

DarkBench: A Benchmark Framework for Darknet Detection Methods

Max Gao
CAIDA/UCSD

Sakshi Deore
UCSD

Reventh Sharma
UCSD

Eric Li
UCSD

Esteban Carisimo
Northwestern University

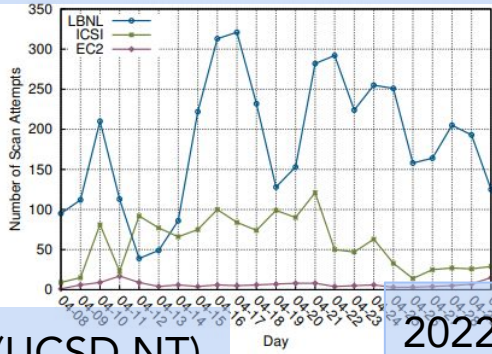
Ricky Mok
CAIDA/UCSD

kc claffy
CAIDA/UCSD

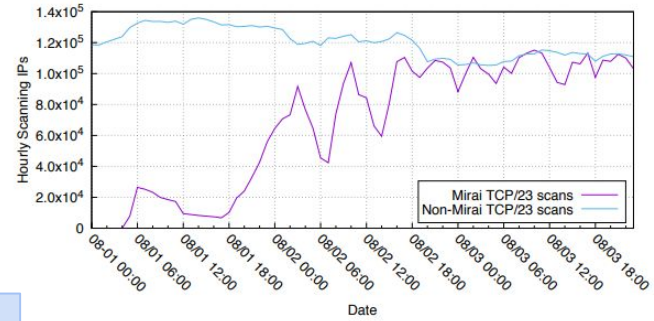


Scanning events motivate detection methods

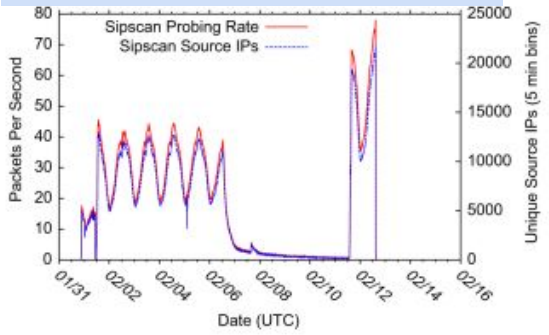
2014 - Heartbleed (LBNL+ICSI)



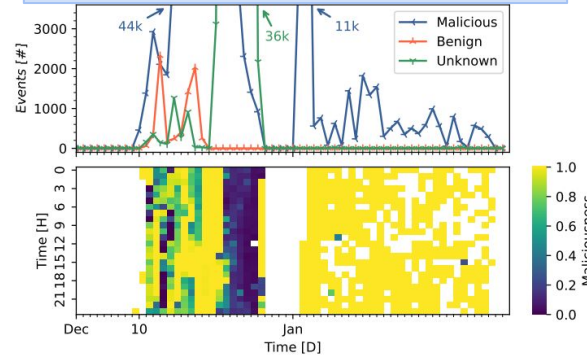
2016 - Mirai (Orion NT)



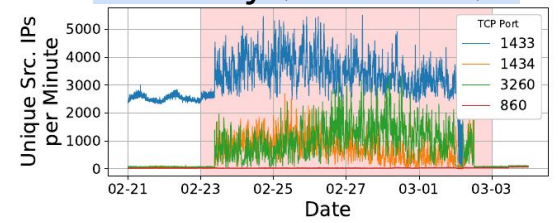
2011 - Sality (UCSD NT)



