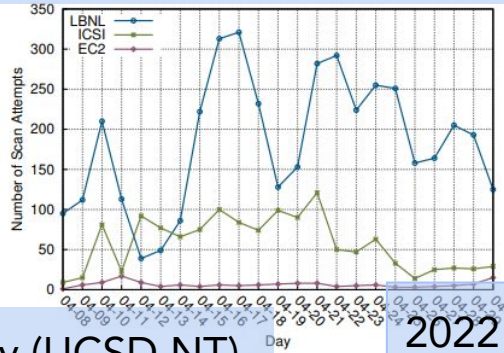
Benchmarking IBR Event Detection Frameworks

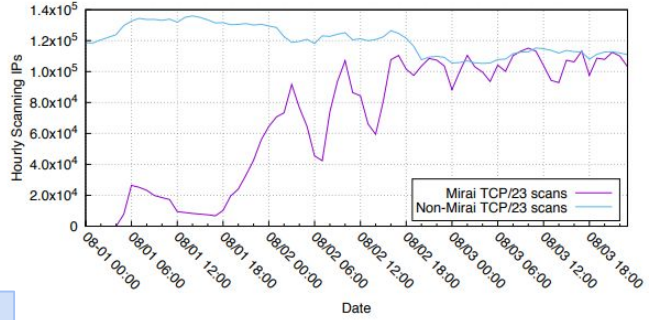
Max Gao
CAIDA/UC San Diego

Scanning events seen by network telescopes

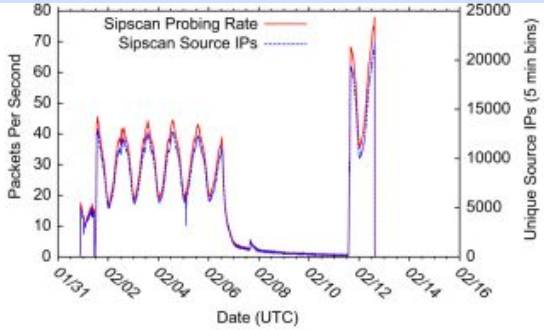
2014 - Heartbleed (LBNL+ICSI)



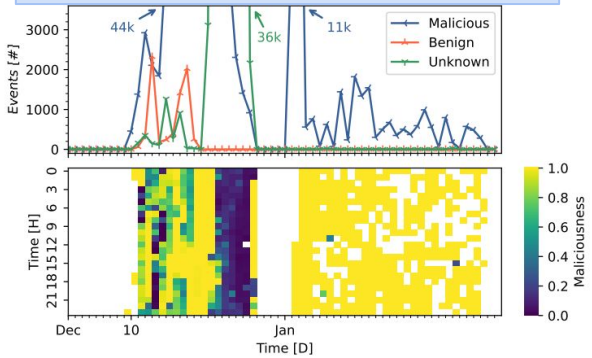
2016 - Mirai (Orion NT)



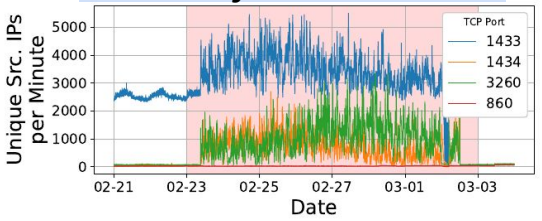
2011 - Salidity (UCSD NT)



