Network Security Modelling with Distributional Data

Subho Majumdar, Splunk Threat Science*
Ganesh Subramaniam, AT&T Data Sci and AI Research

CAMLIS-2022

*Work done while at AT&T Data Sci and AI Research

Motivation

Given the lasting damage of internet security breaches, it is important to monitor IP traffic for malicious activity.

Broad objective: quickly and correctly flag anomalous external IP addresses exhibiting malicious activity patterns in communications with internal devices

- Botnet detection
- Spam/phishing attack detection
- Severity prediction
- Alert prioritization

Challenges for applying ML

There are any challenging statistical problems in network security, as well as unique challenges.

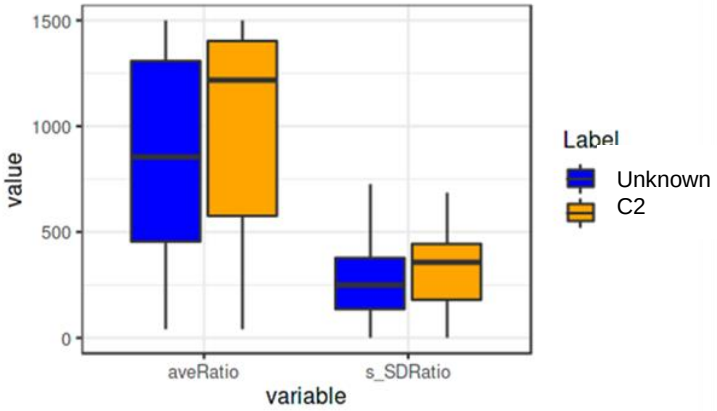
- Extreme rarity of signals
- Data and label quality
- Cost of false negatives <> Cost of verifying positives
- Proper featurization

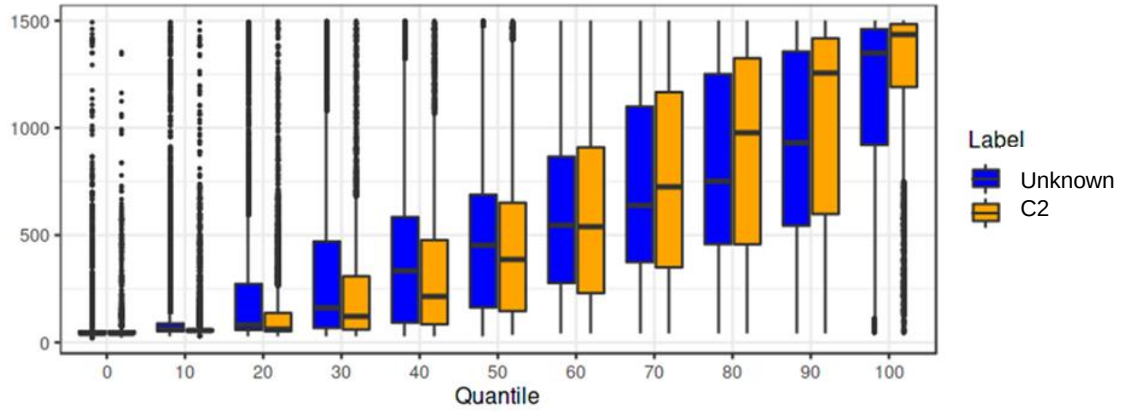
Our focus: botnet detection from NetFlow data.

Summary

We propose quantile-based featurization of distributional flow traffic data.

- Static features such as IP-level average or standard deviation of packet sizes have (a) significant overlap for malicious vs. non-malicious classes, (b) sample size restrictions
- Quantiles of the IP-level features distributions instead distinguishes Command and Control (C2) servers better. Differences are more prominent in tail quantiles.

