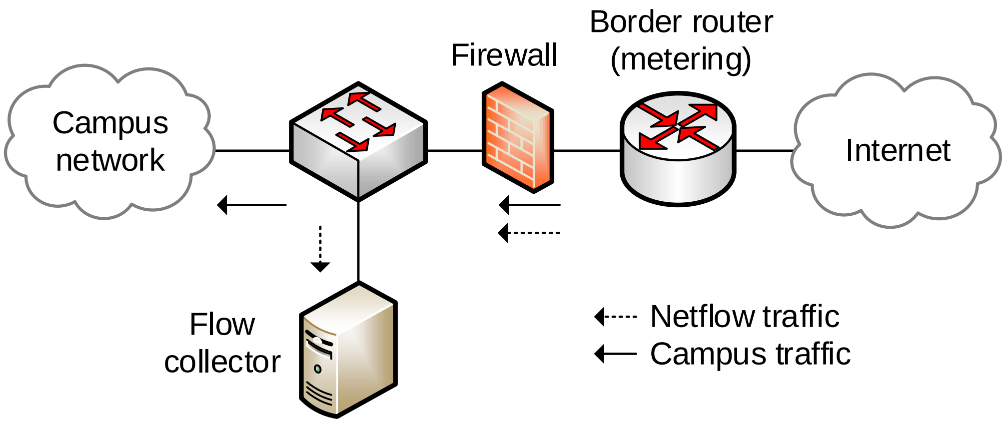
$$X = \begin{cases} 80 \text{ with probability } p_1 = \frac{5}{6} \\ 22 \text{ with probability } p_2 = \frac{1}{6} \end{cases}$$



$$\text{Entropy External Port} = \sum_{i=1}^{2} p_i \log_2\left(\frac{1}{p_i}\right) = \frac{5}{6}\log_2\left(\frac{1}{\frac{5}{6}}\right) + \frac{1}{6}\log_2\left(\frac{1}{\frac{1}{6}}\right) = 0.65$$

Methodology

- Campus network with 15 buildings
- The collector organizes flow data in five-minute time slots
- Traffic data observed during a week is representative of the campus traffic



Methodology

- The entropy of a random variable *X* is:

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

where $x_1, x_2, \ldots x_N$ is the range of values for *X*, and $p(x_i)$ is the probability that *X* takes the value x_i

- For each external (campus) IP address (port) x_i, the probability $p(x_i)$ is calculated as

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

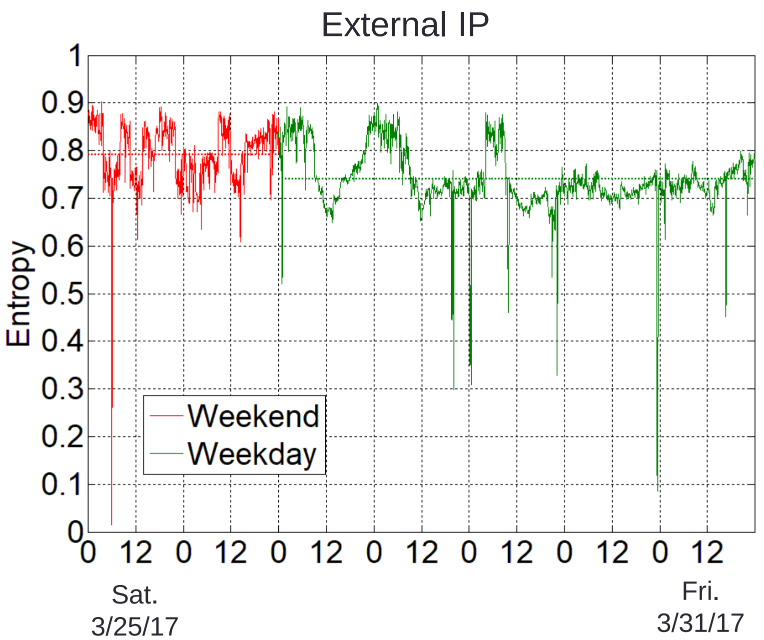
- Entropies are normalized to that of the uniform distribution

