What is Normal?
A Big Data Observational Science Model of Anonymized Internet Traffic

Jeremy Kepner

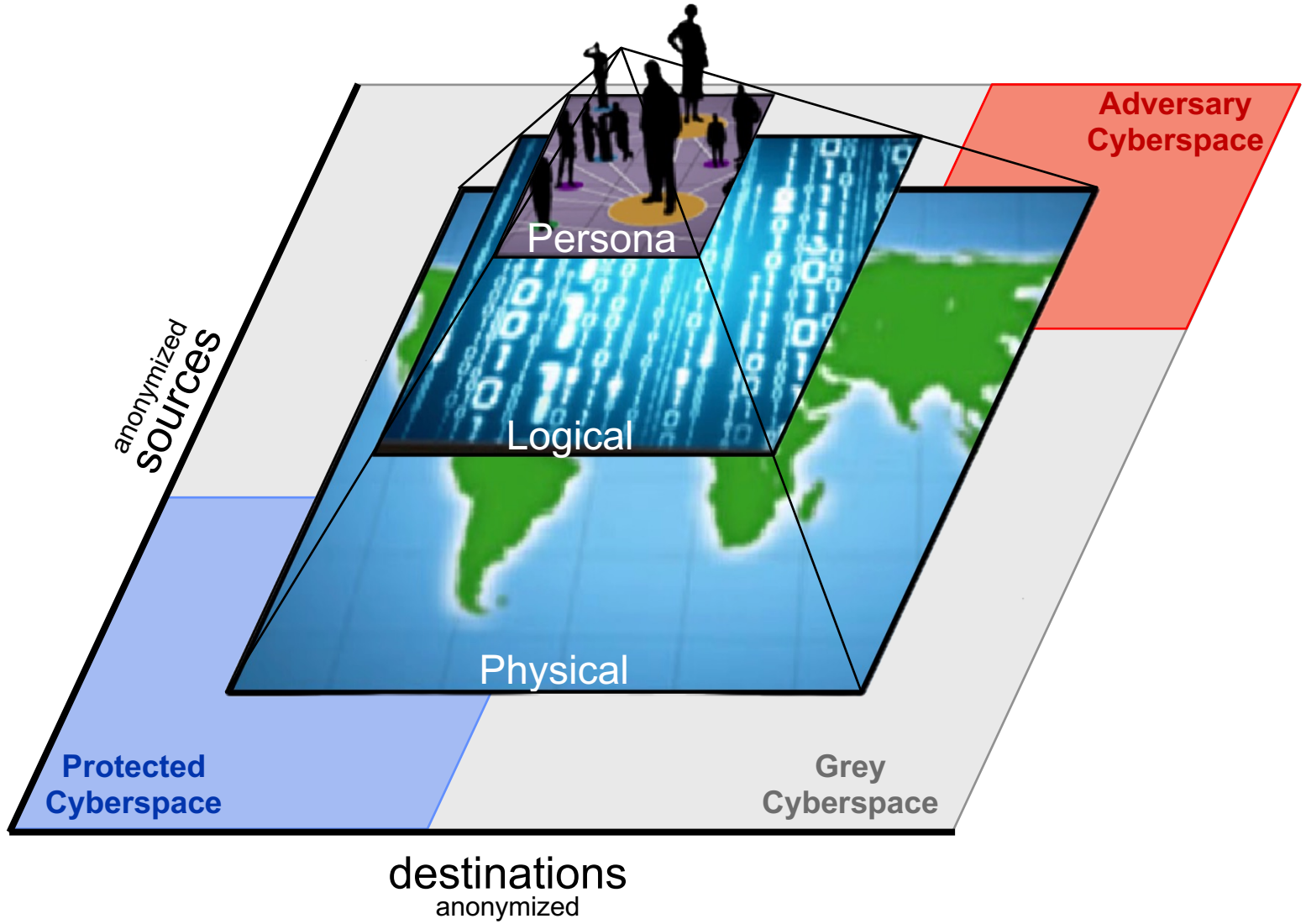
November 2025

MIT LINCOLN LABORATORY
SUPERCOMPUTING CENTER

Research was sponsored by the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Problem We Are Trying To Solve

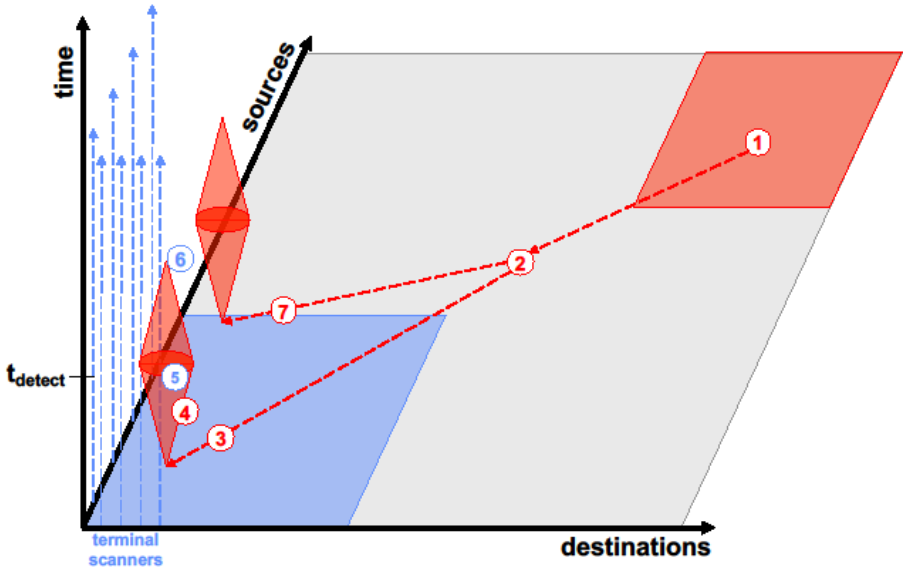
- Problem
  - Need better sensors and AI algorithms to improve networks
- Approach
  - New cyber sensors
  - Novel privacy-preserving hierarchical AI techniques
  - Assembled the largest corpus of publicly available network data to support this research (100+ trillion events)



Expected Impact: Pro-Active Cyber

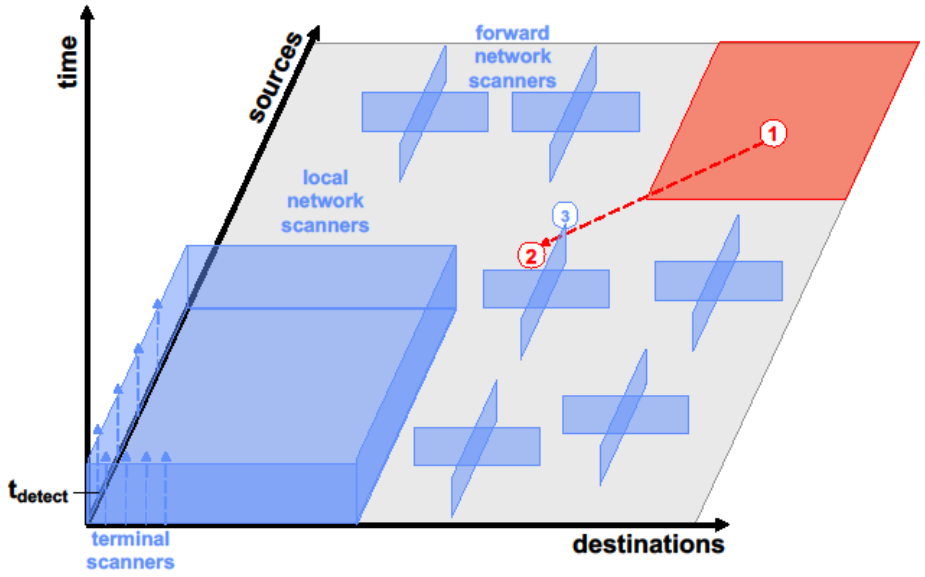
Traditional Cyber

1. Plan
2. Stage
3. Infiltrate
4. Move laterally
5. Detect
6. Cleanse
7. Infiltrate



Pro-Active Cyber

1. Plan
2. Stage
3. Detect



- Move sensors to outside of protected enclave
- Provide lead time to cyber defenders