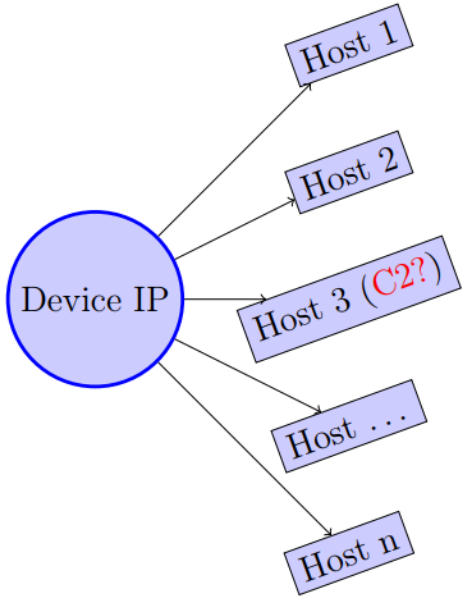


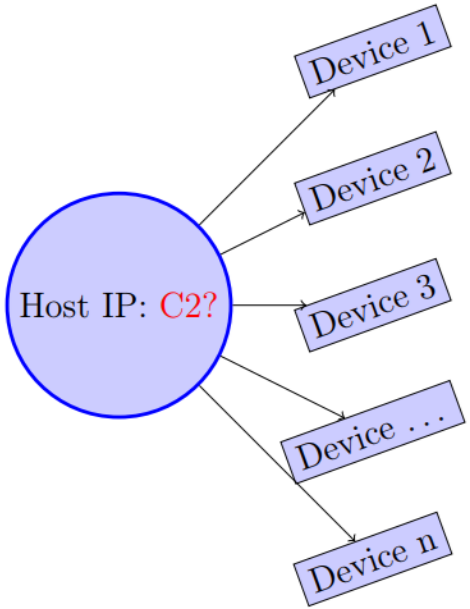


Basics: NetFlow data

Fundamental tool for characterizing IP traffic.

- Source IP address (SIP)
- Destination IP address (DIP)
- Source port
- Destination port
- Bytes transferred
- Packets transferred
- Start Time
- End Time
- IP Protocol number
- Flag

