

Impact of Packet Sampling on Anomaly Detection Metrics

Daniela Brauckhoff*, Bernhard Tellenbach*, Arno Wagner*,
Arukool Lakhina **, Martin May*

*ETH Zurich, ** Boston University



Motivation

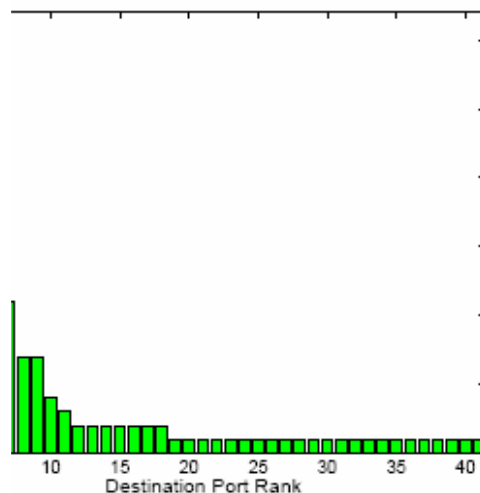
- The general opinion about sampling:
 - With sampling valuable information lost about anomalies
 - But sampling needs to be used anyway...
 - Cannot get unsampled netflow from some routers
- Interesting questions arise:
 - *How much* information is actually lost?
 - Are all *anomalies* equally affected by sampling?
 - Are all *detection metrics* equally affected by sampling?
 - At which *sampling rate* is a certain anomaly still detectable?
 - Can we estimate the *original anomaly size* from a sampled view?

Data & Experiments

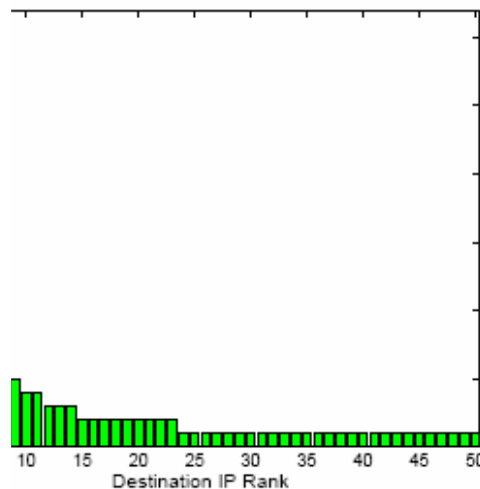
- A week-long dataset of *unsampled Netflow records* from a backbone router of a national ISP
- Known Blaster outbreak in our data
- **Goal:** Study impact of packet sampling on Blaster worm
 - Focus on visibility of Blaster worm
 - Focus on anomaly detection metrics
 - **Bytes, Packets, Flows, Traffic Features, ...**

Entropy as a Detection Metric [LCD:SIGCOMM05]