Traffic Matrix Tutorial

Which attack stage is most relevant to the observed traffic? Hint: youtu.be/_HDplx2MxyA or

stage

infiltrate

move laterally

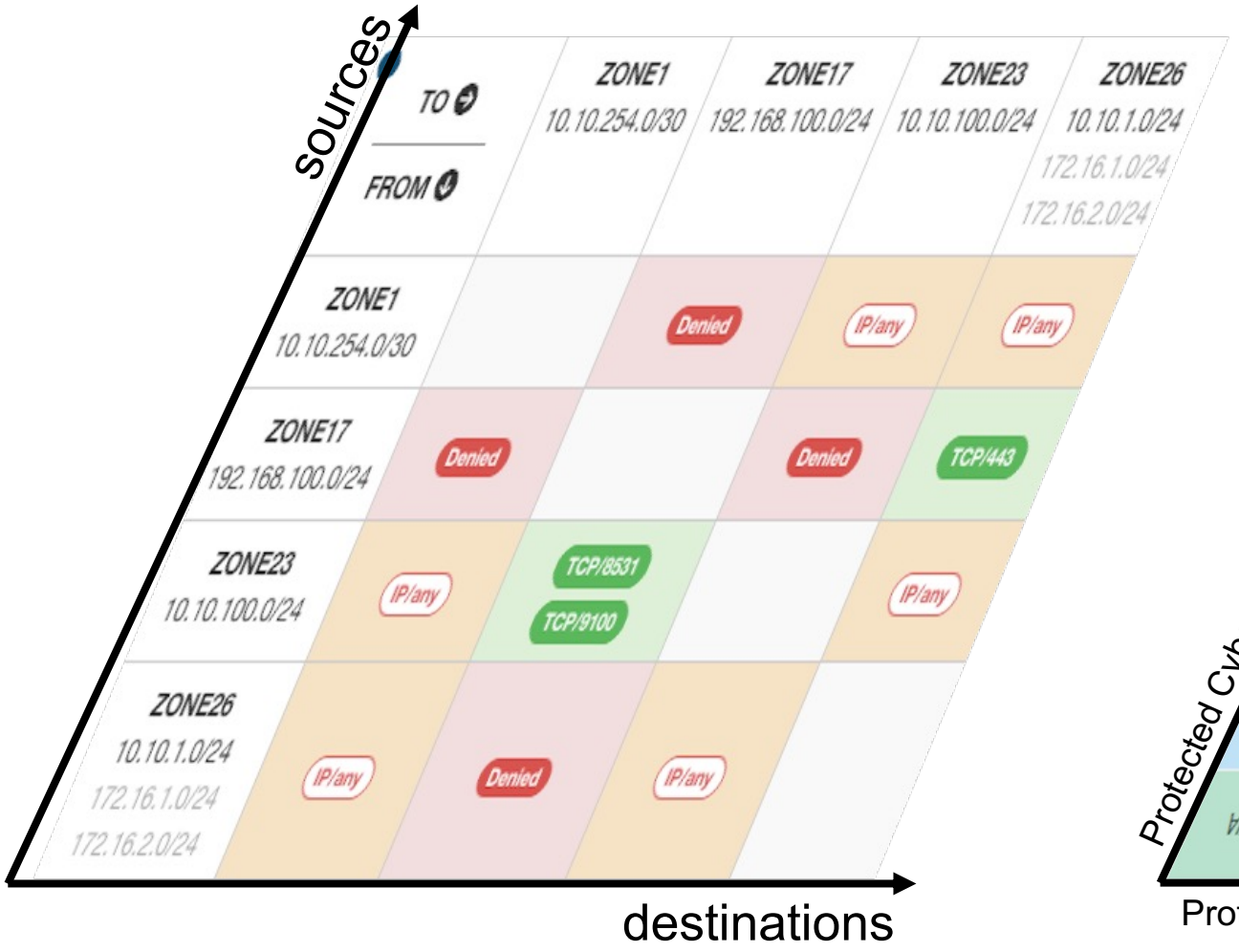
Spacebar - Switch View

Q/E - Rotate screen

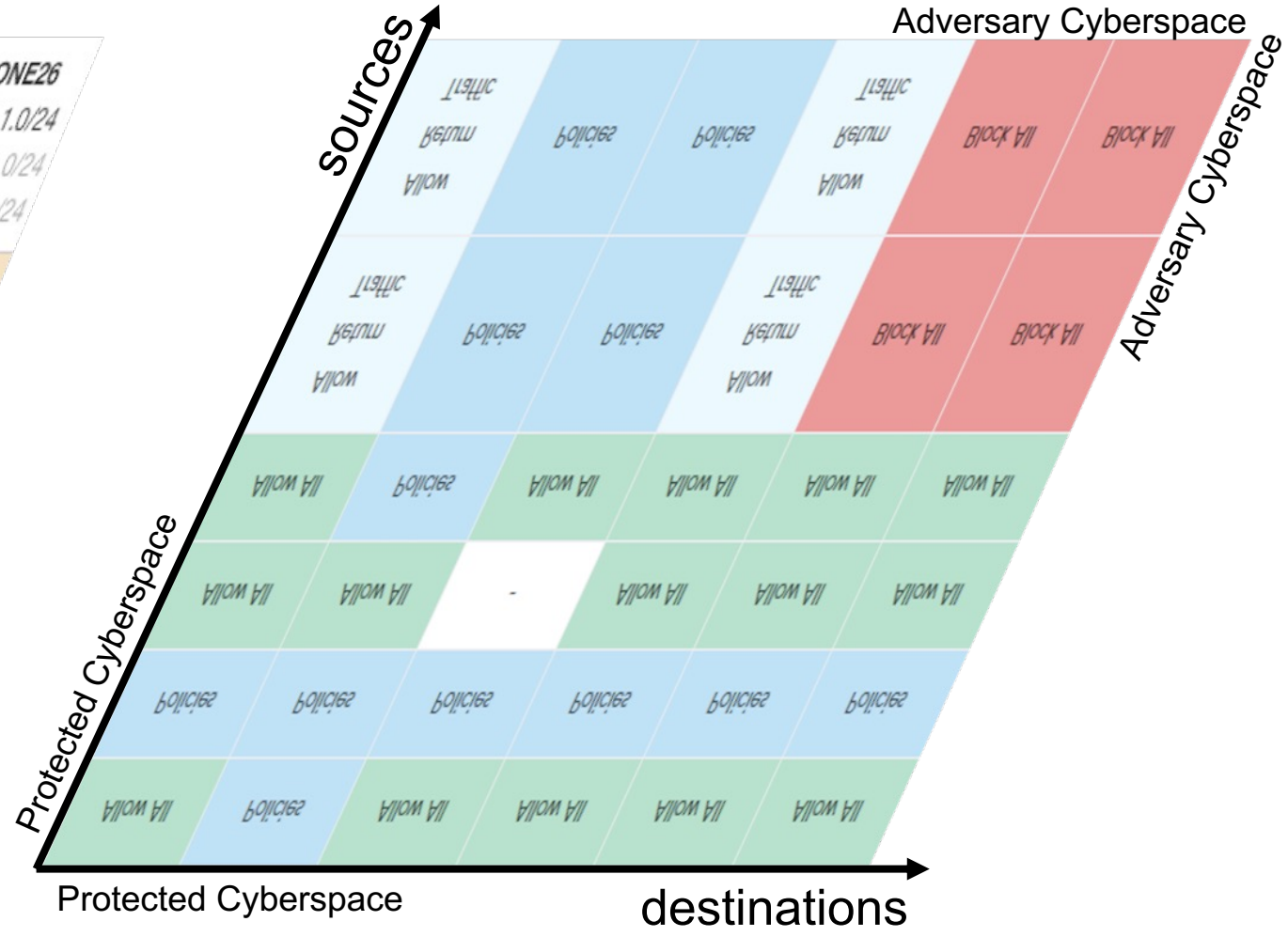
Teaching Network Traffic Matrices in an Interactive Game Environment, IEEE GrAPL 2024