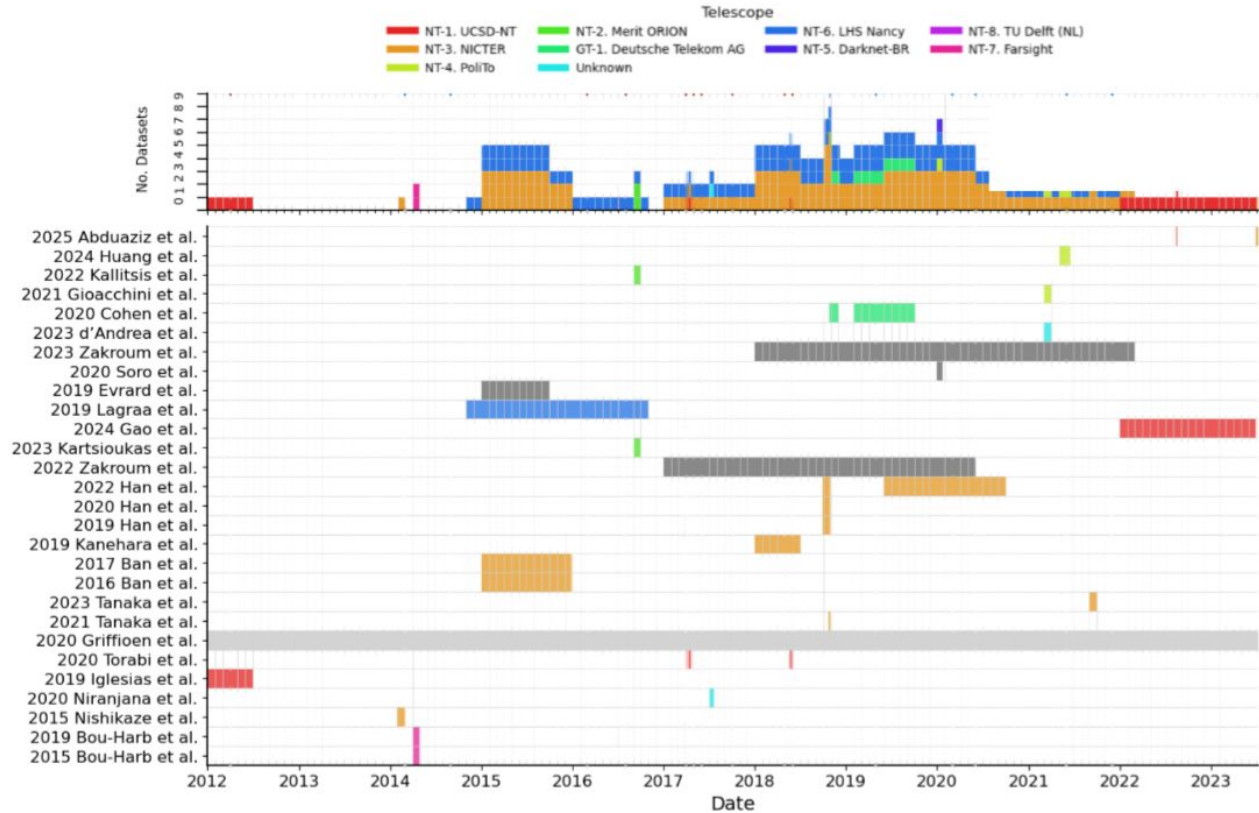
2022 - Log4j2 (UCSD NT)



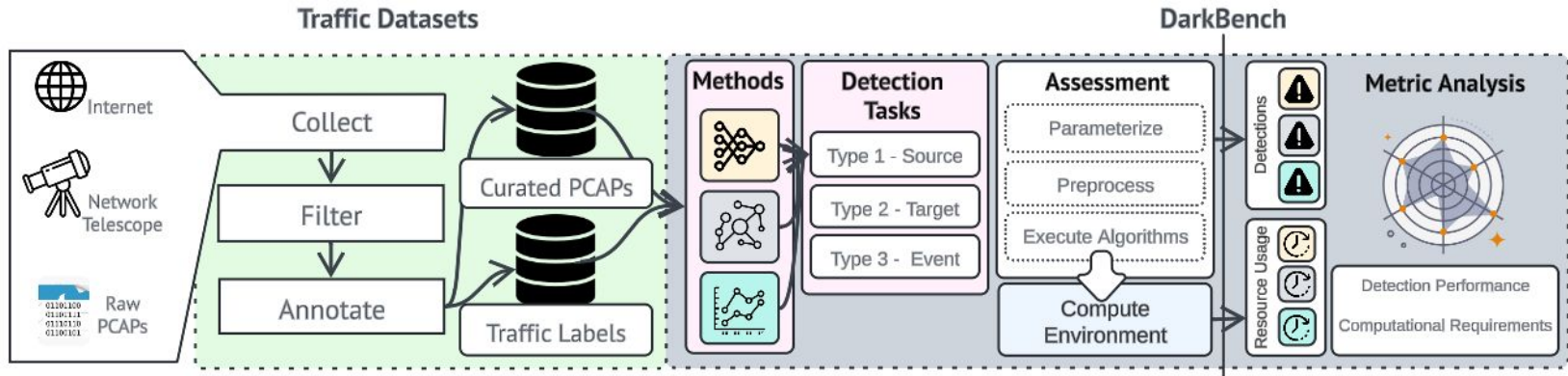
2023 - MSFT Patch Tuesday (UCSD NT)