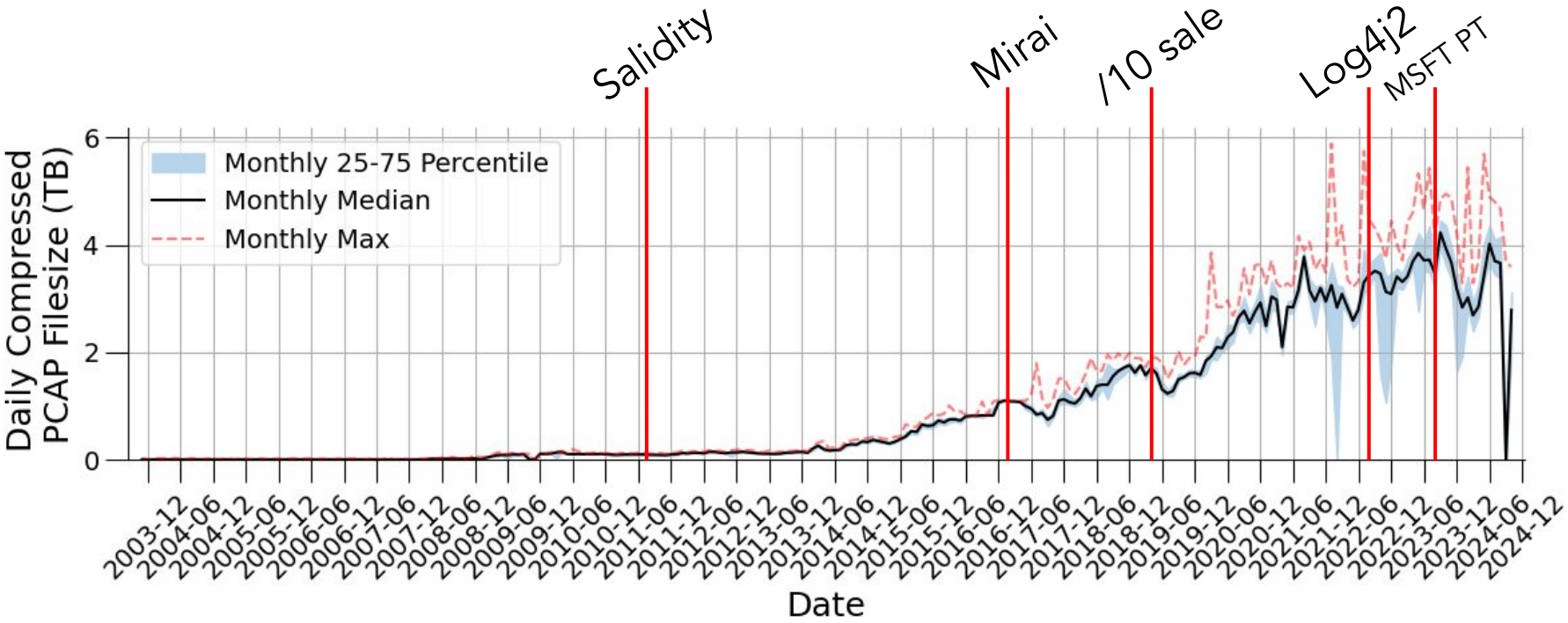
2022 - Log4j2 (UCSD NT)



2023 - MSFT Patch Tuesday (UCSD NT)



UCSD-NT traffic volumes



Benchmark goals

- Assess trade-offs between frameworks
  - detection capabilities
  - computational performance
  - scalability

- Document challenges faced during framework replication and evaluation

Candidate frameworks

- Sample a variety of frameworks from the public domain
  - Graph-based (e.g., [3])
  - Dimensionality-reduction (e.g., [1], *DarkVec* [2])
  - Time series analysis (e.g., *DarkTracer* [4])

- General stages shared across frameworks, but details differ
  - Data Preprocessing
  - Core algorithm
  - Decisioning

