Data ~ Detectability

Haakon Larsen,
Mark Crovella,
Christophe Diot,
Jennifer Rexford
Augustin Soule

Understanding how the nature of the source data impacts network anomaly detectability

Outline

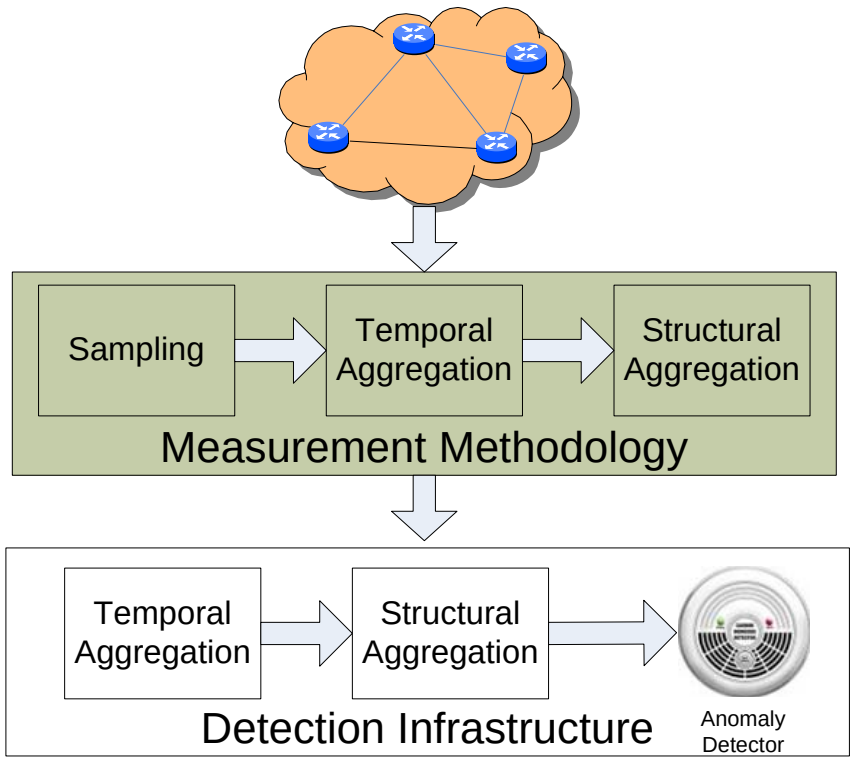
- Background and motivation
  - Anomaly detection “status quo”
  - Why would the source data matter?
- Relevant questions & Approach
  - Impact, tunableness, leveraging address aggregation
- Lessons learned
  - Differences in population, normalization, key aggregation
- Brainstorming!

Anomaly detection

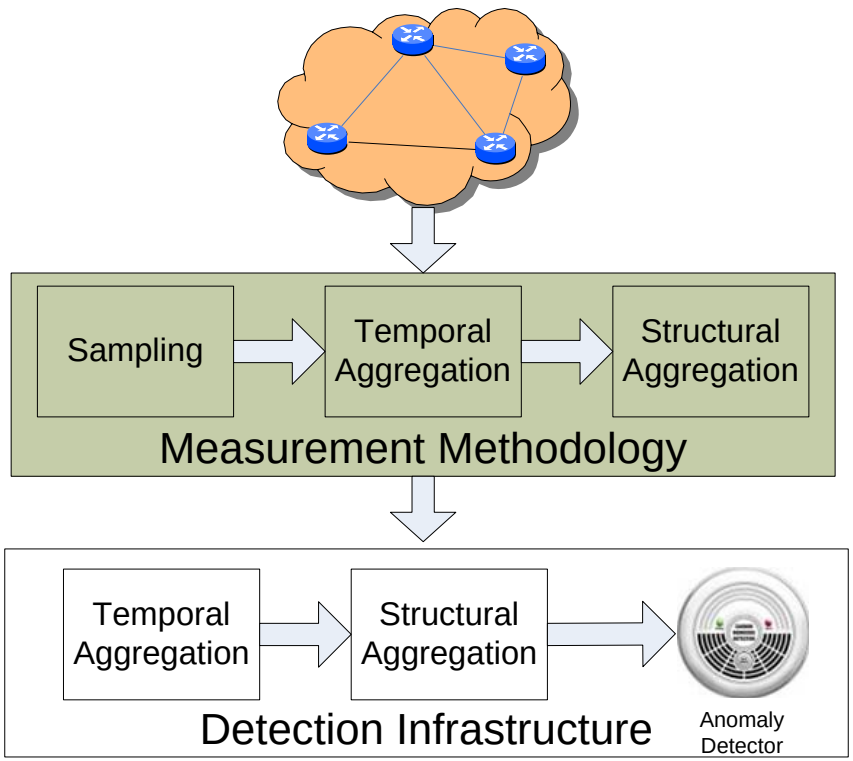
- Analyze traffic traces to discover outliers
  - Change in # bytes, # packets, or # flows
  - Change in source/destination IP/port distributions
- A rich subject area
  - PCA, Kalman filter, signal analysis approaches, signature detection, modeling approaches, etc