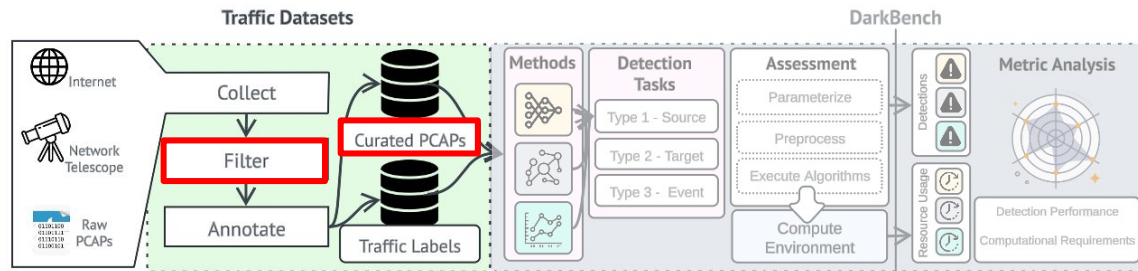
Gaps in comparative method assessments



DarkBench enables systematic, comparative assessments



Openly available, curated darknet data

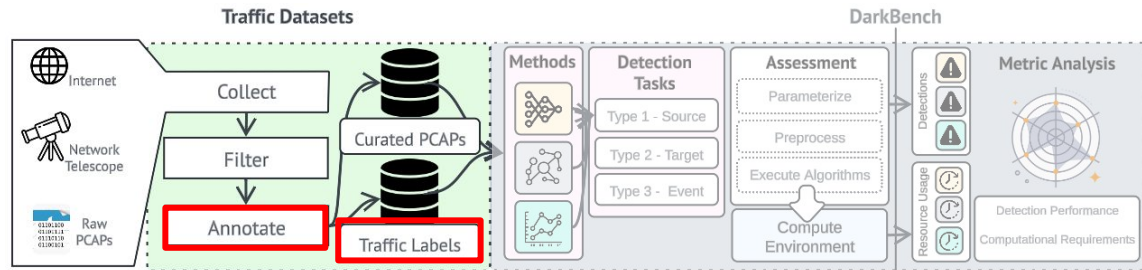


- 3 datasets - UCSD-NT (will be available through OSDF)
 - Filtered for only TCP-SYN traffic
 - Sampled traffic from 5 x /16 subnets

Dataset	Total GB
i. Aug. 1-3, 2016	41.99
ii. Mar. 27-31, 2023	752.33
iii. Apr. 1-4, 2025	449.00