Industry Firewall Zone Matrix Adoption

Ubiquiti UniFi



Dragos Network Perception



https://help.ui.com/hc/en-us/articles/115003173168-Zone-Based-Firewalls-in-UniFi
https://www.network-perception.com/kb/zone-matrix

Anomaly Detection and Observational Science

- **Core challenge for anomaly detection systems is adequate models of normal**

  The concept of normality. It is one of the main steps to build a solution to detect network anomalies. The question ``how to create a precise idea of normality?" is what has driven most researchers into creating different solutions through the years. This can be considered as the main challenge related to anomaly detection and has not been entirely solved yet.[1]

- **Reproducible observations of cyberspace have been recommended as a core foundation for the science of cyber-security**

  The highest priority should be assigned to establishing research protocols to enable reproducible [observations][2]

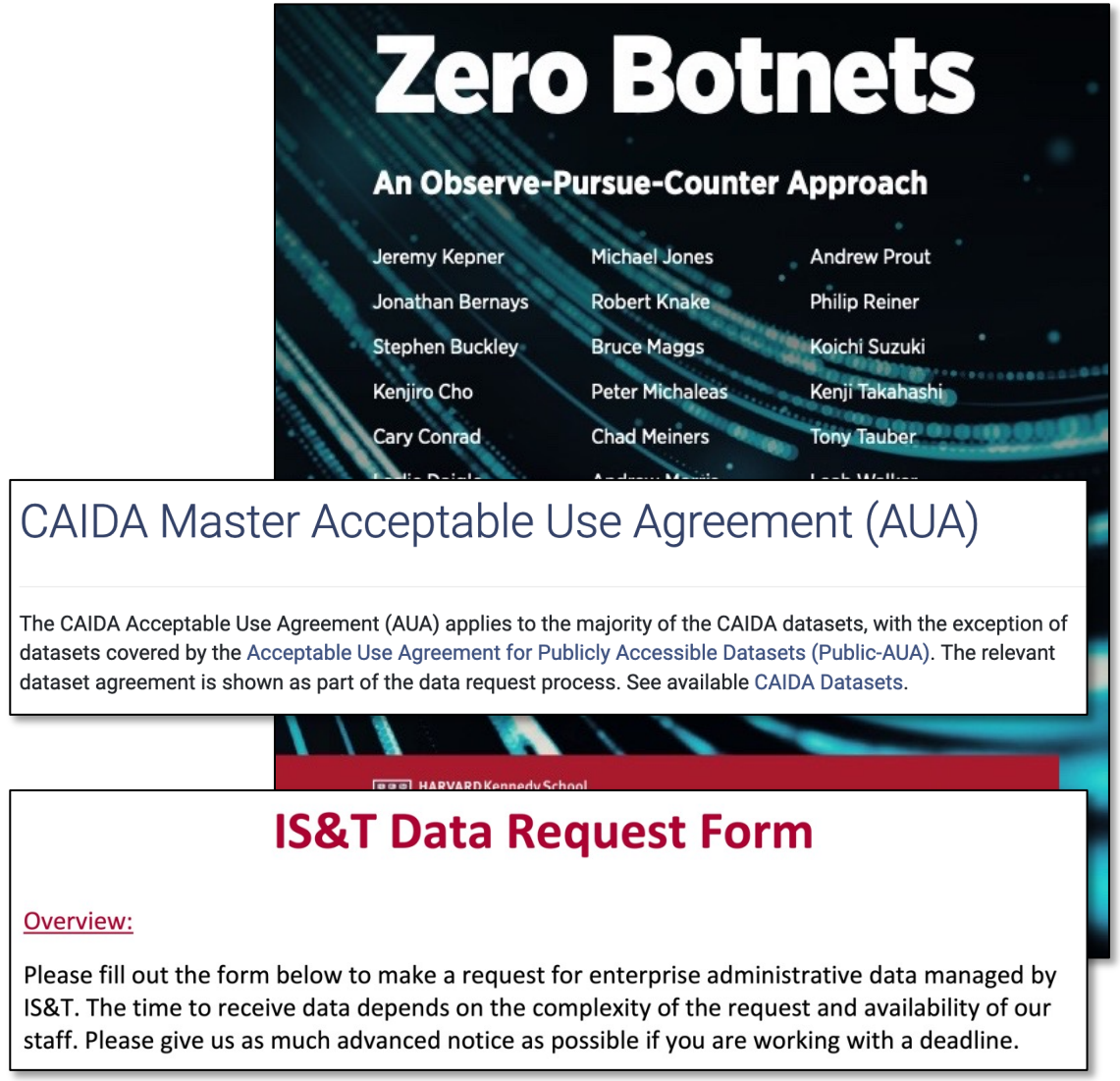
Other domains (land, sea, undersea, air, and space) rely on detailed observational science models of their environment to understand what is normal

[1] *A Comprehensive Survey on Network Anomaly Detection,* Fernandes et al, Telecommunication Systems, 2019
[2] *Science of Cyber-Security*, JASON, 2010

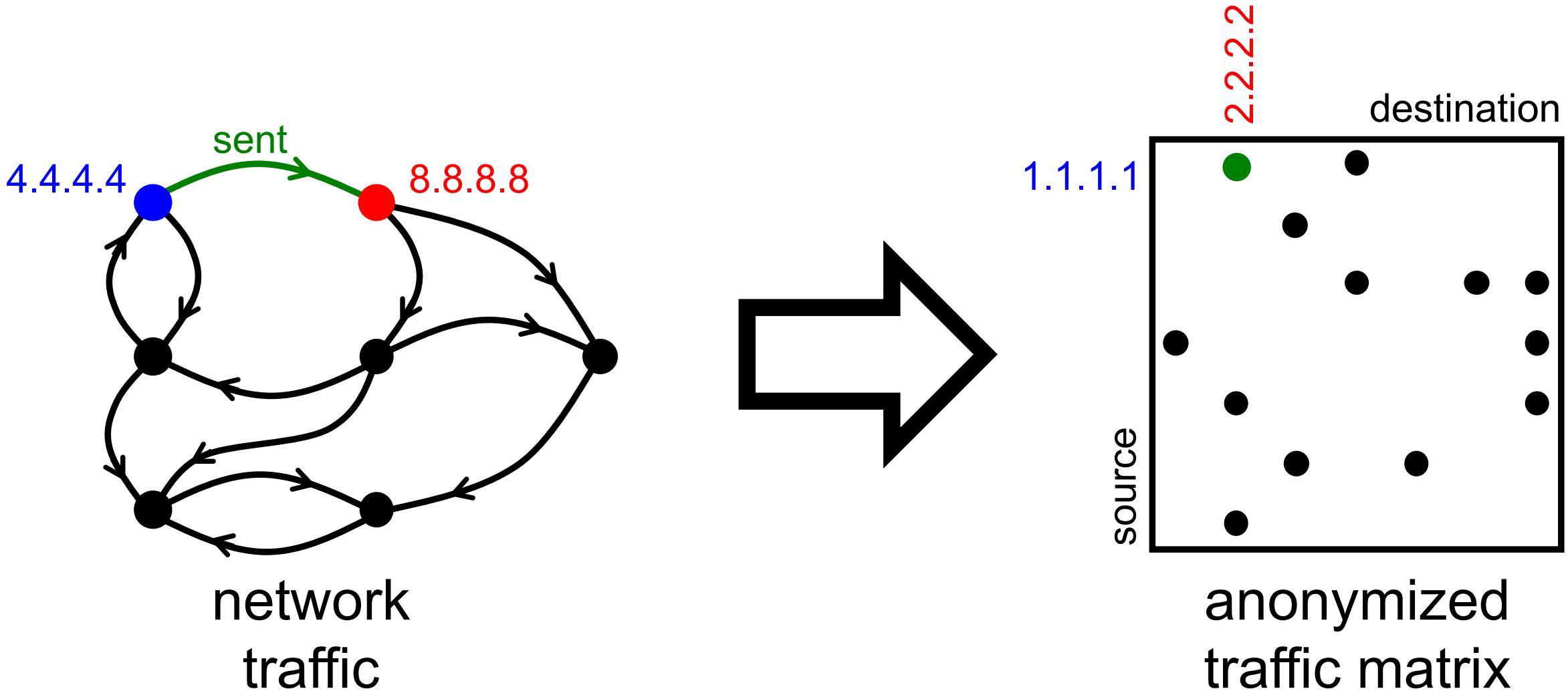
AI Ready Privacy Preserving Data Request Agreements

