Labeling Network Telescope Data: Challenges and New Directions

Michalis Kallitsis
Merit Network, Inc. , University of Michigan
Joint work with Stilian Stoev, Zakir Durumeric, John Yen,
Vasant Honavar, Dinghao Wu, Rupesh Prajapati

DINR 2023
Feb 22, 2023

Challenges in Labeling Network Telescope Data

- **Darknet traffic definition:** traffic destined to an *unused* but *routed* address space
- Darknets observe *unidirectional traffic*, completely passive operation
  - No payload can be collected in TCP traffic (about 90% of all traffic is TCP)
- *Large volumes* of data
  - More than 100GB compressed PCAP per day
- Complex traffic, *dynamically changing* based on new vulnerabilities found, new malware, etc.

Labeling efforts at Merit's Network Telescope

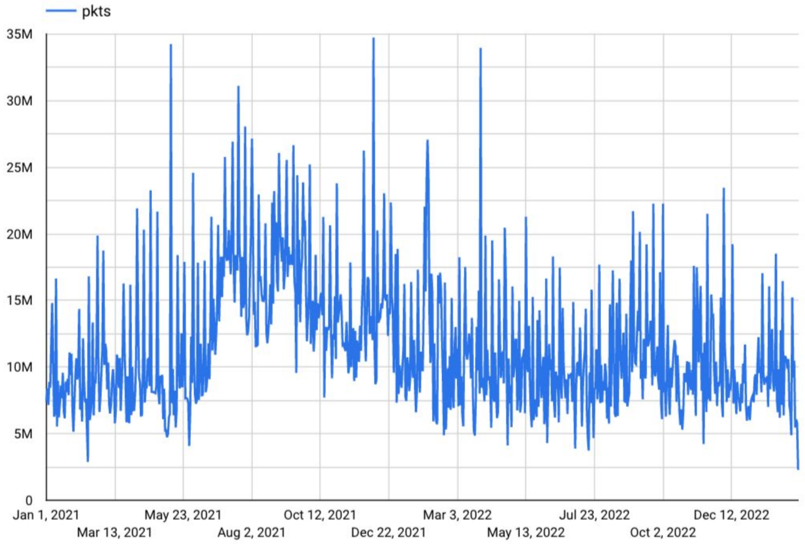
A. Labeling by ***traffic type*** (e.g., backscatter versus scanning)
   a. This can help us quickly identify randomly-spoofed denial-of-service attacks (RsDoS)

B. Labeling by known ***fingerprints***
   a. Mirai
   b. Masscan
   c. Zmap

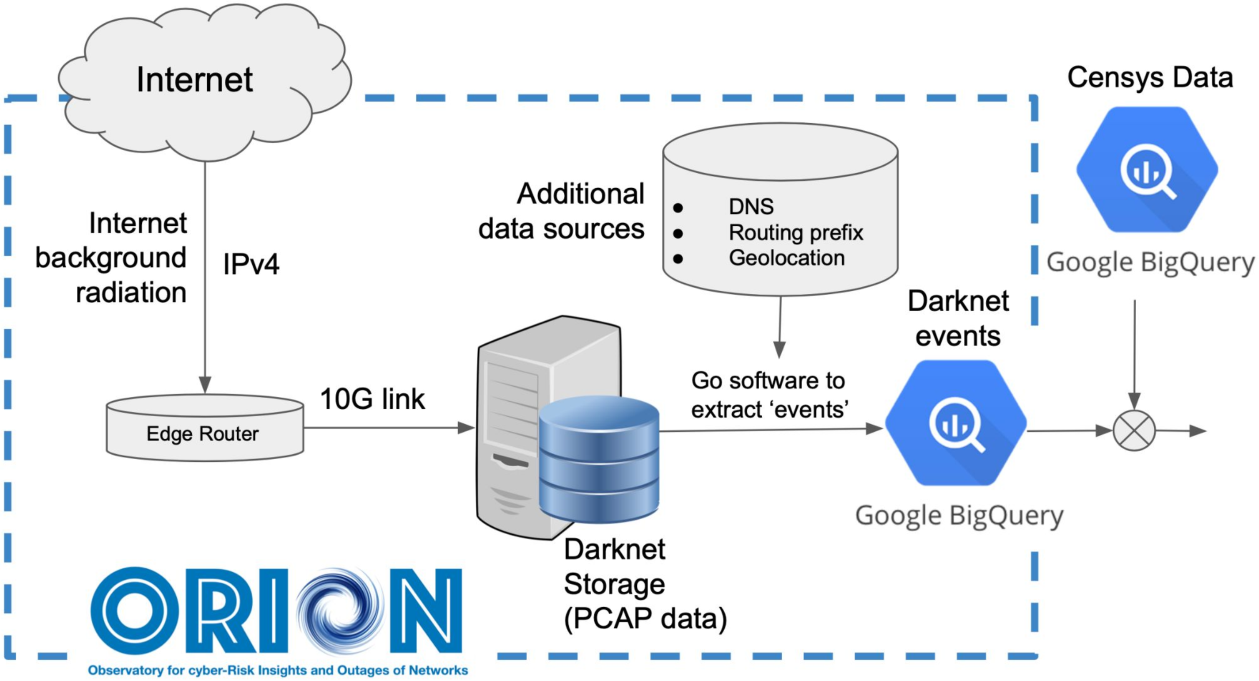
C. Using unsupervised machine learning techniques to ***cluster the data***
   a. Cluster data on, e.g., daily basis