**Dispersed
Histogram**
High Entropy



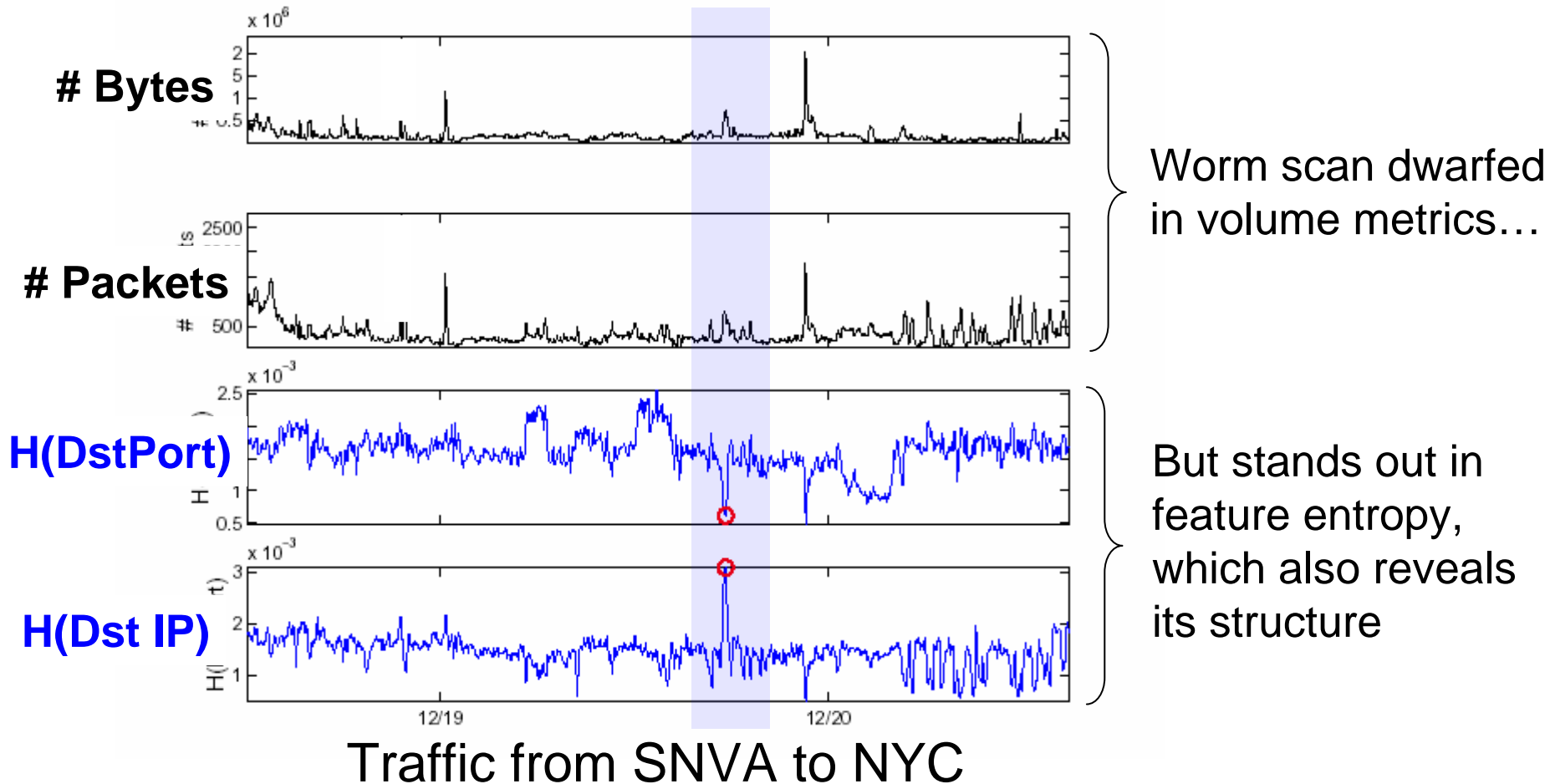
**Concentrated
Histogram**
Low Entropy



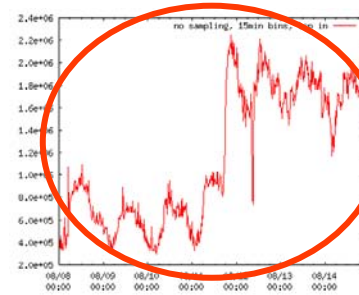
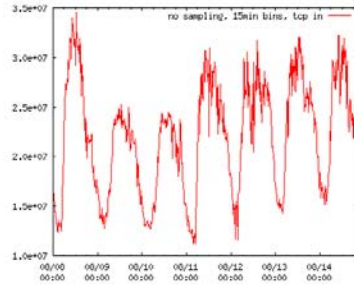
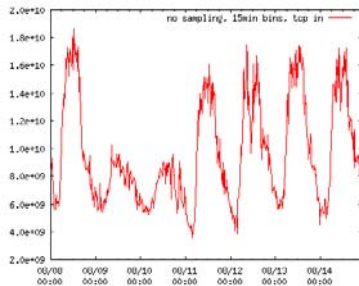
Summarize using
sample entropy of
histogram X

typical Traffic

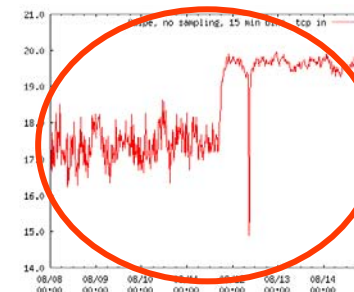
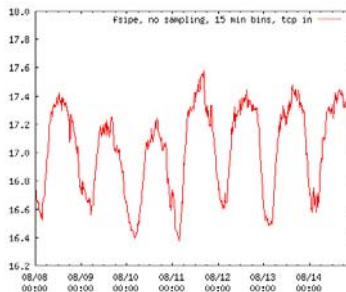
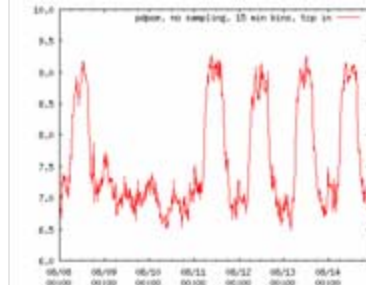
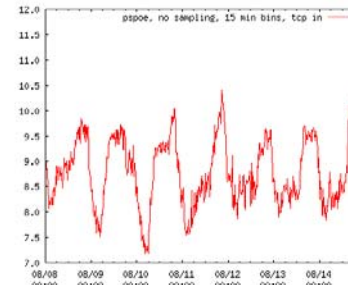
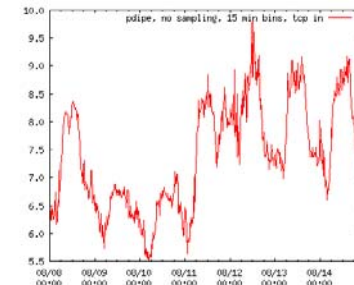
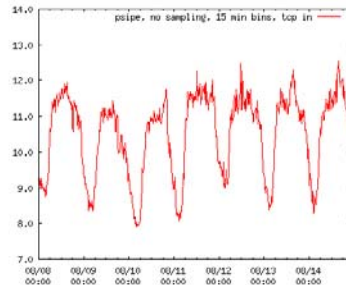
The Power of Entropy