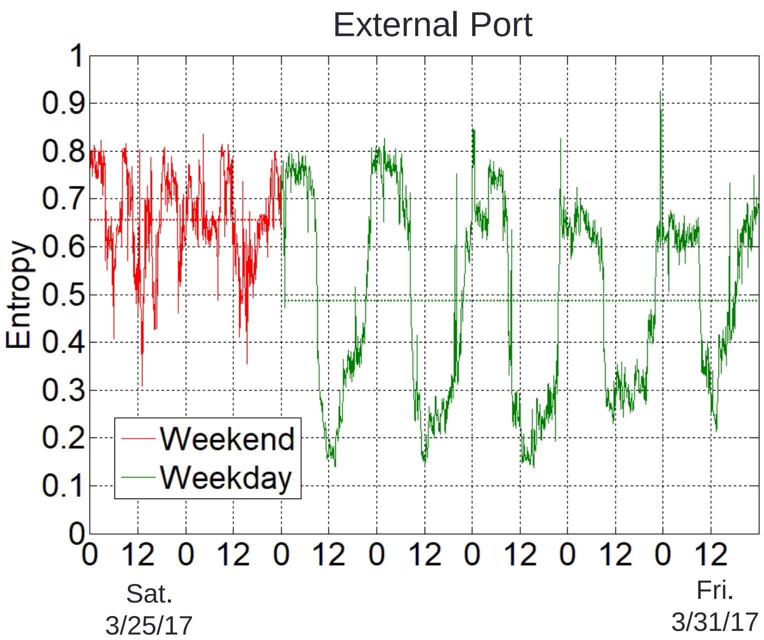
Methodology

- 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

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

Results





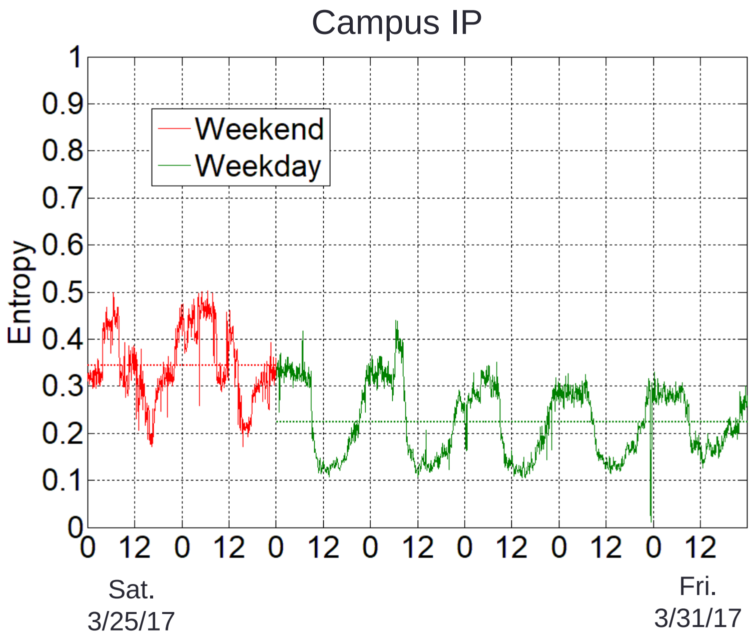
External IP

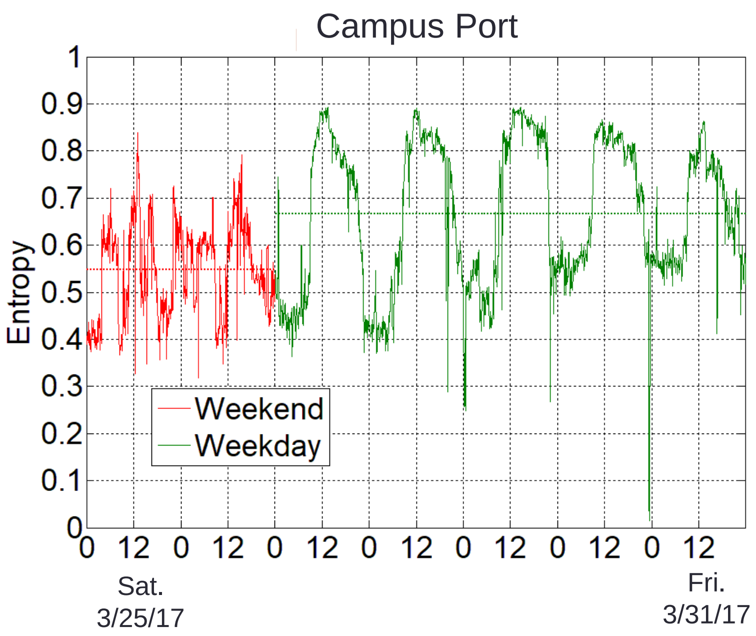
- In general, high entropy, ‘many’ external IP addresses
- External IPs dispersed in the Internet
- Abnormal low entropy points
- Entropy near zero (no uncertainty of the external IP address), or ‘very low’ level (few external IP addresses dominate the distribution)

External port

- Higher entropy during the night, weekends
- Low entropy during the day, noon
- Large volume of http flows when students are on campus (less uncertainty/entropy on external port)
- Abnormal high entropy points
- Entropy widely varies over ‘hours’ but not over very short time periods