Reproducible, programmatic label definitions

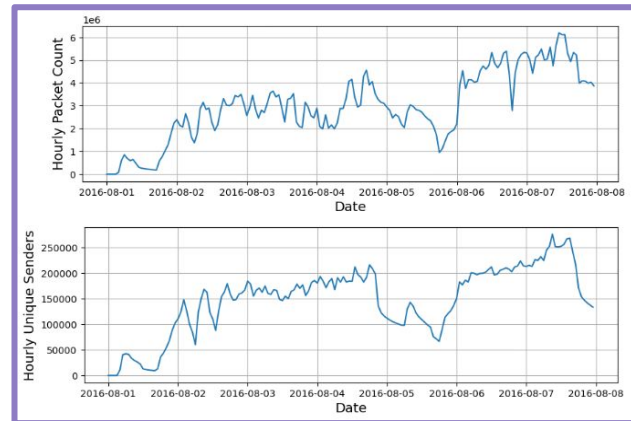
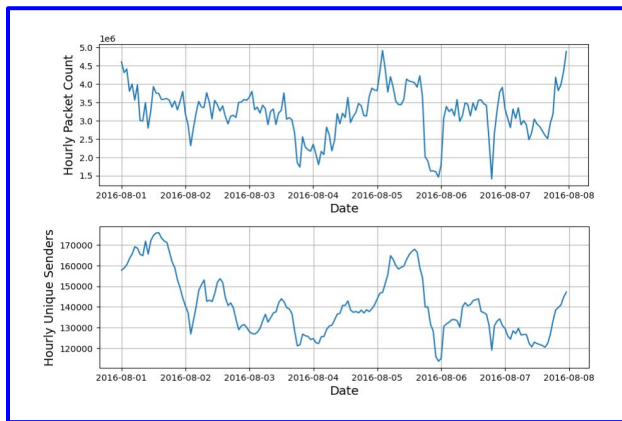


- Packet Fingerprints

- Mirai
- Nmap
- Zmap*
- Unicorn

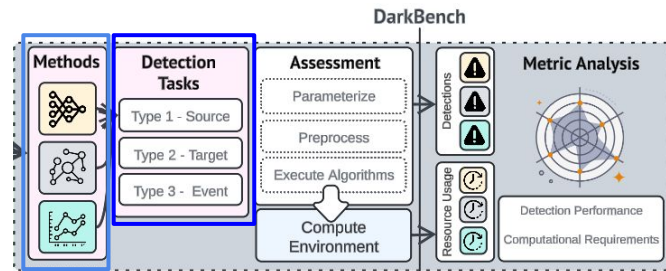
- Acknowledged Scanners

- Censys
- Shodan
- Stretchoid
- ...



* up until ZMap v4.1.0

Handful of task definitions for many methods



- Benchmarks methods on **three types of detection tasks**
 - Source-based
 - Target-based
 - Event-based
- Accommodates a **variety of methods**
 - representation learning (e.g. *DarkVec* [1], *Kallitsis* [2])
 - time series (e.g., *DarkGLASSO* [3], *DarkSIM* [4])
 - graph-mining

[1] Gioacchini et al., *DarkVec: automatic analysis of darknet traffic with word embeddings*. ACM CoNEXT, 2021.

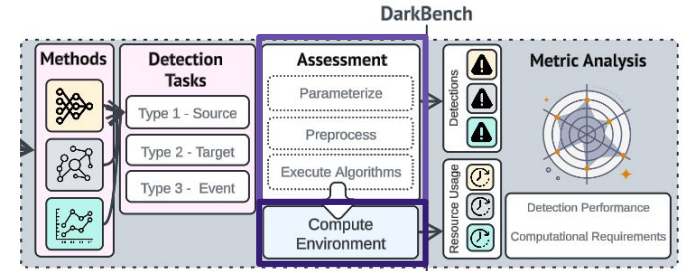
[2] Kallitsis, et al., *Detecting and Interpreting Changes in Scanning Behavior in Large Network Telescopes*. IEEE Transactions on Information Forensics and Security, 2022

[3] Han et al., *Real-time detection of malware activities by analyzing darknet traffic using graphical lasso*. IEEE TrustCom/BigDataSE, 2019.

[4] Gao et al., *DarkSim: A similarity-based time-series analytic framework for darknet traffic*. ACM IMC, 2024.