Which AD metrics to look at?



Flow counts



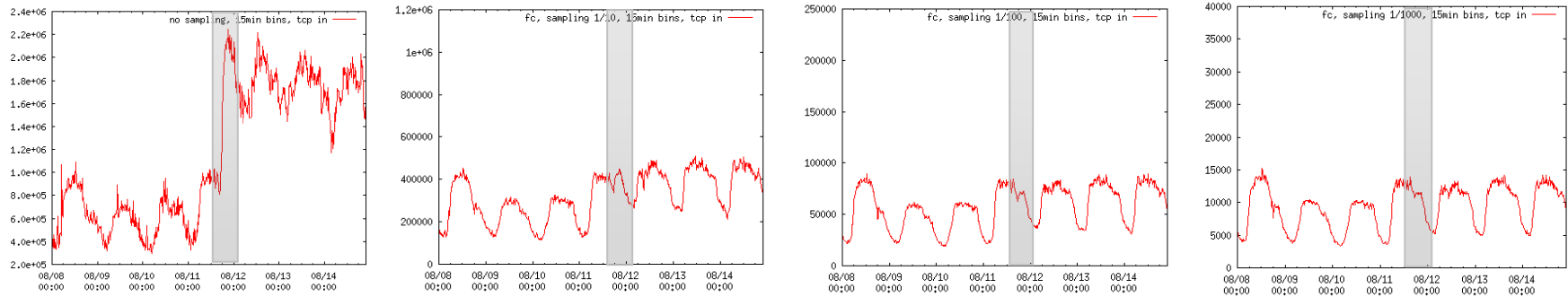
Flow destination IP entropy

Methodology: Packet Sampling

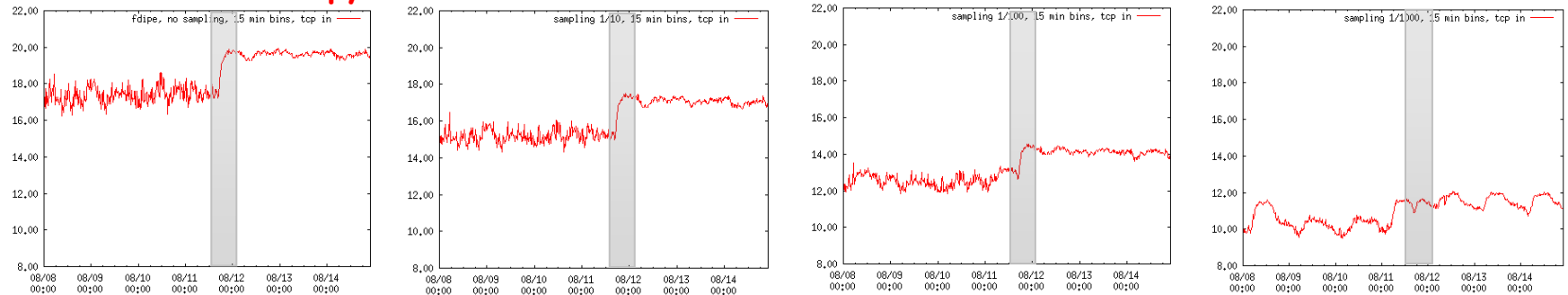
- Determine the *packet size* (bytes) and *timestamps* for individual packets in the flow trace
- Each packet of a flow is recorded in it's own flow record with
 - $packet_size = flow_size/num_packets$ (average packet size)
 - timestamp randomly chosen within flow bounds
- Randomly sample every 10th, 100th, 250th, and 1000th packet
 - Not exactly what Cisco does, but pretty close...