What about the data?

- Sampling
- Temporal aggregation
- Address aggregation
- Aggregation into link or OD-flow traffic matrixes
- (Let's call these "data parameters")



What about the data?



- The top half is a given in the measurement infrastructure
  - i.e. data parameters we must deal with
  - Exists for a reason
- The latter half is post-processing performed for anomaly detection

Abilene versus Géant

- Our focus is on detectability within a single network
- But the two networks highlight need for adaptive tuning of your approach

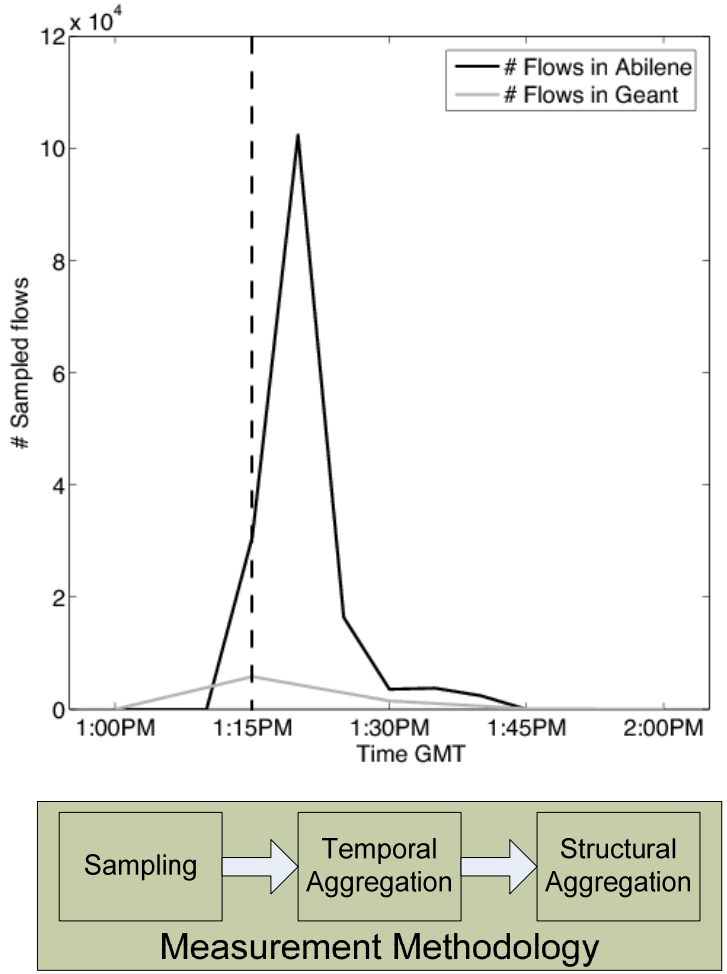
	Abilene	Géant
Sampling Rate	1 / 100	1 / 1000
Temporal Aggregation	5 min bins	15 min bins
Address Aggregation	/21	None

Outline

- Background and motivation
  - Anomaly detection “status quo”
  - Why would the source data matter?
- Relevant questions & Approach
  - Impact, tunableness, leveraging address aggregation
- Lessons learned
  - Differences in population, normalization, key aggregation
- Brainstorming!

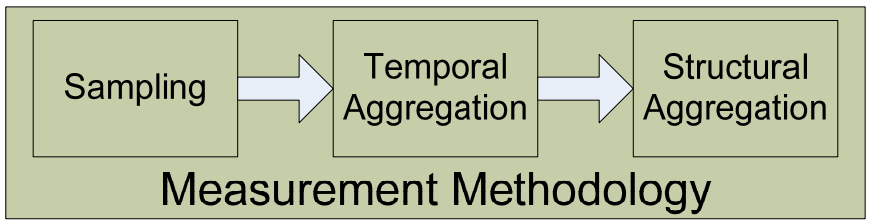
Question: impact

- How does parameter X impact detectability of anomaly Y?
- Can we then recommend which dimension to trade off for computational complexity?