Disclaimer: Discussion is non-DNS focused

- Although with appropriate adjustments everything discussed applies to DNS data



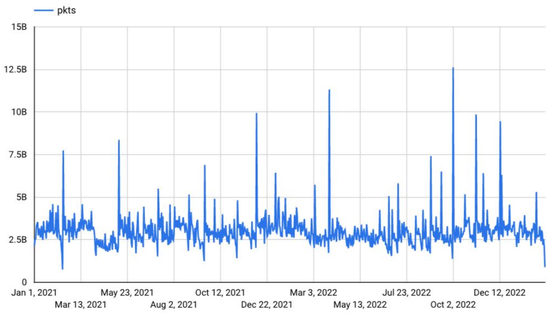
ORION's near-real-time data pipeline



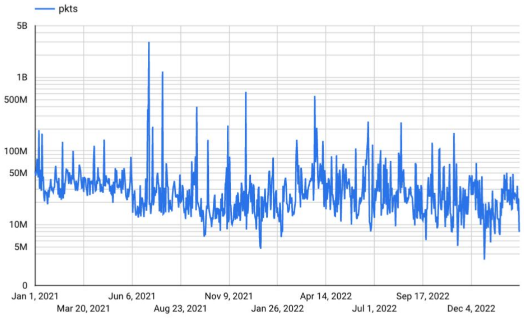
A. Labeling by Traffic Type

- Protocol fields in TCP and ICMP can help us glean insights about various traffic types
  - Scanning attempt: TCP SYN, ICMP Echo Request
    - UDP is usually scanning but we have seen events of DDoS attacks against our Darknet!
  - Backscatter: TCP SYN+ACK or TCP RST

ORION Scanning Packets / Day

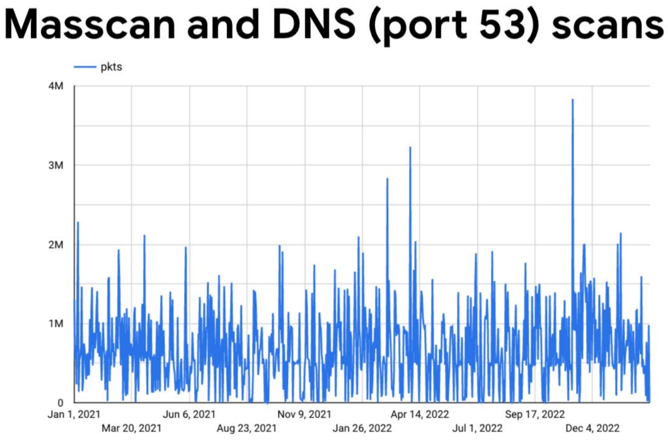


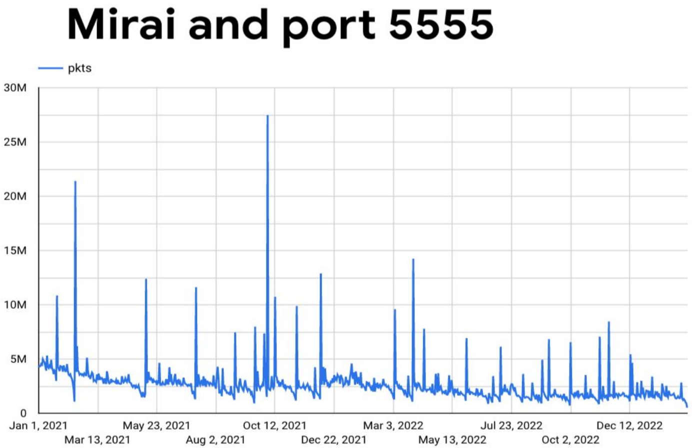
ORION Backscatter Packets / Day



B. Labeling by Known Fingerprints

- Telltale fingerprints: Mirai [Antonakakis et al.], Zmap and Masscan [Durumeric et al.]
  - Remark #1: Zmap's latest release has changed their fingerprint
  - Remark #2: Some of this labels might happen by chance (1 / 2**16 chance to assign Zmap label erroneously)