- Network operators/owners collect, analyze, and share data only to improve network operations/security
  - Data recipients inherit that intent

- CAIDA & MIT have pioneered new approaches that approve appropriately anonymized data with clear who/what/where/why/how agreements
  - MIT now has the largest collections of data because of our unique ability to analyze these data at scale while protecting privacy

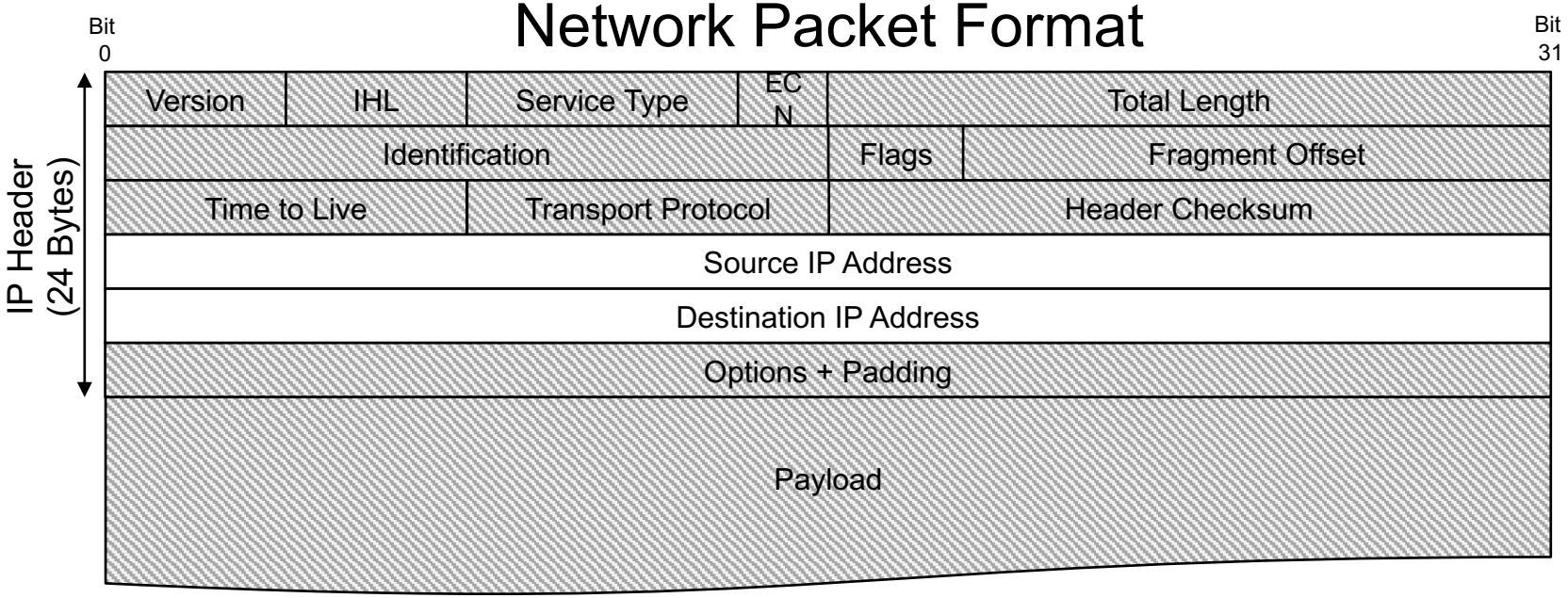


Network Traffic Messages to Anonymized Traffic Matrix



Keep Only Source/Destination IPs

Network Packet Format



Source IP Address	Destination IP Address

Sample Packet

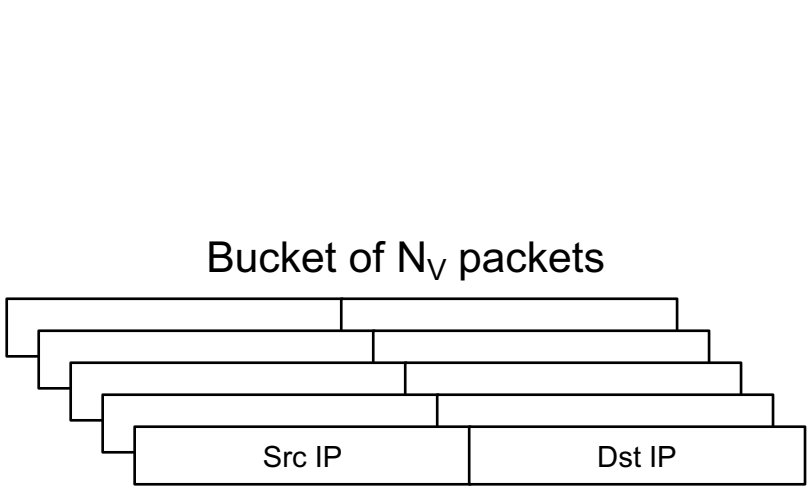
```
45 00 01 13 00 00 40 00 40 06 18 46 c0 a8 c8 1a
34 3d 64 9f c5 2e 01 bb df ea e7 4f de 89 d2 e5
80 18 08 00 22 96 00 00 01 01 08 0a 49 67 5e 0a
86 23 3f 61 17 03 03 00 da 14 20 61 3a df ec a7
ff ba 29 05 86 7a e7 76 d1 de 2a a3 4e ae fa ea
bb fd 01 71 fe 8e 93 de ab 20 7a b5 8d 81 a2 e2
48 69 0b 35 f5 e1 8f e9 31 37 4b 4e 0f 97 44 ac
4e b3 f5 11 dd 0e 0e 49 4e 95 79 77 e9 a1 b3 7a
e5 9e 8b 88 80 38 4a 0f a5 a5 e9 67 ef 7e b2 59
4e 0e 8f b0 ea a4 38 e0 e9 de 1e d4 f5 80 f1 d5
a8 db 95 04 a0 5f 50 d6 6e 97 9b 9f dd 82 14 26
a8 e8 fe 47 02 05 2d 33 28 37 87 3e 9e e9 f4 d5
65 84 45 78 e0 23 3f 34 94 2e 99 3e d5 05 ac 25
65 6b 78 11 fd 7b 08 a1 6e f3 47 62 3a e7 8a 43
5a 77 ef 6a 88 d0 46 a9 de 19 34 05 4e 69 b2 40
56 a9 8f f5 d4 d6 21 e0 16 77 e9 35 a9 d0 20 4d
ef e2 d0 8f f4 15 7e 3a f6 d2 93 aa 65 34 be b4
a7 d0 7a
```