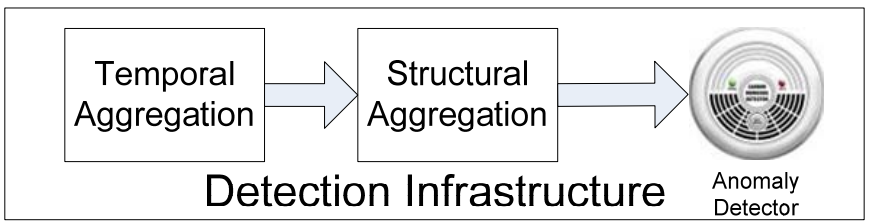


Question: tuning

- Given a measurement methodology
  - Unalterable data parameters



- How do we optimally tune the detection algorithm
  - e.g. thresholds for anomaly detection algorithms



Approach

- Data used
  - Abilene and Géant networks
  - NetFlow traffic traces from 11/2005
- Vary data parameters
- Anomaly Detection
  - Calculate entropy of IP/port distributions
  - Analyze time-series with Kalman filter & PCA
- Validation…

Outline

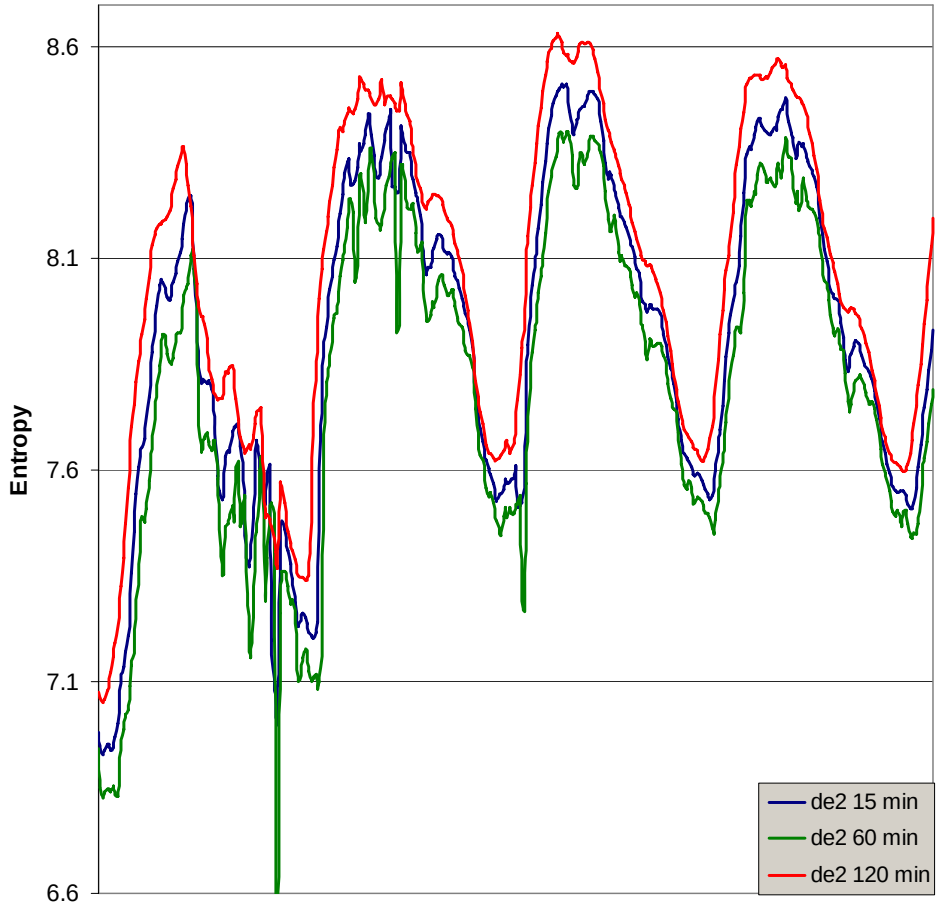
- Background and motivation
  - Anomaly detection “status quo”
  - Why would the source data matter?
- Relevant questions & Approach
  - Impact, tunableness, leveraging address aggregation
- Lessons learned
  - Differences in population, normalization, key aggregation
- Brainstorming!