Timeseries of Detection Metrics

Flow counts



Flow dst IP entropy



unsampled

10

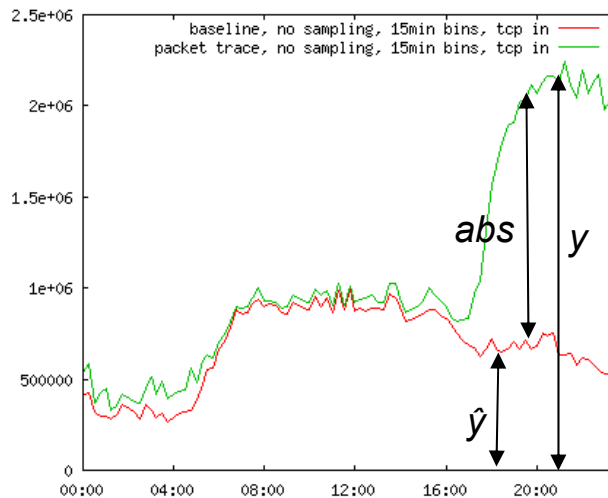
100

1000

Methodology: Determine the Baseline

- AD algorithms measure distance from (predicted) baseline to (actual) observed metrics
- Each AD method uses it's own handcrafted algorithm to determine the baseline model
- Since we know the anomaly very well we can construct an „*ideal baseline*“ by removing all blaster packets from the observed trace
 - *Heuristic: blaster packet = packet with destination port 135, protocol TCP, and length of 40, 44, 48 bytes*
- One baseline per metric and *sampling rate*

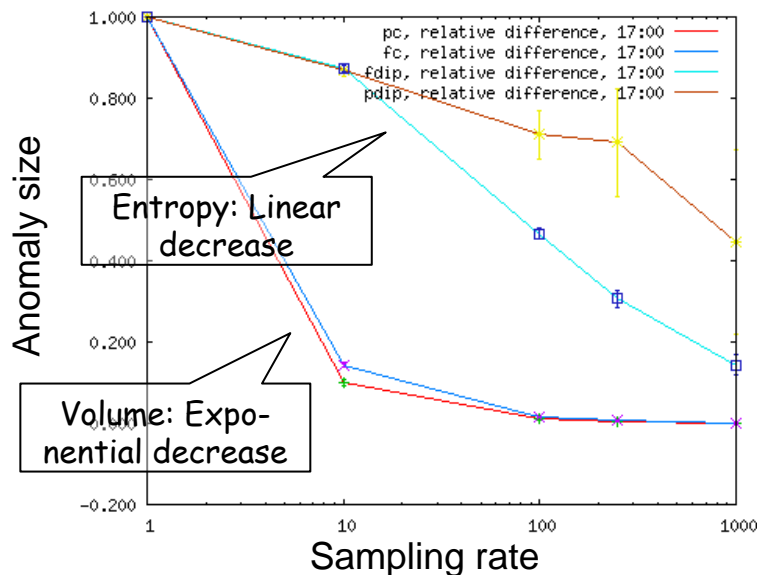
Methodology: Measure anomaly distance



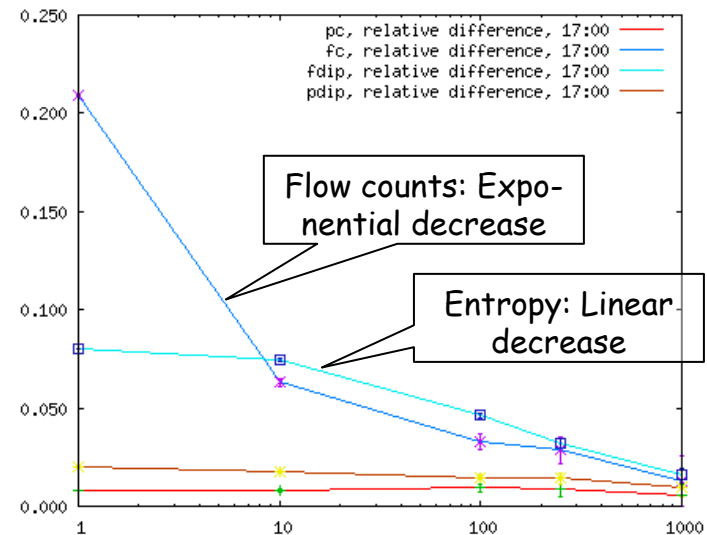
- Absolute difference between trace y and baseline \hat{y}
 - $abs = y - \hat{y}$
- Absolute difference normalized to the baseline \hat{y}
 - $rel = (y - \hat{y}) / \hat{y}$

Anomaly Distance vs Sampling Rate

Absolute distance



Relative distance



Q: What do these distance measures tell us?