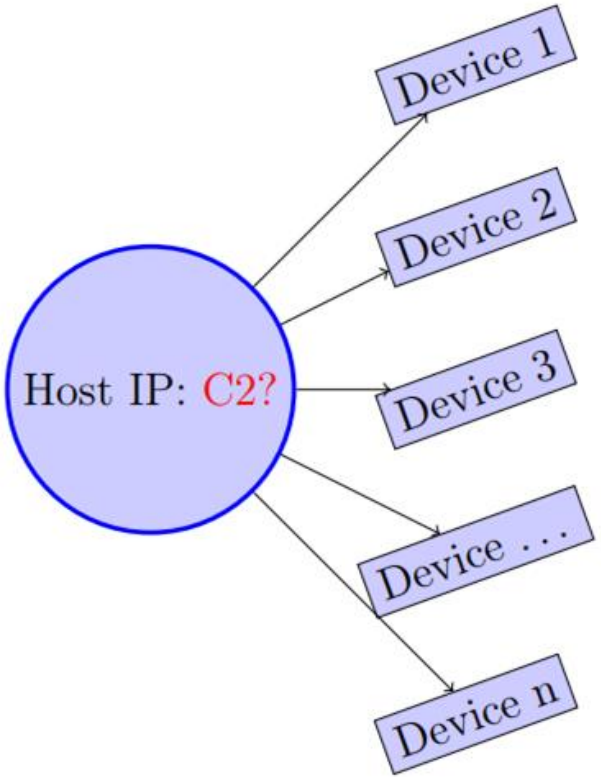




Basics: Botnet

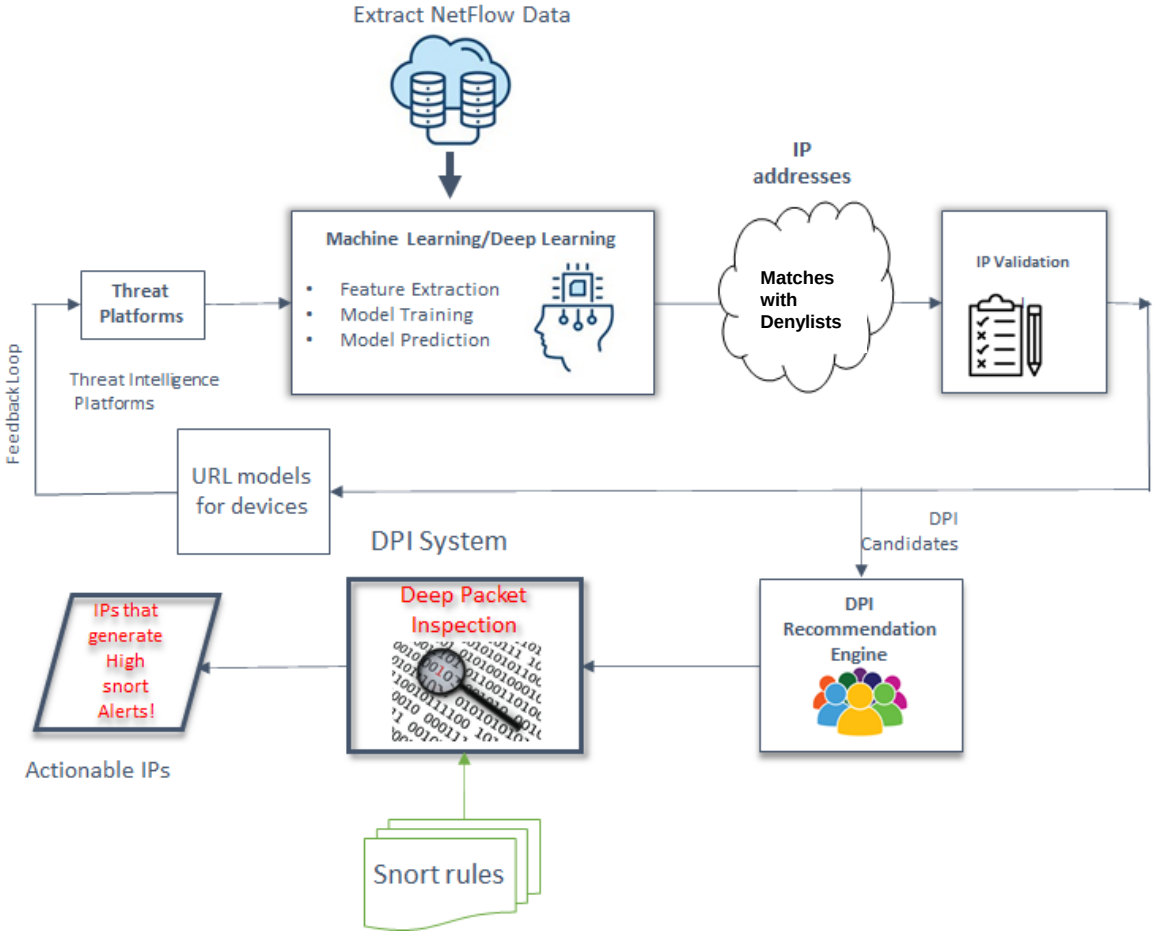
A network of compromised devices called bots and one or more Command & Control (C&C or C2) servers.

- Botmaster authors a malware that operates on each bot
- Botmaster infects devices with the malware
- Controls the bots through C2 server



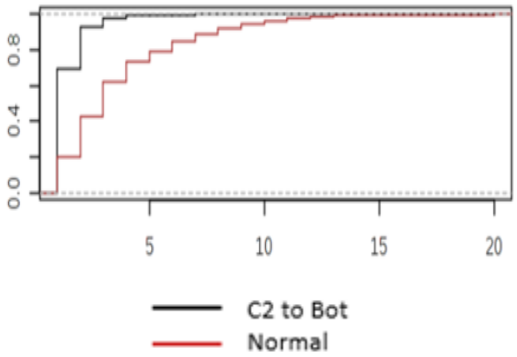
Compared to previous work, our work is host-centric, i.e. identifies if a host is C2 or not.