Lessons learned thus far

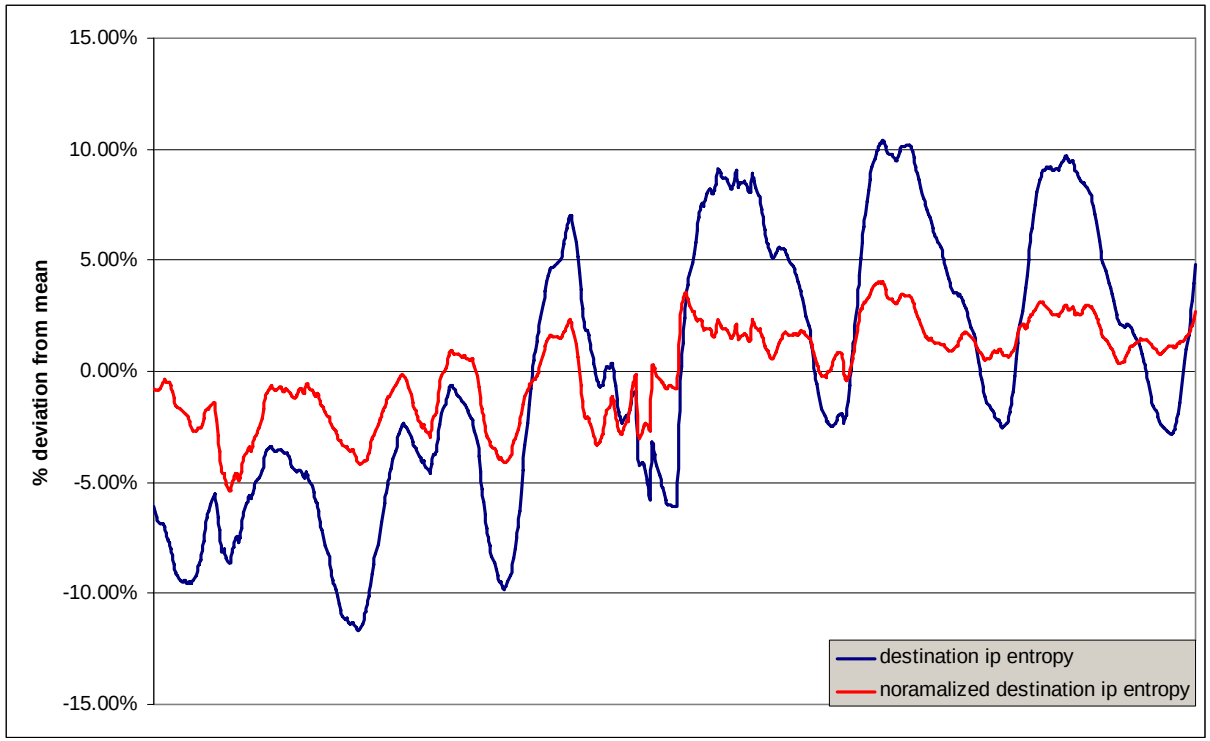
- Géant and Abilene are very different
  - In more ways than data parameters
- Entropy is sensitive to volume shifts (e.g. diurnal trends)
  - Need for normalization
- Analyze multiple aggregation levels!

Need for entropy normalization

- There are diurnal trends in the data
- Entropy is vulnerable to changes in number of base symbols
- Can be dealt with by normalization, PCA, or Kalman filters