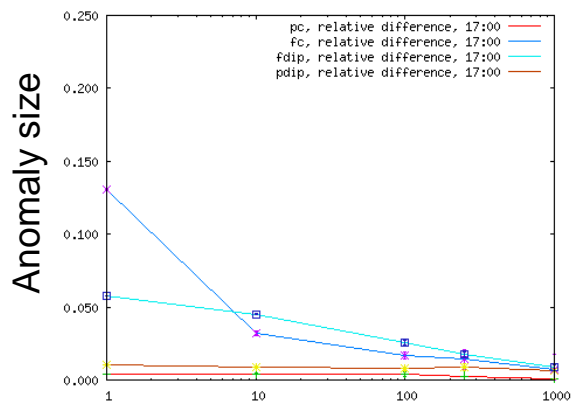
A: In this scenario entropy is less disturbed by sampling...

Scaling the Blaster Worm

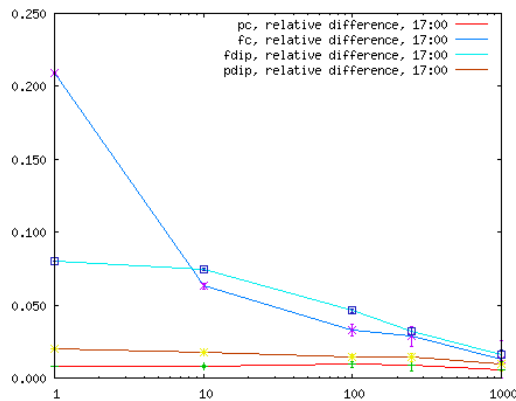
- *Identification* of Blaster packets based on heuristic
 - dst port, packet size, tcp
- *Amplification* of the Blaster worm
 - Insertion of new packets with same src IP, and dst IP randomly selected from SWITCH IP range
- *Attenuation* of the Blaster worm
 - Randomly throwing out of some of the Blaster packets (e.g., select each packet with probability of 50%)

Relative Distance for Scaled Blaster

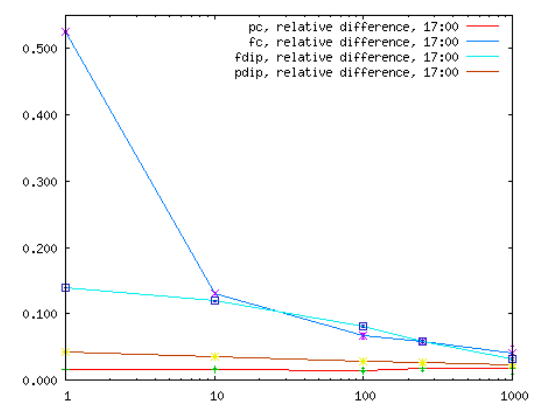
Scaling factor: 0.5



Scaling factor: 1



Scaling factor: 2



Q: What do these scaled distance measures tell us?

A: For faster and slower Blaster-like worms, entropy is less disturbed by sampling than flow counts...

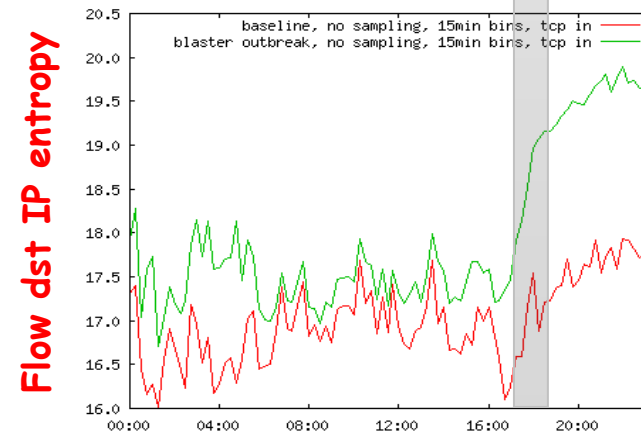
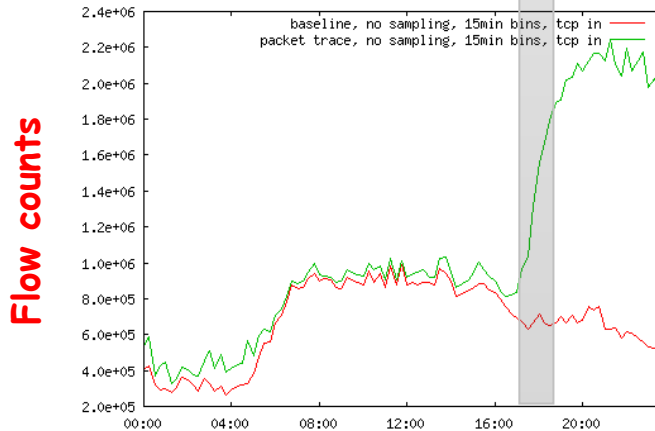
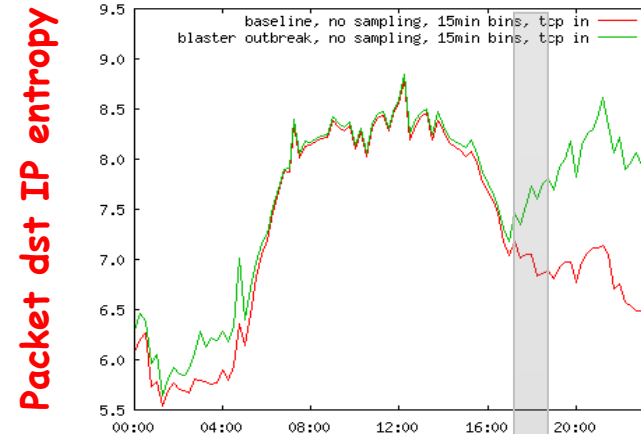
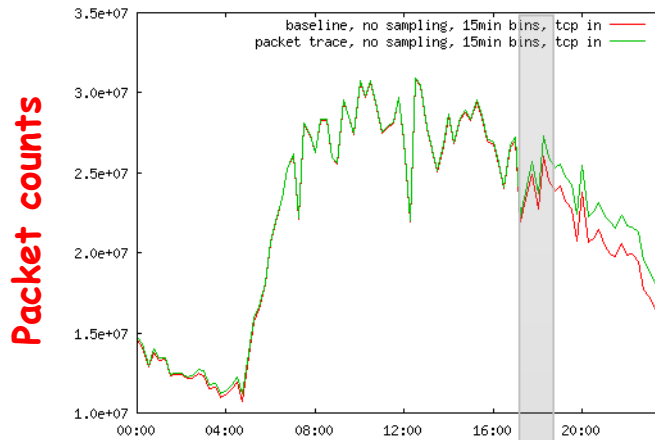
Conclusion and Future Work