192.168.200.26	52.61.100.159

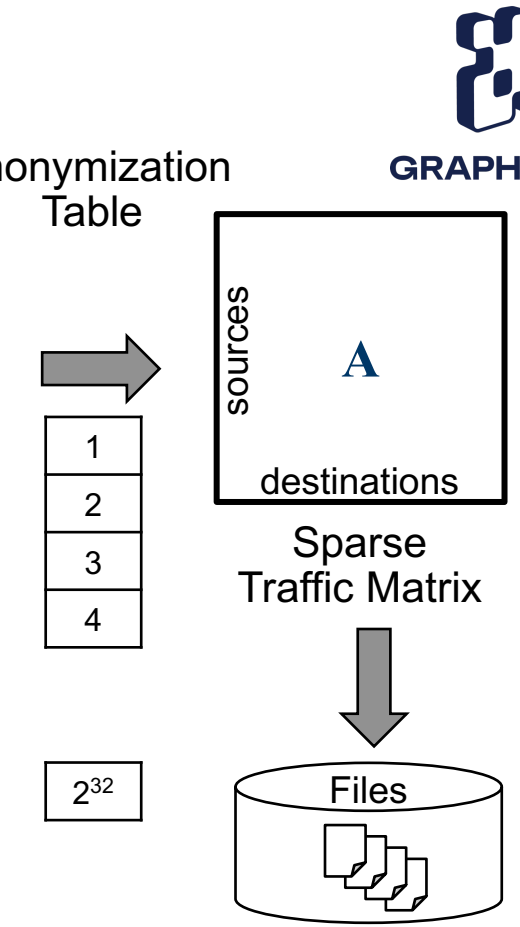
ECN: Explicit Congestion Notification
IHL: Internet Header Length

GraphBLAS on the Edge: Anonymized High Performance Streaming of Network Traffic, Jones et al, IEEE HPEC 2022 Outstanding Paper Award

Construct Anonymized Traffic Matrix



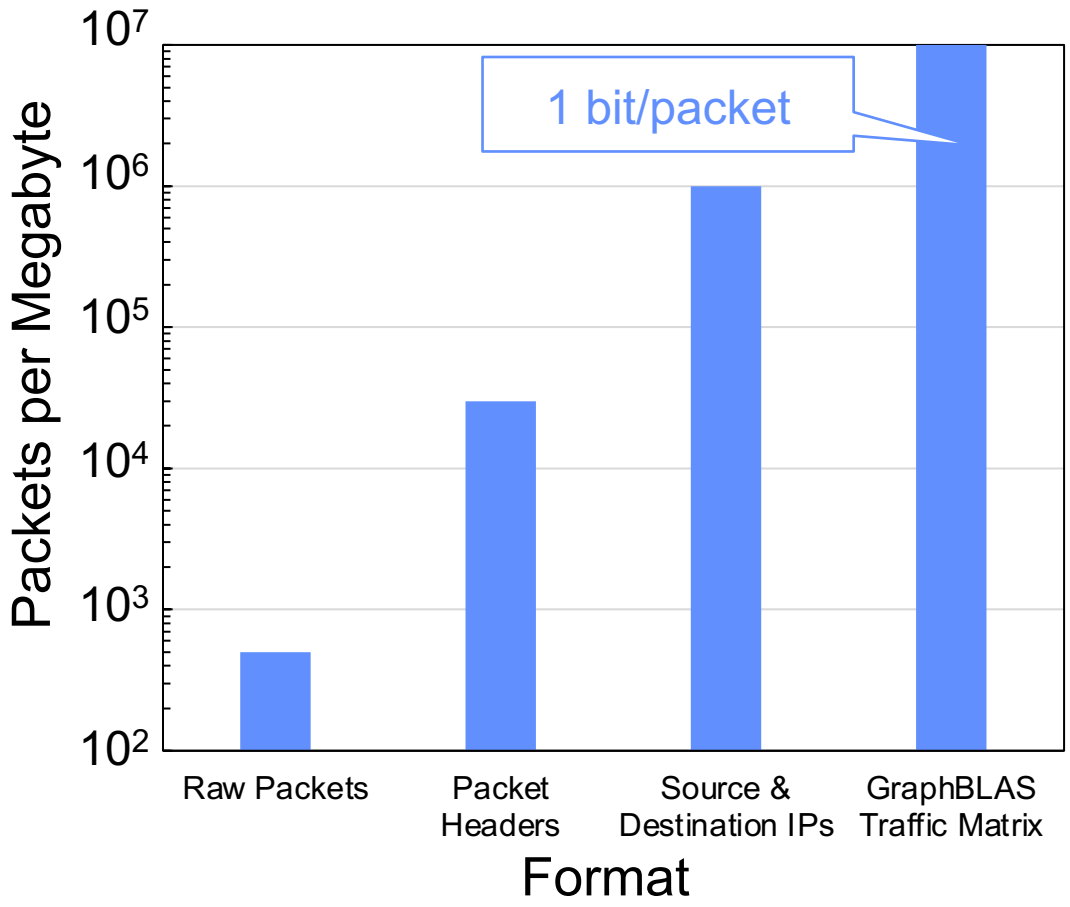
Timestamp	Source IP	Destination IP
t0	Source IP1	Destination IP1
t1	Source IP2	Destination IP2
t2	Source IP3	Destination IP3
t3	Source IP4	Destination IP4
t4	Source IP5	Destination IP5



GraphBLAS on the Edge: Anonymized High Performance Streaming of Network Traffic, Jones et al, IEEE HPEC 2022 Outstanding Paper Award