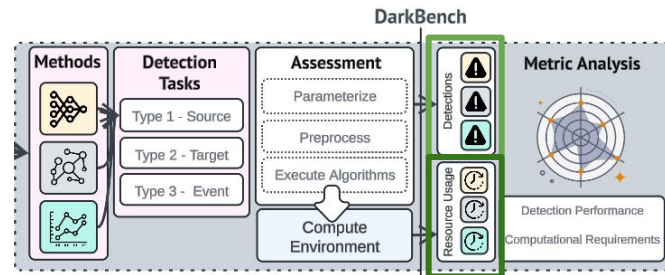
Standardized components of method assessment

- Functional stages shared across methods
 - Method parameterization
 - Traffic preprocessing
 - Algorithm execution



- Method implementations run over a common computational environment

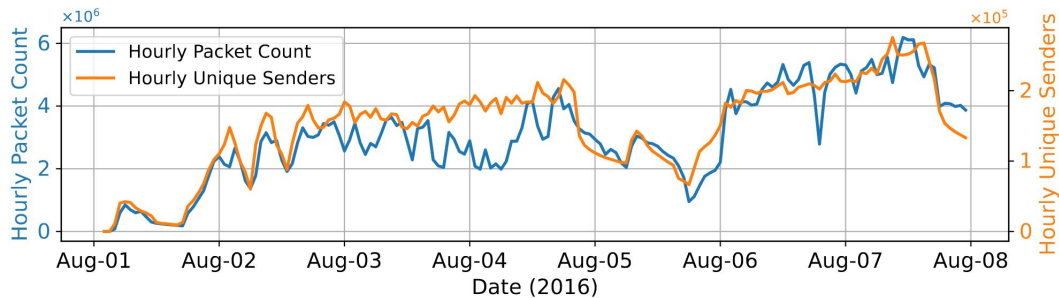
Canonical metrics for interpretable method comparisons



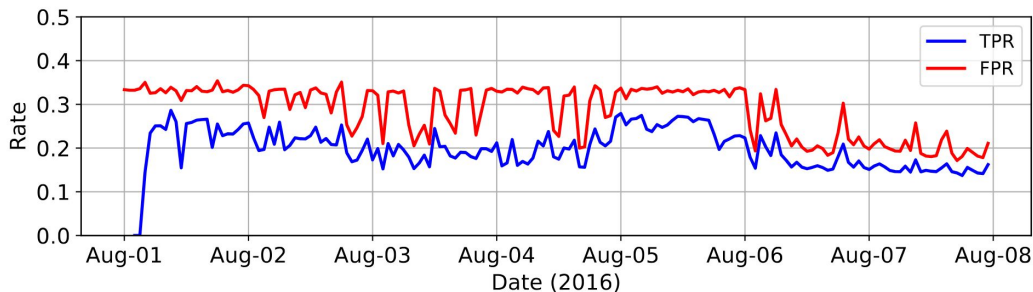
- Detection performance measured under two scenarios
 - Canonical classification metrics - Labeled traffic
 - Set (dis)similarity - Unlabeled traffic
- Computational performance measured at stage-wise granularity
 - Runtime
 - Memory Usage

Preliminary Results - Detection Performance

Mirai-hosts



DarkVec



Next steps

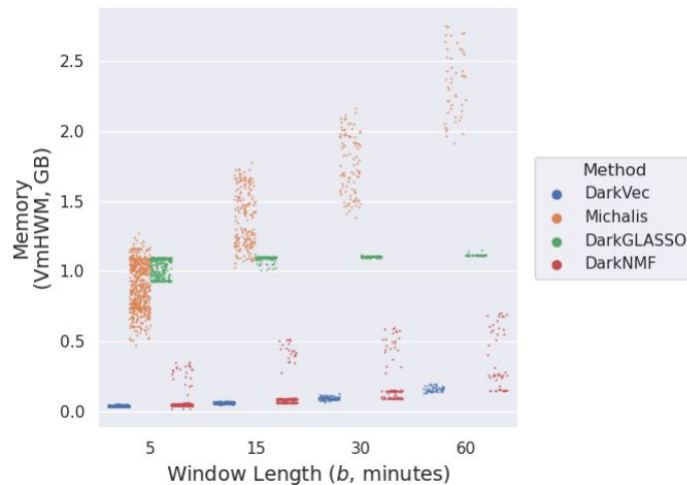
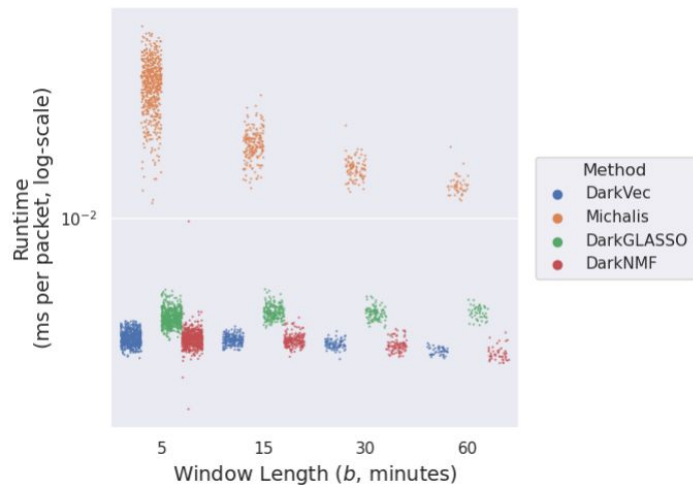
- Preliminary evaluation of the framework
 - 4 methods
 - 3 curated datasets
- Datasets distributed through OSDF
- Open-source framework and method implementations for reproducibility