Results





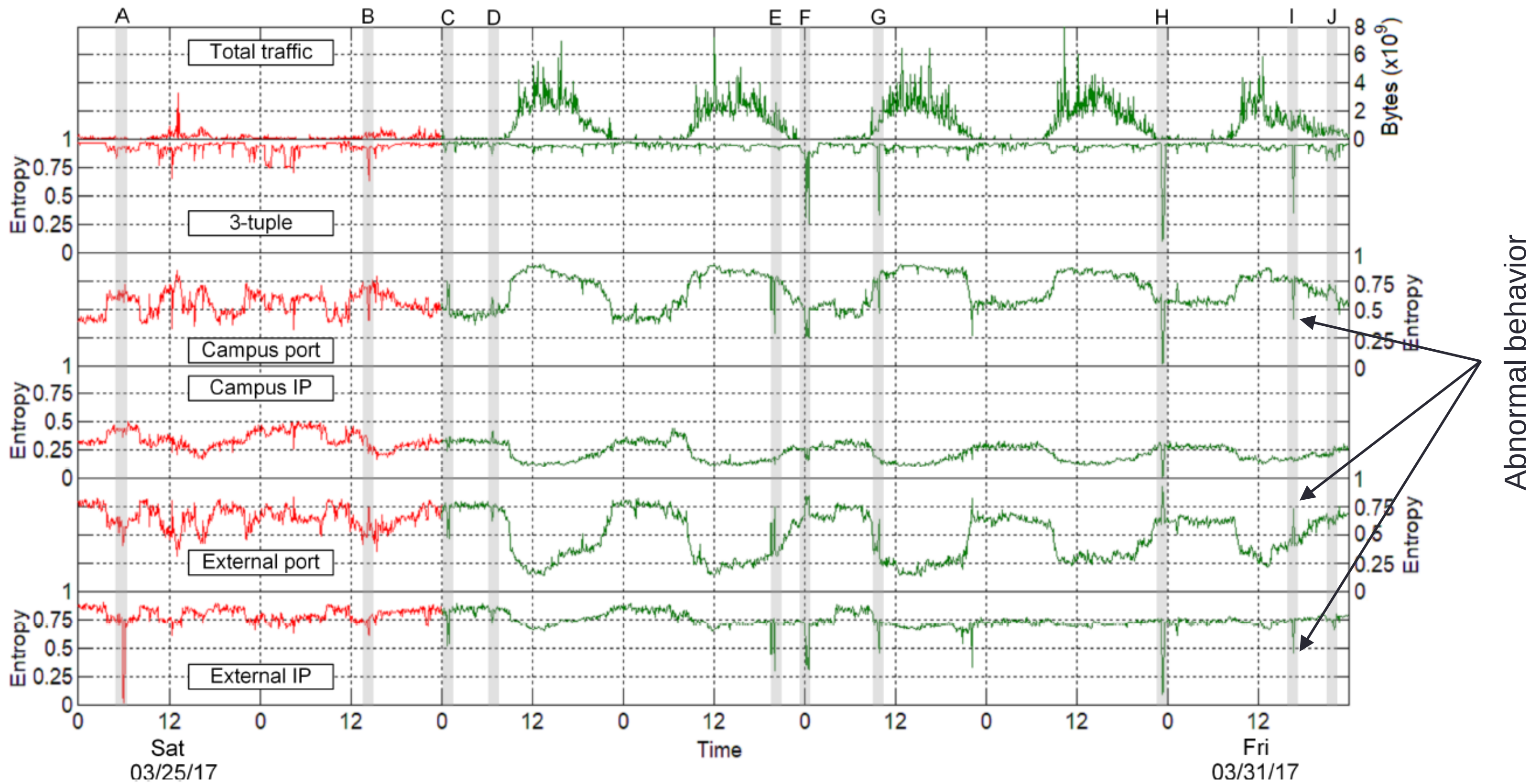
Campus IP

- In general, low entropy, 'few' IP addresses on campus
- A handful of public IP addresses used for regular Internet connectivity (NAT operation)
- Higher entropy on weekends and at night
- Lower entropy when students are on campus
- Entropy varies over 'hours' but not over very short time periods