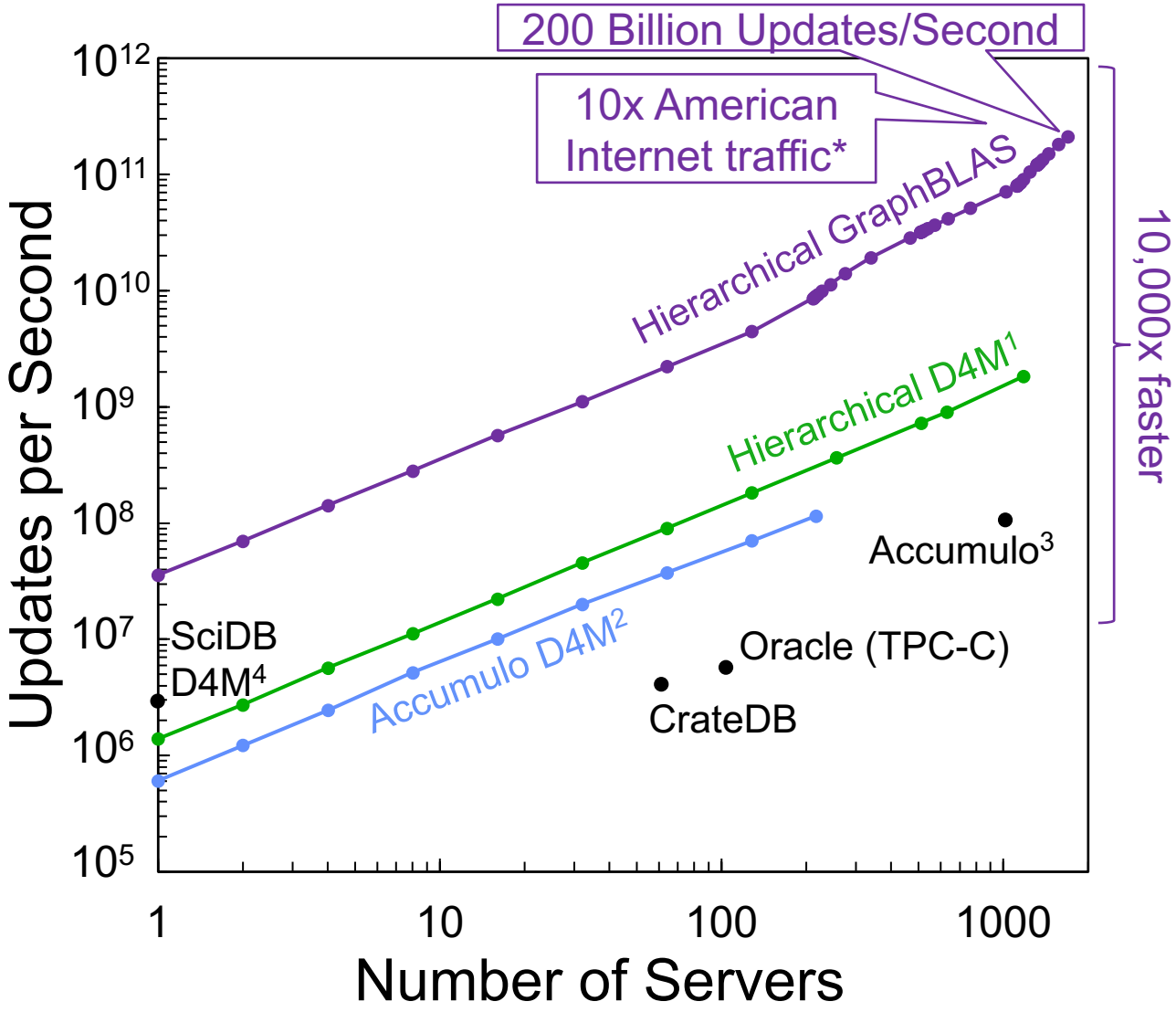
Performance Breakthroughs

Compression



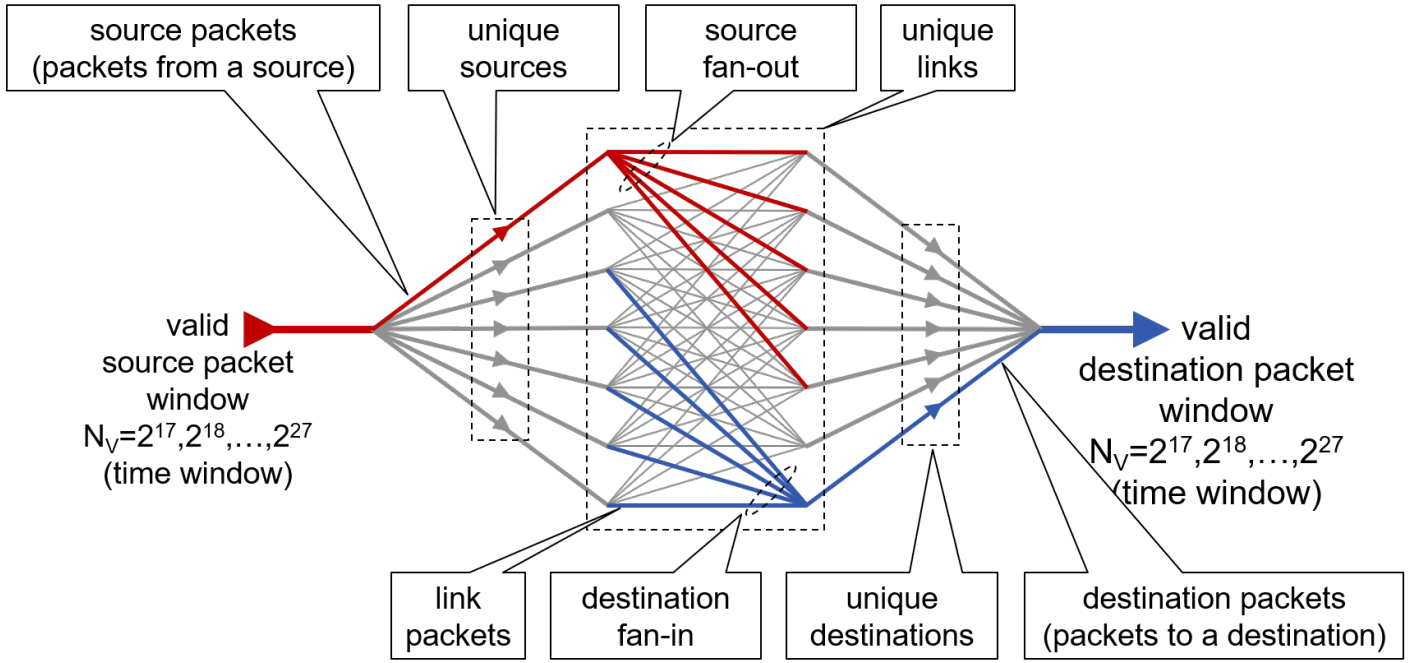
GraphBLAS Traffic Matrices
3,000x compression & 10,000x faster

Speed



*Cisco Visual Networking Index: Forecast and Trends, 2018–2023; [1]*Vertical, Temporal, and Horizontal Scaling of Hierarchical Hypersparse GraphBLAS Matrices*, Kepner et al, IEEE HPEC 2021; [2]*Streaming 1.9 Billion Hypersparse Network Updates per Second with D4M*, Kepner et al, IEEE HPEC 2019; [3]*Achieving 100,000,000 database inserts per second using Accumulo and D4M*, Kepner et al, IEEE HPEC 2014; [4]*Benchmarking Apache Accumulo BigData Distributed Table Store*, Sen et al, IEEE Big Data Congress 2013; [5]*Benchmarking SciDB Data Import on HPC Systems*, Samsi et al, IEEE HPEC 2016