- What did we learn?
 - Some metrics are more resilient to sampling than others
 - Flow DST IP entropy is most resilient to sampling for Blaster-type anomalies (in our traces)
- What still needs to be studied...
 - Other types of anomalies, anomaly intensities
 - Other distance metrics (considering a metrics' variance)
 - Different bin sizes
 - Further anomaly metrics
 - Anomaly detectability at different sampling rates

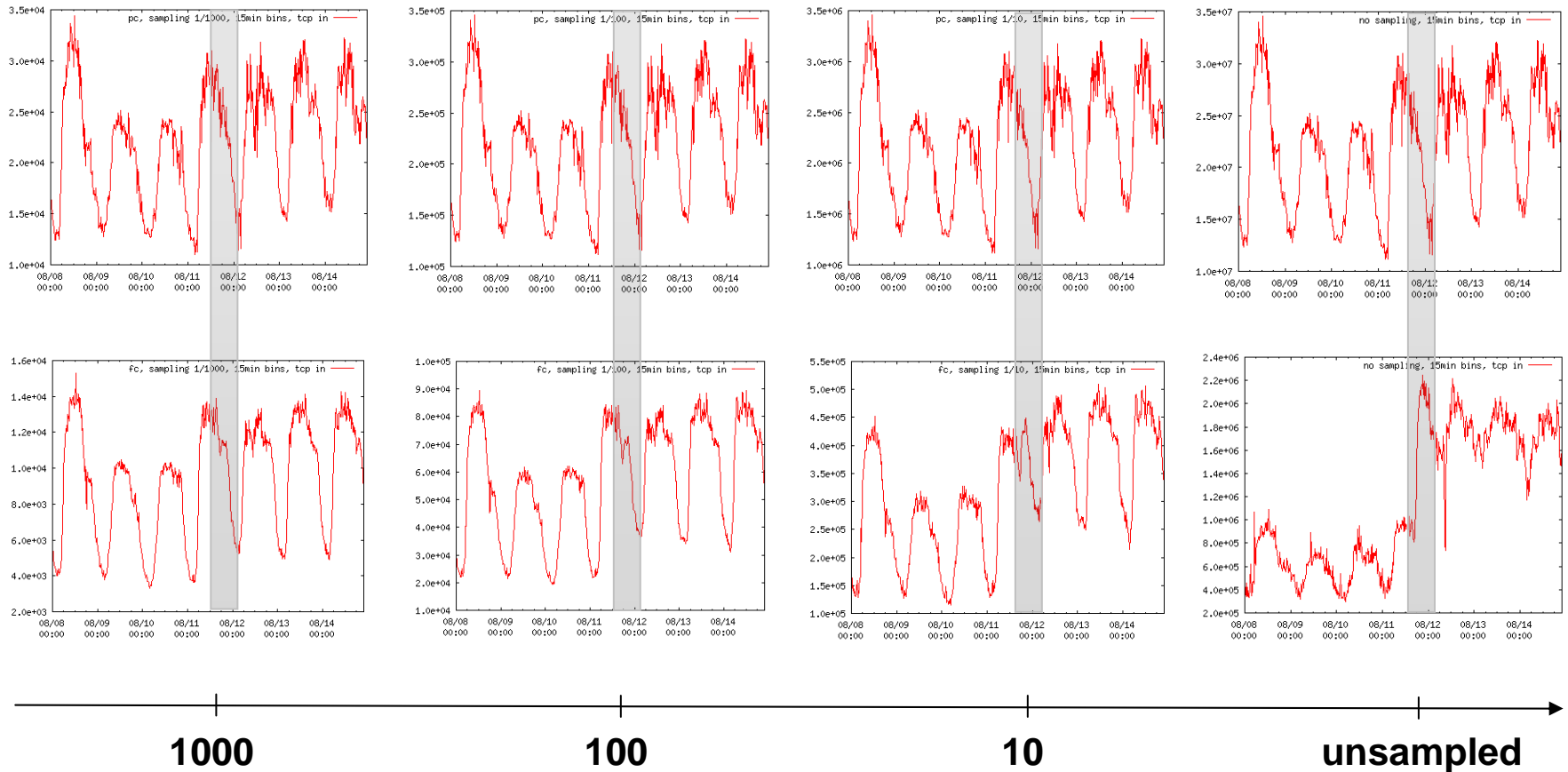
Questions?