Thanks! Questions?
magao@ucsd.edu

This work is based on research sponsored by U.S. NSF grants OAC-2319959, OAC-2531134. The views and conclusions are those of the authors and do not necessarily represent endorsements, either expressed or implied, of NSF.



DarkBench preliminary results - detection performance



DarkBench -

- Sample a variety of methods 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

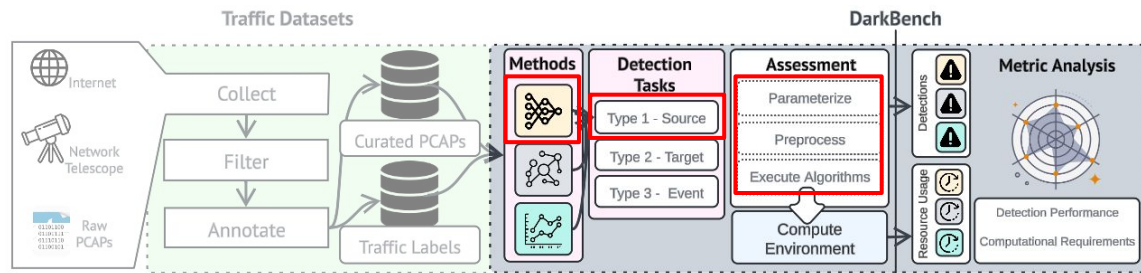
[1] M. Kallitsis, et al., *Detecting and Interpreting Changes in Scanning Behavior in Large Network Telescopes*. IEEE Transactions on Information Forensics and Security, 2022

[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.

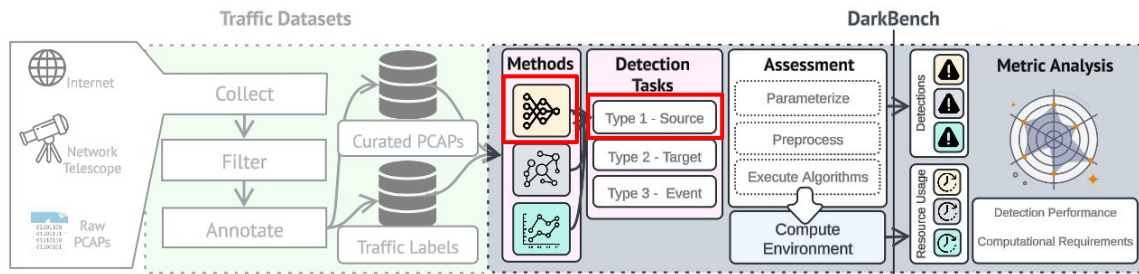
[4] C Han, et al., *Dark-TRACER: Early Detection Framework for Malware Activity Based on Anomalous Spatiotemporal Patterns*. IEEE Access, 2022

DarkBench pilot evaluation



- 4 detection methods
- 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
- SDSC Expanse used

DarkBench pilot evaluation



- 4 detection methods
- 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
- Run implementations on