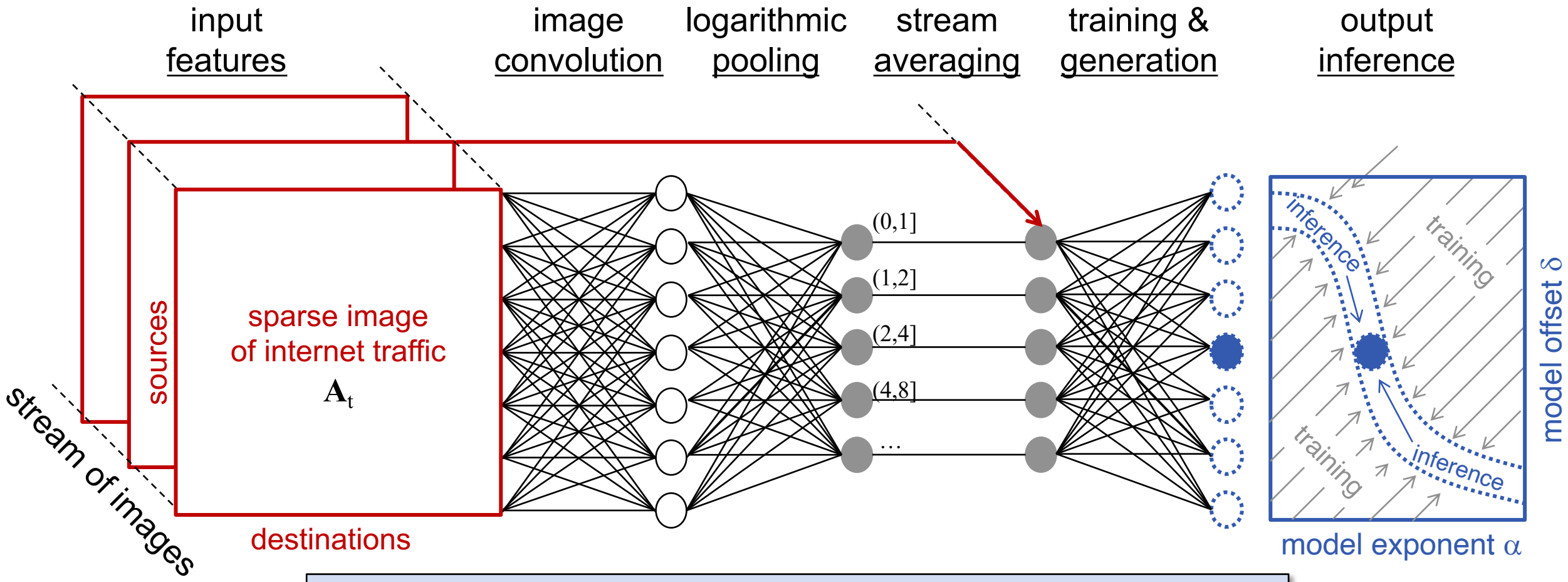
GraphBLAS Hypersparse Math



Aggregate Property	Summation Notation	Matrix Notation
Valid packets N_V	$\sum_i \sum_j \mathbf{A}_t(i,j)$	$\mathbf{1}^\mathsf{T}\mathbf{A}_t\mathbf{1}$
Unique links	$\sum_i \sum_j \lvert\mathbf{A}_t(i,j)\rvert_0$	$\mathbf{1}^\mathsf{T}\lvert\mathbf{A}_t\rvert_0\mathbf{1}$
Link packets from i to j	$\mathbf{A}_t(i,j)$	$\mathbf{A}_t$
Max link packets ($d_{\max}$)	$\max_{ij} \mathbf{A}_t(i,j)$	$\max(\mathbf{A}_t)$
Unique sources	$\sum_i \lvert\sum_j \mathbf{A}_t(i,j)\rvert_0$	$\mathbf{1}^\mathsf{T}\lvert\mathbf{A}_t\mathbf{1}\rvert_0$
Packets from source i	$\sum_j \mathbf{A}_t(i,j)$	$\mathbf{A}_t\mathbf{1}$
Max source packets ($d_{\max}$)	$\max_i \sum_j \mathbf{A}_t(i,j)$	$\max(\mathbf{A}_t\mathbf{1})$
Source fan-out from i	$\sum_j \lvert\mathbf{A}_t(i,j)\rvert_0$	$\lvert\mathbf{A}_t\rvert_0\mathbf{1}$
Max source fan-out ($d_{\max}$)	$\max_i \sum_j \lvert\mathbf{A}_t(i,j)\rvert_0$	$\max(\lvert\mathbf{A}_t\rvert_0\mathbf{1})$
Unique destinations	$\sum_j \lvert\sum_i \mathbf{A}_t(i,j)\rvert_0$	$\lvert\mathbf{1}^\mathsf{T}\mathbf{A}_t\rvert_0\mathbf{1}$
Destination packets to j	$\sum_i \mathbf{A}_t(i,j)$	$\mathbf{1}^\mathsf{T}\lvert\mathbf{A}_t\rvert_0$
Max destination packets ($d_{\max}$)	$\max_j \sum_i \mathbf{A}_t(i,j)$	$\max(\mathbf{1}^\mathsf{T}\lvert\mathbf{A}_t\rvert_0)$
Destination fan-in to j	$\sum_i \lvert\mathbf{A}_t(i,j)\rvert_0$	$\mathbf{1}^\mathsf{T}\mathbf{A}_t$
Max destination fan-in ($d_{\max}$)	$\max_j \sum_i \lvert\mathbf{A}_t(i,j)\rvert_0$	$\max(\mathbf{1}^\mathsf{T}\mathbf{A}_t)$

GraphBLAS hypersparse traffic images enable efficient computation of network quantities in C/C++/Python/Julia/Matlab/Octave

Graph Neural Network Analysis of Large-Scale Internet Traffic



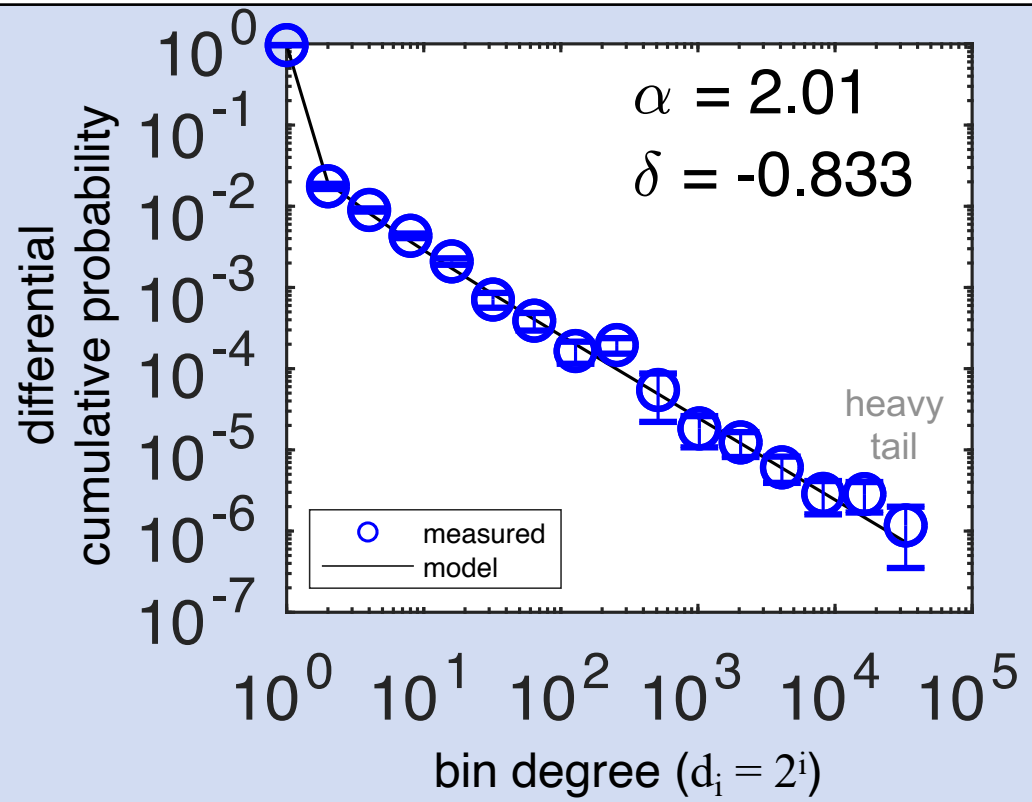
- **Traffic matrices are sparse images of internet traffic**
- **Used to train statistical models of network quantities**

$$p(d) \propto 1/(d+\delta)^{\alpha}$$

heavy tail model

Hypersparse Neural Network Analysis of Large-Scale Internet Traffic, IEEE HPEC 2019

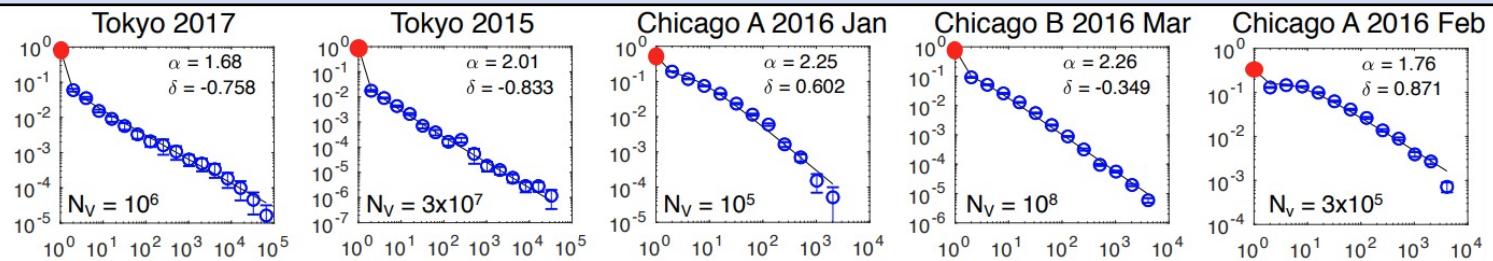
High Precision Heavy Tail Internet Background Model