C. Labeling using AI/ML techniques

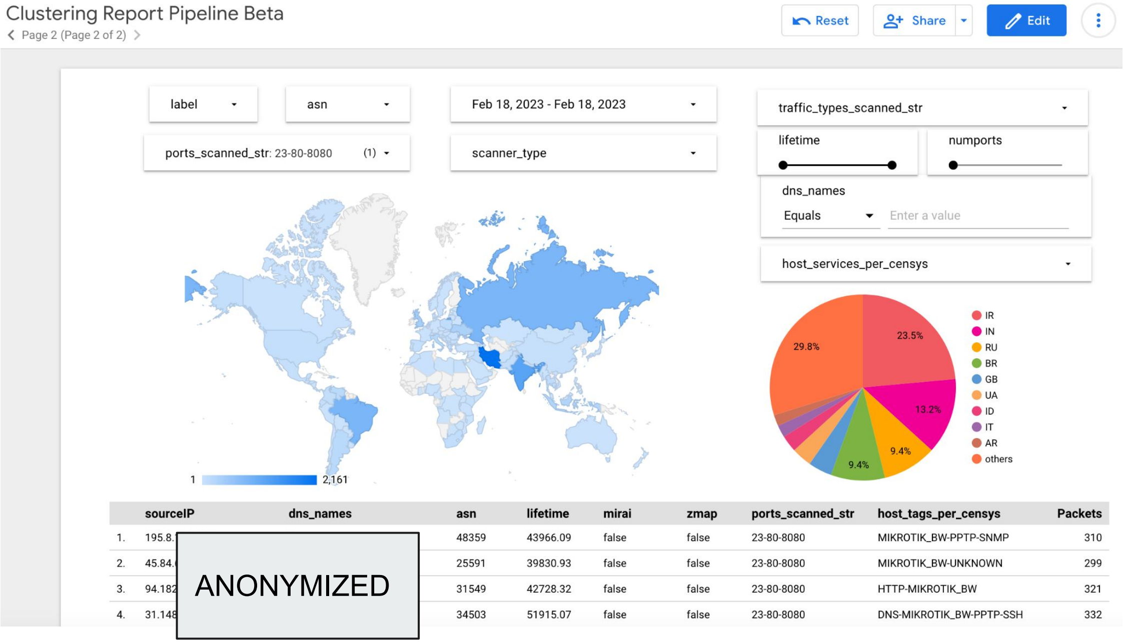
- Leverage *unsupervised learning techniques* to find clusters in data
- Cluster the *Darknet IPs* based on some network *features*
- Key step #1: engineering meaningful features to characterize IP behavior
  - Examples: set of ports scanned, scanning intensity
- Key step #2: encode these features into a space of *embeddings*
  - Lower dimensional space to perform the clustering on
  - Initial space has both *numeric* and *categorical* features
- M. Kallitsis, R. Prajapati, V. Honavar, D. Wu and J. Yen, "Detecting and Interpreting Changes in Scanning Behavior in Large Network Telescopes," in IEEE Transactions on Information Forensics and Security, vol. 17, pp. 3611-3625, 2022, doi: 10.1109/TIFS.2022.3211644.

Cluster Identification: Case Study 2022-02-20

TABLE V: Cluster Inspection (2022-02-20).

Description	# of Clusters	# of Senders
Mirai-related	70	108,912
Unknown	67	76,525
SMB	20	23,700
Heavy Scanners	19	2,377
ICMP scanning	5	2,619
Ack Scanners	4	795
SSH scanning	4	2,635
censys.io	3	147
TCP/3389 (RDP)	2	1,482
UDP/5353	2	3,212
Backscatter (DDoS)	2	815
TCP/6379 (Redis)	1	437
Normshield	1	253
TOTAL	**200**	**223,909**