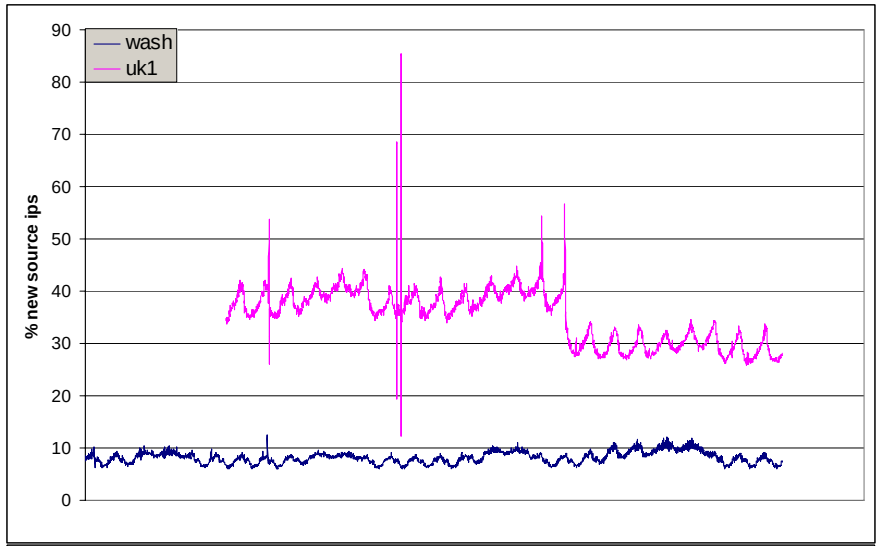
Effectiveness of normalization

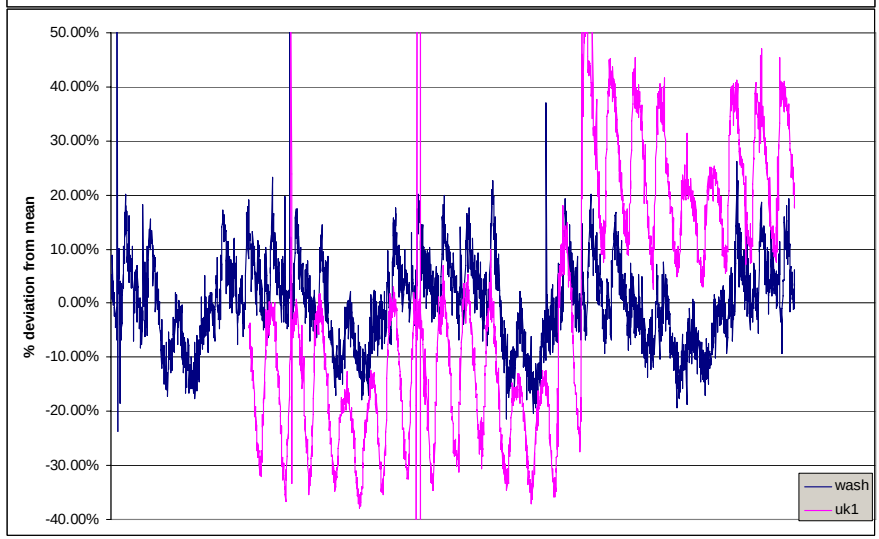


- Standard deviation as a fraction of the mean is reduced by up to a factor of 2

Different IP populations

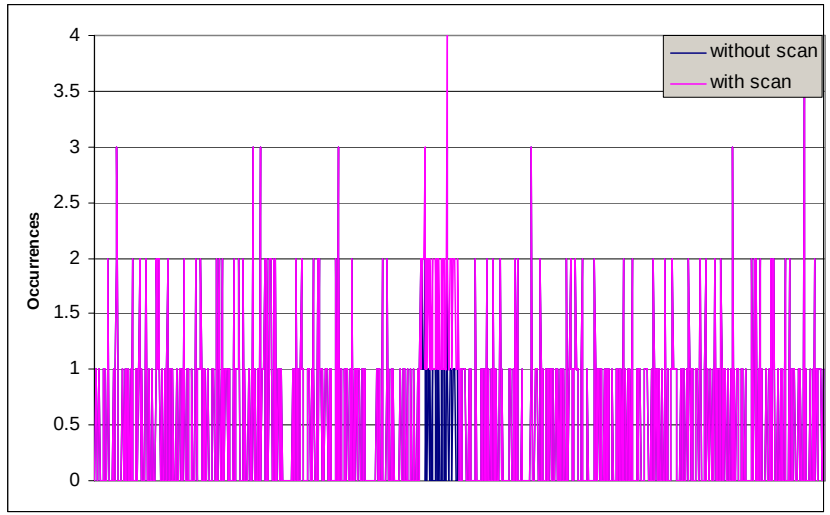
- Abilene is not a transit network
  - Not connected to the Internet!
- Only U.S. universities and research institutions
- Very stable IP population