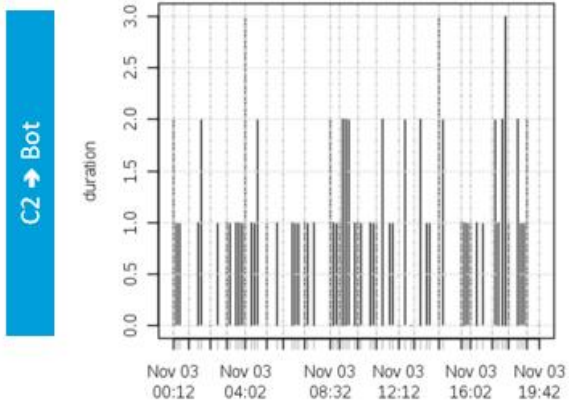
ML pipeline for botnet detection

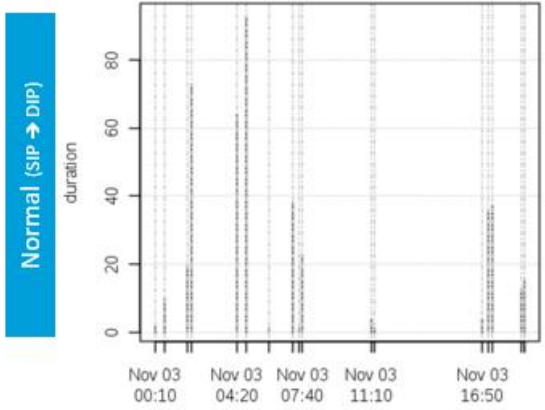


Why distributional data?

- Raw NetFlow data is SIP <> DIP connection-level, so for each SIP there is a distribution of observation for features.
- Need to summarize that to SIP level to use as input features to predict if a SIP is a C2.
- SIP-centric feature distributions show distinct differentiating patterns.







Data

- Daily flow data processed from numerous traffic domains associated with several classes of service for a telecommunications solution provider
- Based on active malware sample traffic traces observed in the network within past 30 days belonging to several malicious botnet families, some IP addresses are labeled ‘malicious’, and others as ‘unknown’.
- To mitigate the class imbalance, we sampled 1000 IP addresses from the ‘unknown’ class for every day in Dec 2021, took all IP addresses associated with the ‘malicious’ class, and used the traffic flowing through these IPs to construct our training dataset.
- This hand-constructed training data had ~17% ‘malicious’ and the rest ‘unknown’ labels.
- All traffic from Jan 2022 were used as the test dataset.

Features

Flow size features

Static feature summaries

bytes transferred

packets transferred

Avg bytes to packet ratio

Packets per flow

Bytes per flow

Beaconing features

Beaconing indicators

Periodicity of inter-arrival times

Std dev of inter-arrival times

Time gaps

Distributional features

Distributional summaries of flow features

Distributional features

Consider three input features such as packets, bytes, packets-tobytes ratios have multiple observations per IP. Denote their distributions for a device as $\mathcal{D}_p$, $\mathcal{D}_b$, $\mathcal{D}_r$, respectively. We model C2 status of a device as a function of the device-level distributions of these features:

$$\mathbb{I}(\text{malicious}) = f(\mathcal{D}_p, \mathcal{D}_b, \mathcal{D}_r).$$

Conventional analyses are based on static summaries:

$$\mathbb{I}(\text{malicious}) \simeq f((\mu(\mathcal{D}_p), \sigma(\mathcal{D}_p)), (\mu(\mathcal{D}_b), \sigma(\mathcal{D}_b)), (\mu(\mathcal{D}_r), \sigma(\mathcal{D}_r)).$$