Campus port

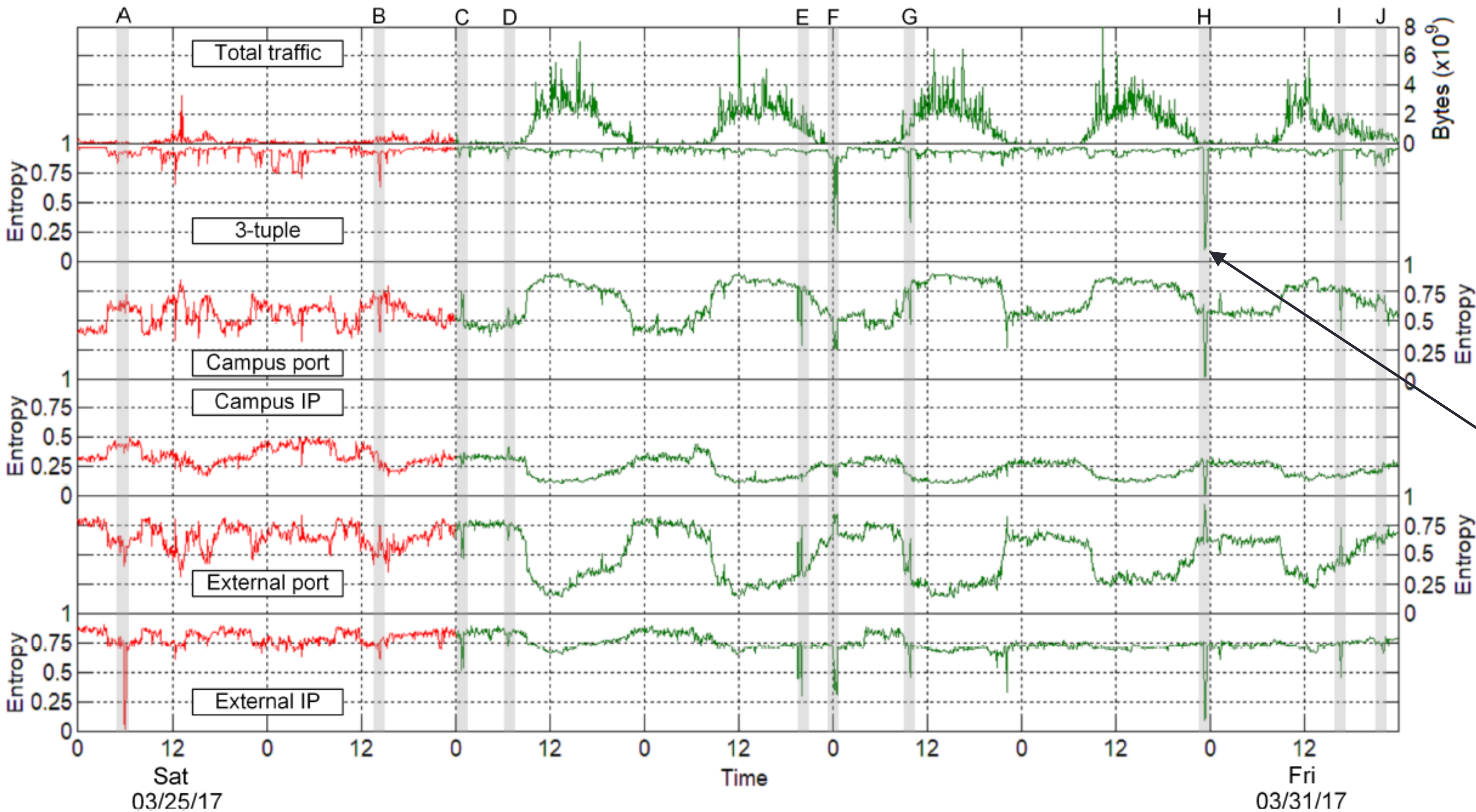
- Lower entropy at night
- High entropy (close to uniform distribution) at noon
- Dynamic / ephemeral ports used by browsers when students connect to the Internet
- Abnormal low entropy points
- Entropy widely varies over 'hours' but not over very short time periods

Results

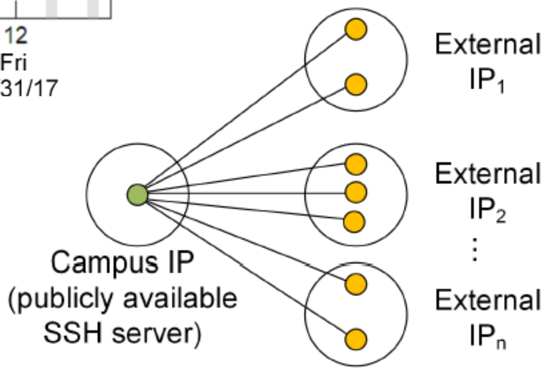


- Anomalies are detected by a single feature or by correlating multiple features
- E.g., event I: low campus port's entropy, high external port's entropy, low external IP's entropy

Results



Distributed brute force attack



Correlation of Entropy Time-series

- Campus IP – Campus port
  - The lower the number of campus IP addresses (most traffic coming from a NAT address), the higher the number of campus ports (ephemeral ports by browsers)

	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

Correlation of Entropy Time-series

- Campus port – External port
  - The higher the number of campus ports (more connected students – more ephemeral ports), the lower the number of external ports (http / https)

	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

Conclusion

- In a NATed environment, entropies may widely vary. E.g.,
  - External and campus ports vary from below 0.2 to above 0.8 (in a normalized entropy scale 0-1)
  - Campus IP address varies from 0.1 to 0.4
- Despite the wide range of values, building a granular (small time slots) entropy characterization helps to detect anomalies
- Strong correlation exists between entropy time-series
- Future work includes anomaly detection algorithms to exploit flow entropy characterization