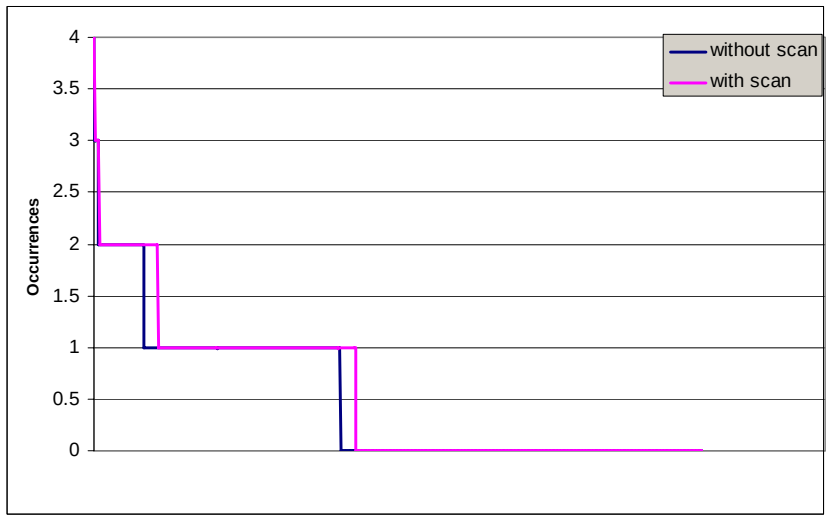




Leveraging key aggregation

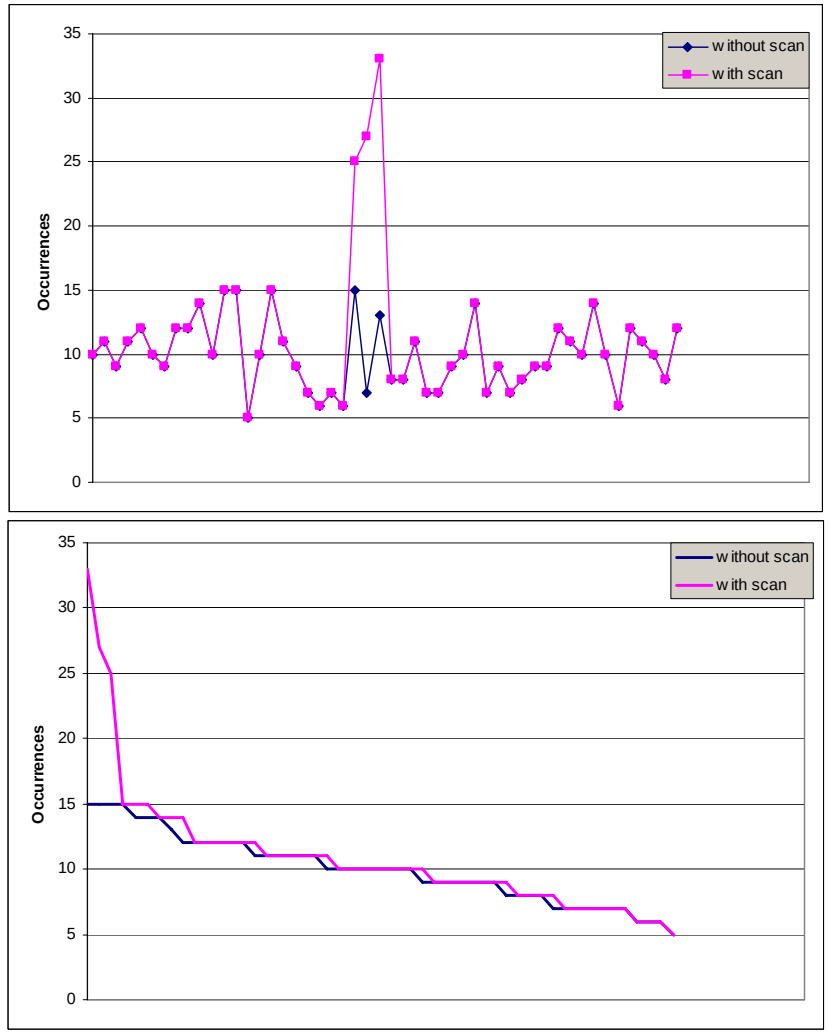
- An IP scan will add a small amount of traffic to each key scanned
- The increase in total keys may increase entropy *slightly*
- But we can *see* that it's an attack!