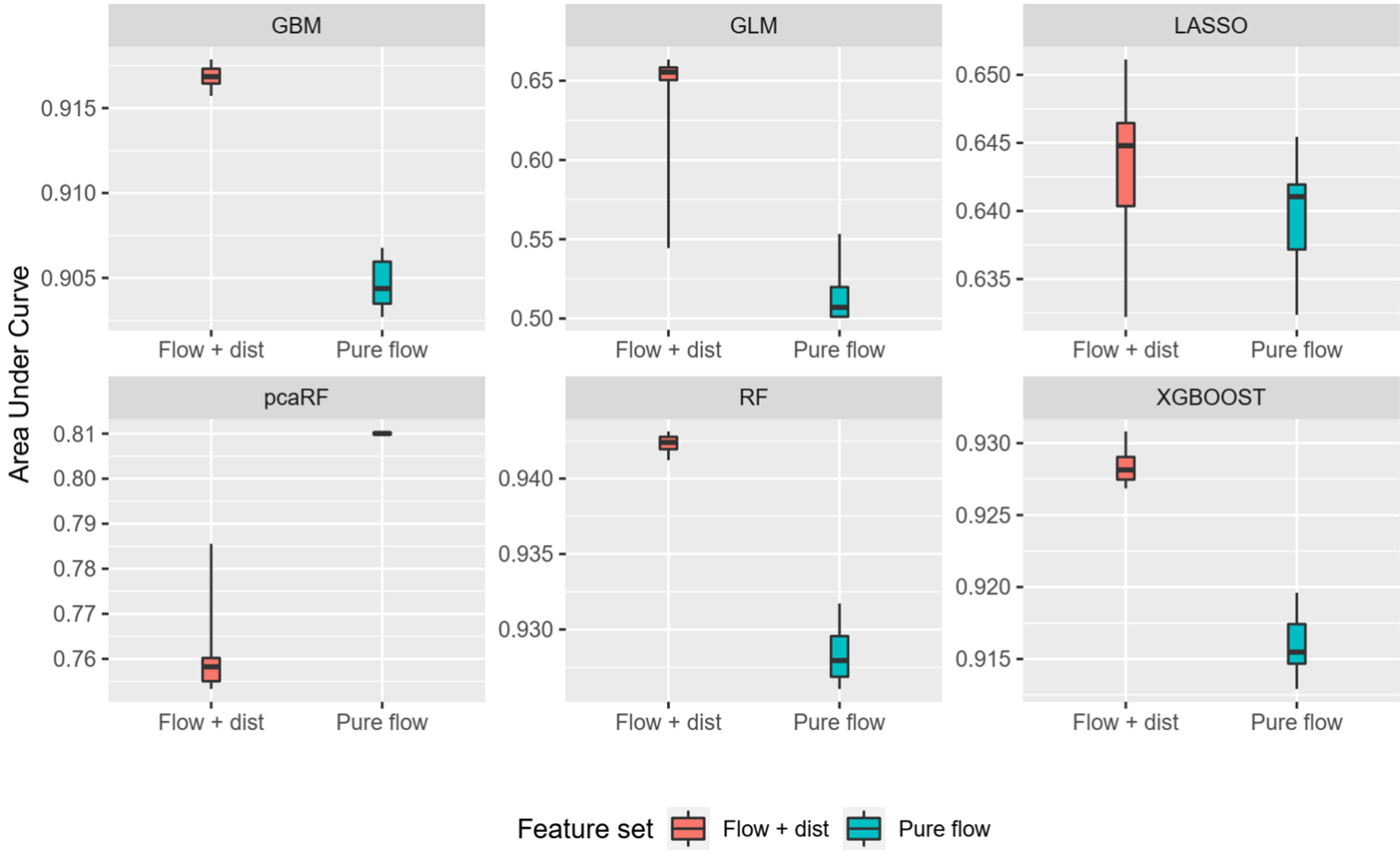
We propose using a wider spectrum of features, obtained using quantiles of each feature:

$$\mathbb{I}(\text{malicious}) \simeq f(G(\mathcal{D}_p), G(\mathcal{D}_b), G(\mathcal{D}_r)),$$

Where G = (μ, σ, Q) indicating the the vector transformation giving n pre-defined quantiles from a distribution.