Daniela Brauckhoff
ETH Zurich, Switzerland
brauckhoff@tik.ee.ethz.ch

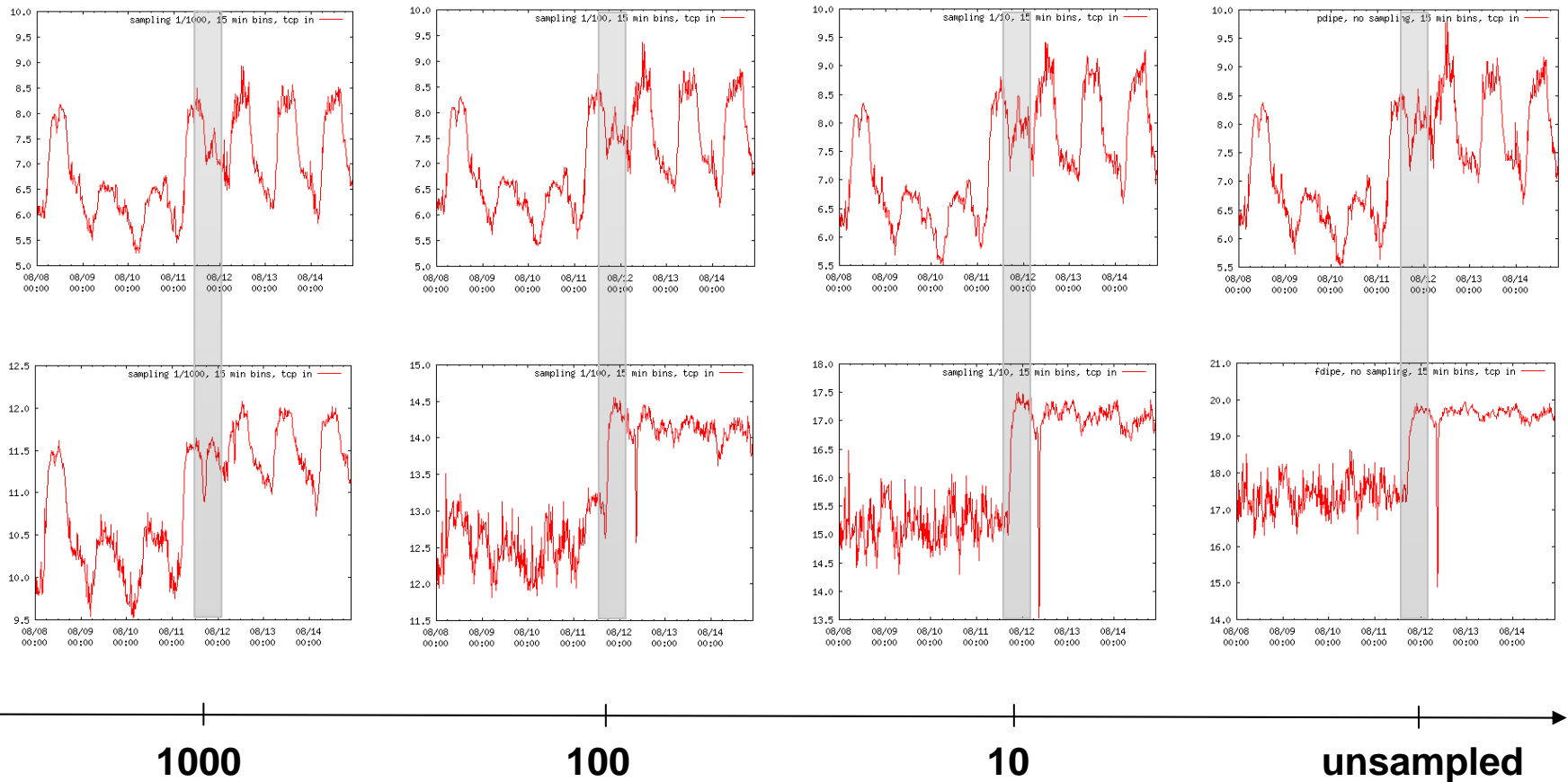
Baselines for AD Metrics (unsampled)