Clustering Dashboard



Detect Temporal Darknet Changes via Clustering

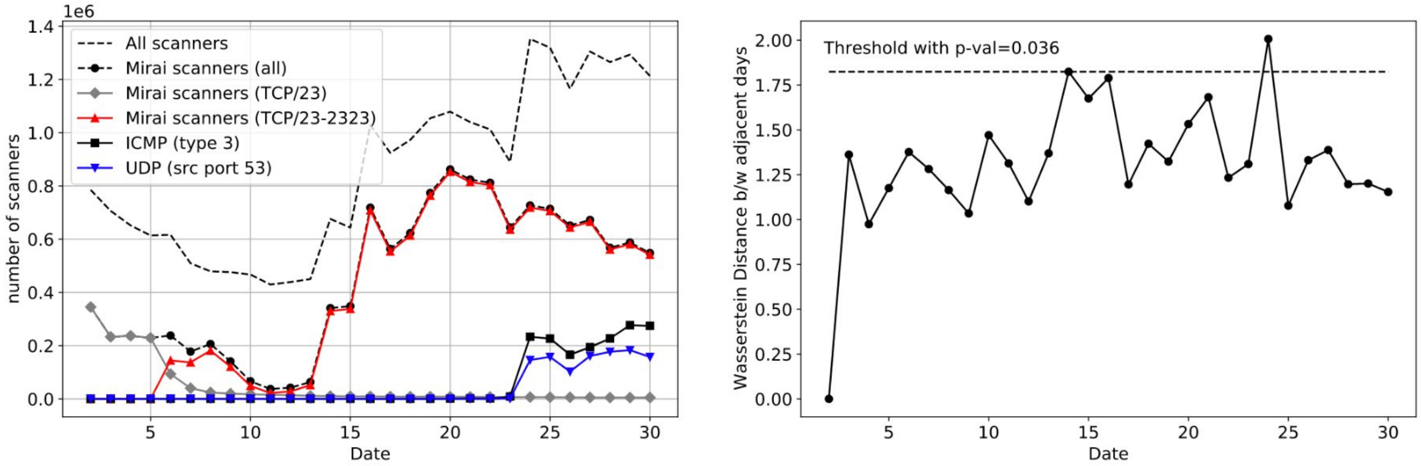


Fig. 1: (Left panel) Scanning traffic at Merit's Darknet (a /10 Darknet, back then) for September 2016. Notice the expansion of the Mirai botnet, namely the addition of TCP/2323 in the set of ports scanned. The figure considers scanners emitting at least 50 packets per day. (Right panel) Detection of temporal changes in the Darknet using the Wasserstein distance.

Interpret Clustering Changes via Optimal Transport

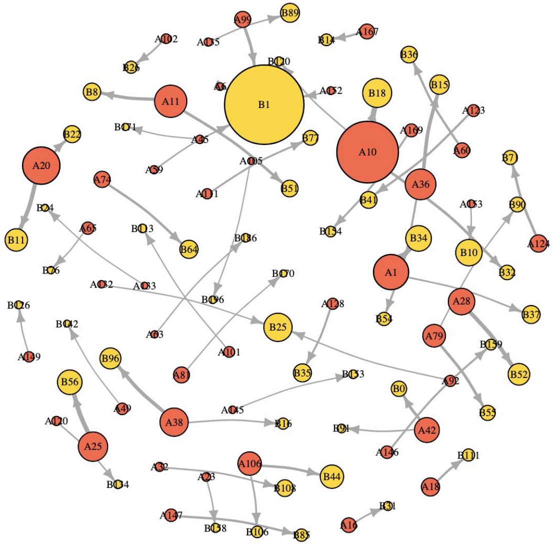


Fig. 6: Optimal transport plans for Sept. 13–14. Only edges with $\gamma^*_{uv} \geq 0.01$ are shown.

- Leverage the outcome of *Optimal Transport Plan* (see Earth Mover's Distance problem too) to interpret clustering changes
- Identify if the change is due to changes in *existing / known scanners* or due to a new *emerging vulnerability*!

Conclusions and Next Steps

- ORION network telescope's labeling efforts
  - Traffic types, known fingerprints, AI/ML methods (clustering)
- How might we *integrate* more data?
  - What other labels exist?
    - GreyNoise data, others?
    - How can we link Darknet data to specific vulnerabilities / CVEs?
- How might we *share* data and who are potential "consumers"?
  - Threat intelligence sharing protocols [Kampanakis]
    - TAXII (Trusted Automated Exchange of Indicator Information
    - STIX (Structured Threat Information Expression)
- Allow others to "plug" their code into our analysis pipeline
- Darknet data for research:
  - NSF CLASSNET: https://comunda.isi.edu/
  - https://www.merit.edu/initiatives/orion-network-telescope/