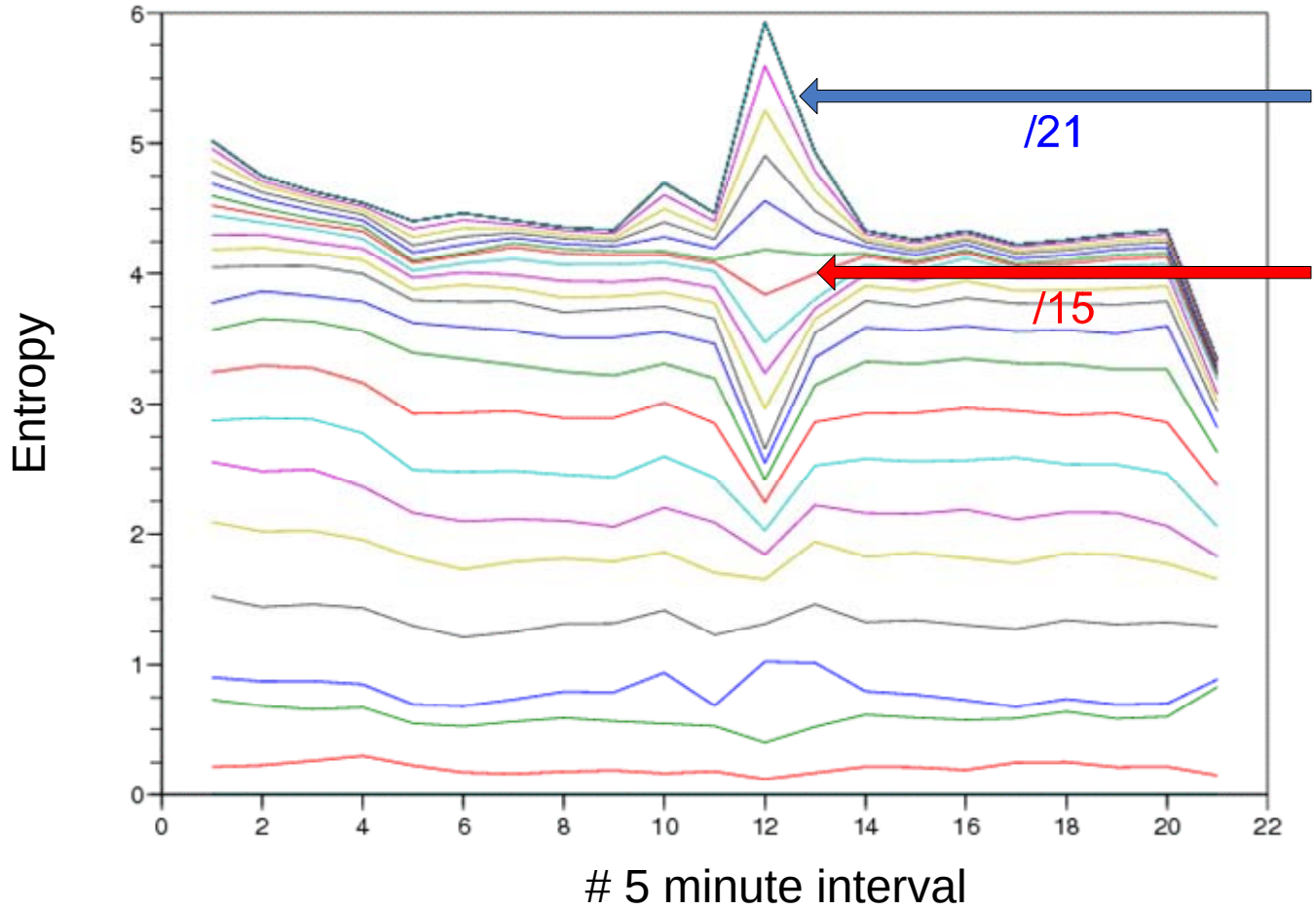


Leveraging key aggregation

- Aggregate into IP prefixes
- A scan that is dispersed at /32 may concentrate at /24
- Decrease in entropy
- Useful for diagnosis also?



Key aggregation for diagnosis



Outline

- Background and motivation
  - Anomaly detection “status quo”
  - Why would the source data matter?
- Relevant questions & Approach
  - Impact, tunableness, leveraging address aggregation
- Lessons learned
  - Differences in population, normalization, key aggregation
- Brainstorming!

Validation

- Manually label a trace
- Run a suite of outlier detectors on a trace to find a set S of possible anomalies
  - Suite S can be pruned by a domain expert
- Model anomalies & detection in a simulation or emulation environment
- Inject anomalies into a trace according to some model

- Thank you!

(Especially Mark Crovella,
Christophe Diot, Jennifer Rexford,
and Augustin Soule)