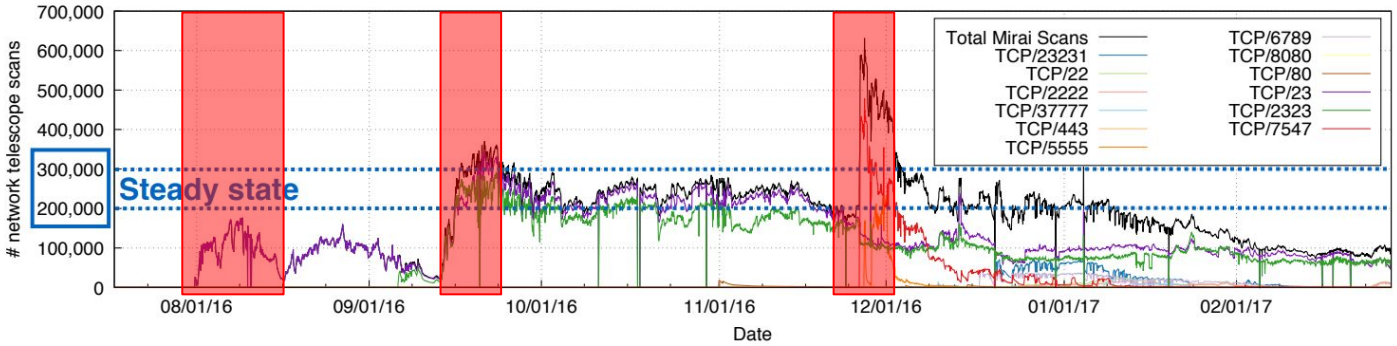
[1] M. Kallitsis, et al., *Detecting and Interpreting Changes in Scanning Behavior in Large Network Telescopes*. IEEE Transactions on Information Forensics and Security, 2022
[2] Luca Gioacchini et al., *DarkVec: automatic analysis of darknet traffic with word embeddings.* CoNEXT, 2021.
[3] Sofiane Lagraa, et al., *Deep Mining Port Scans from Darknet*. International Journal of Network Management, 2019.
[4] C Han, et al., *Dark-TRACER: Early Detection Framework for Malware Activity Based on Anomalous Spatiotemporal Patterns*.IEEE Access, 2022

Framework evaluation

- Devise metrics for comparing frameworks
  - Detection capabilities
    - Standard classification metrics (e.g., TP, FP)
  - Computational performance
    - CPU + Memory usage
    - File sizes
  - Scalability
    - Varied traffic volumes, traffic complexity
- Perform offline analysis on UCSD-NT historical traffic
  - Implement frameworks using standardized tooling
  - Run implementations on Expanse's infrastructure

What about ground truth?

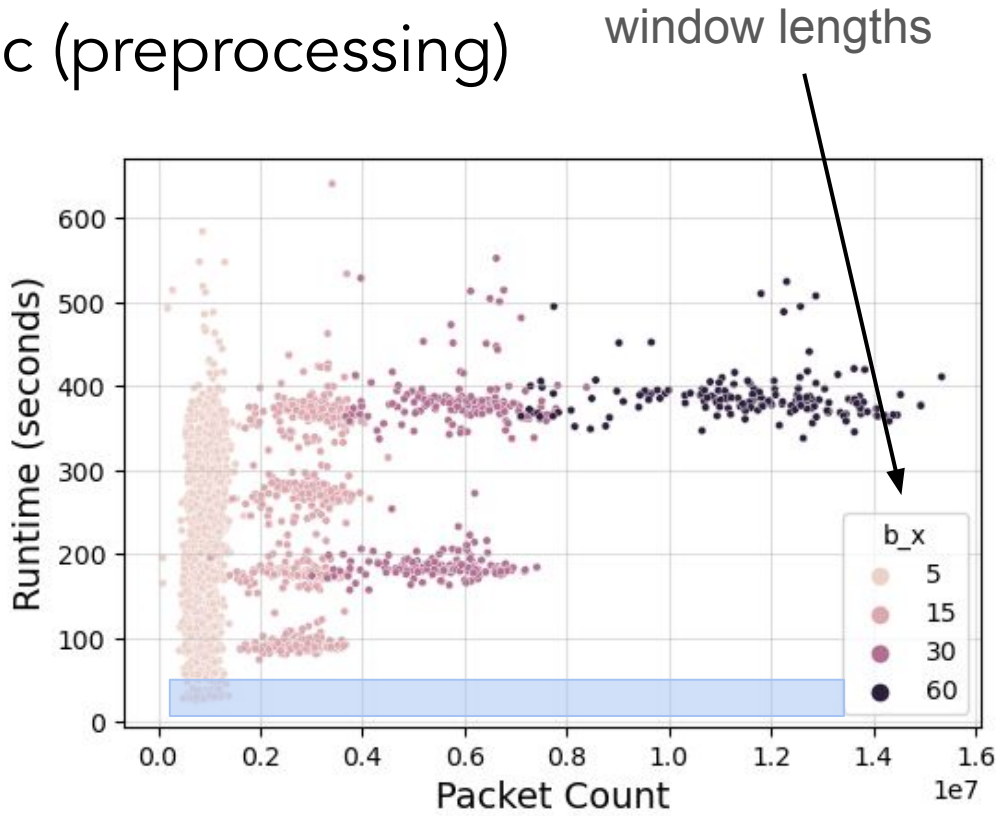
- Packet Fingerprints
  - Mirai
  - Zmap
  - Nmap
  - Unicorn
- Acknowledged Scanners
  - ASes / subnet ranges
- Historical Events
  - Mirai
  - Log4j2