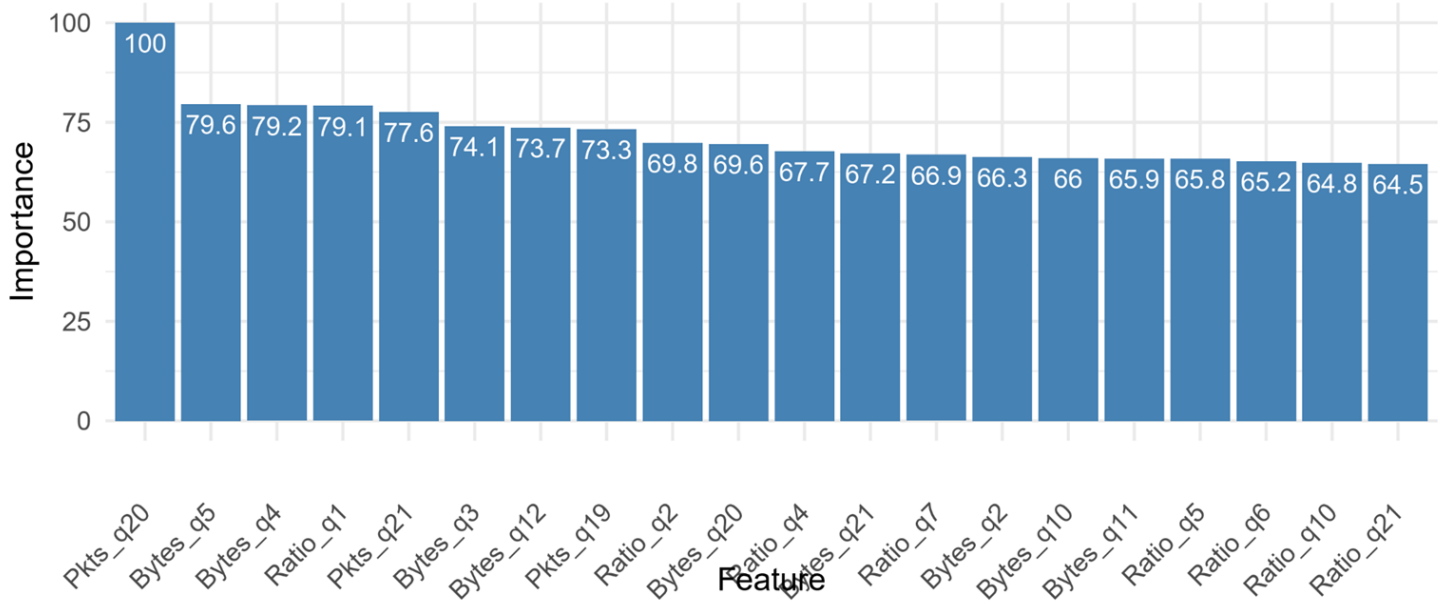
We take n = 20: performance plateaus at higher values.

Results: performance metrics



Results: variable importance

- All the top important features belong to the quantile feature set underline their informativeness in our IP classification scenario.
- Most of the top 20 quantiles are correspond to either of the tails: 14 of the 20 quantiles lie outside the Inter-quartile range, i.e. 25th and 75th quantiles.



Future work

- Sophisticated ensembles for prediction, e.g. model averaging
- Deep-learning methods
- Separate models for separate malware families, while tackling data challenges

THANK YOU!