Volume Time Series

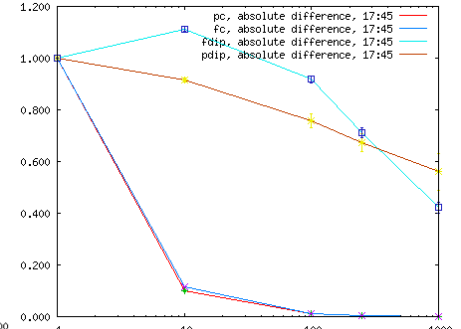
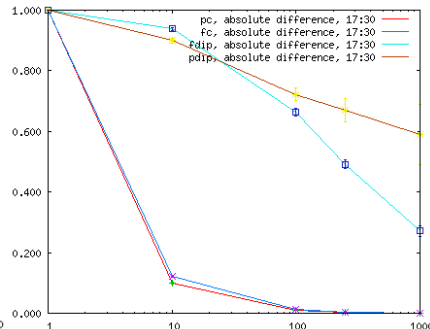
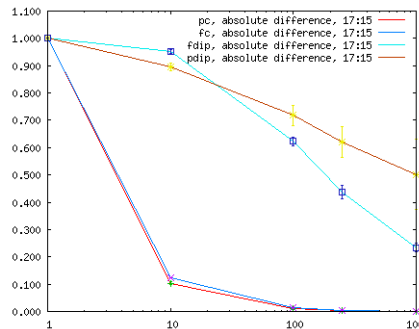
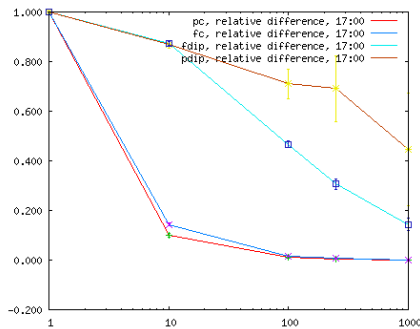


Entropy Time Series

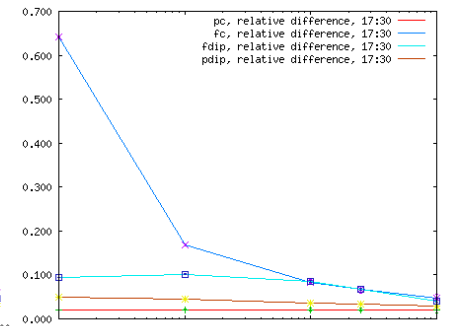
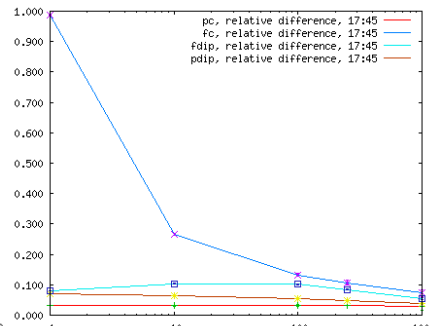
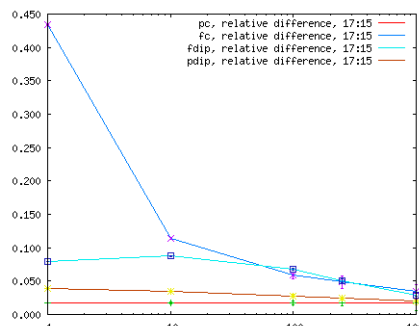
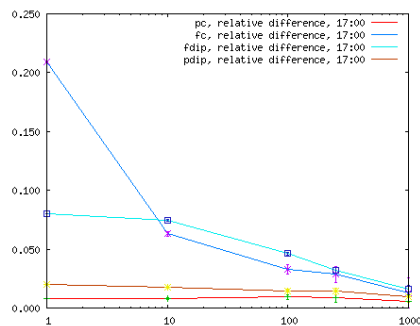


Anomaly Distance vs Sampling Rate

Absolute distance



Relative distance



17:00

17:15

17:30

17:45