Preliminary results: DarkVec (preprocessing)

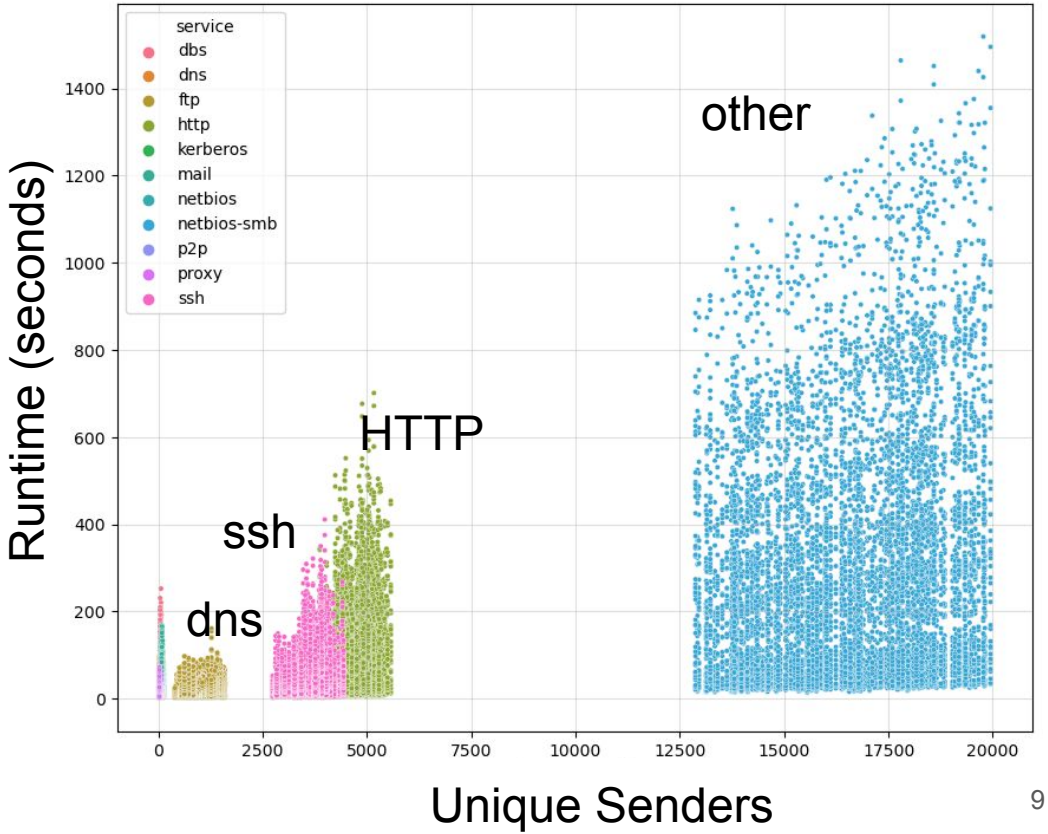
- Converting windowed PCAPs to sender sequences

- Longer window, more packets, higher preprocessing times



Preliminary results: DarkVec (training)

- Training DarkVec corpuses from sender sequences
- Runtime scale in relation to unique sender count
- More unique senders require higher training times (though noisy)



Next steps

- Replicate published algorithms with our reference datasets
  - Using full/subset of the data
  - Capture performance metrics
- Compare the “clustering”/ detection results across different algorithms
  - Evaluate accuracy, recall, … using our “ground-truth” labels