- **Fit measured binned probability with modified Zipf-Mandelbrot distribution**

$$p(d) \propto 1/(d + \delta)^{\alpha}$$

- **Characterizes each distribution with two parameters**



Hypersparse Neural Network Analysis of Large-Scale Internet Traffic, Kepner et al., HPEC 2019: https://arxiv.org/abs/1904.04396

Holistic Internet Background Model

- **Given a window with N_V incoming packets, the probability of seeing a source emitting d packets being observed again at time t by a different observer is proportional to**

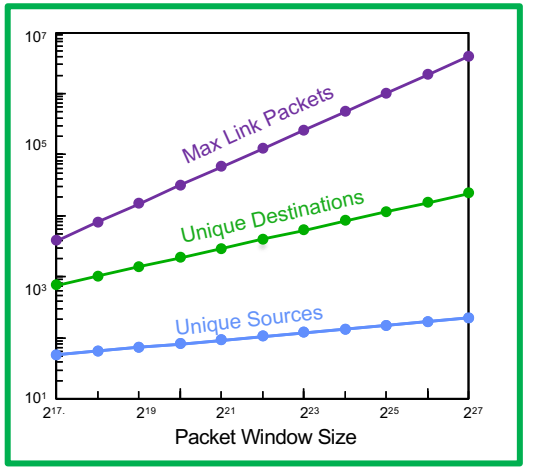
$$N_V^{\,a}$$

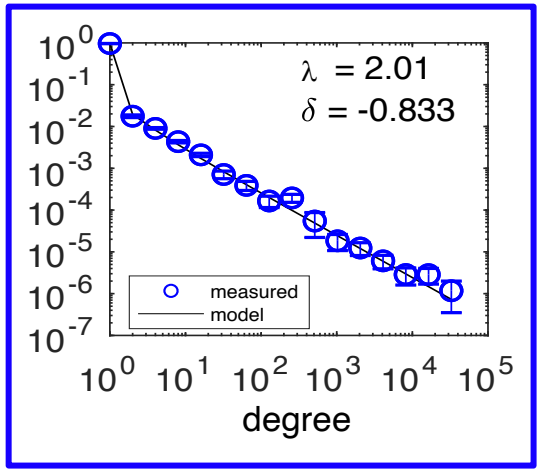
$$1/(d+\delta)^{\lambda}$$

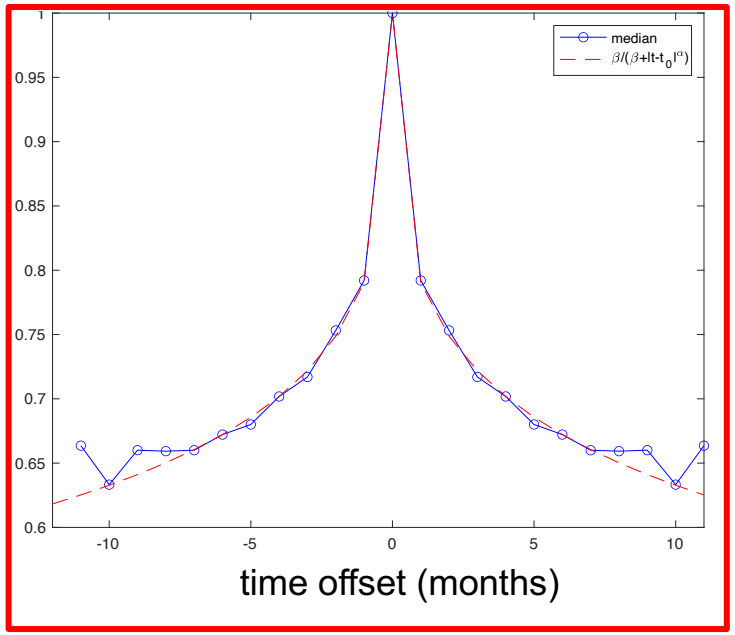
$$\beta/(\beta + t^{\alpha})$$

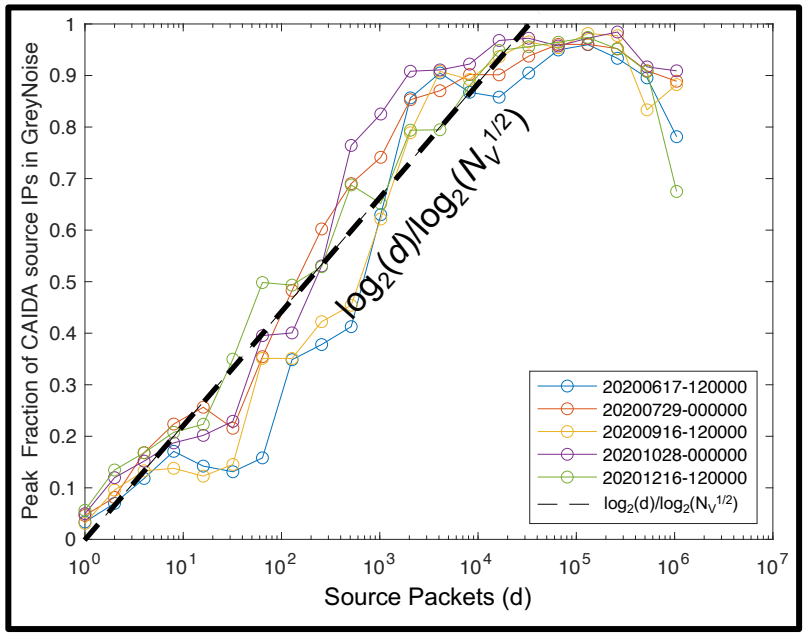
$$\log_2(d)/\log_2(N_V^{1/2})$$

- **where a, λ, δ, α, and β, are site specific model parameters that vary slowly in time**









What is Normal? A Big Data Observational Science Model of Anonymized Internet Traffic, IEEE HPEC 2024

AI Ready Anonymized Network Sensing Challenge

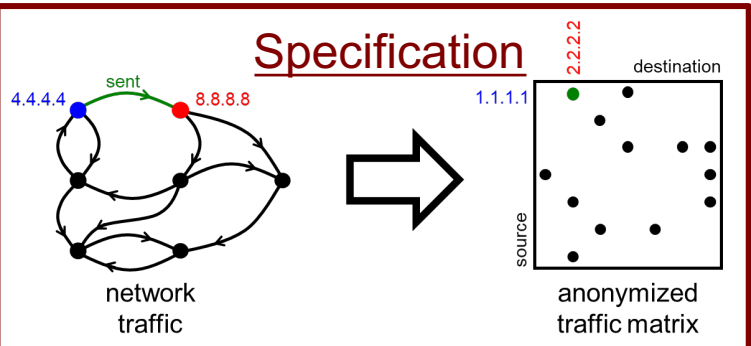
- IEEE High Performance Extreme Computing (HPEC) Conference -

GraphChallenge

Motivation | Challenges | Data Sets | Scenarios | Submit | News | Champions | Contact

Home

Specification



Code

/ src / pcap / pcap2grb.c

msjones0 5 months ago

601 lines (522 loc) · 20.2 KB

Code | Blame

```
1   #define _GNU_SOURCE 1
2
3   #include <GraphBLAS.h>
```

Data

caida RESOURCE CATALOG

Anonymized Network Sensing Graph Challenge Dataset

dataset:2024_ieee_anonymized_nsg_challenge

overview | related | cite | access | README.txt

New! Anonymized Network Sensing Graph Challenge This challenge constructs and analyzes anonymized traffic matrices from network packet capture (PCAP) data to enable open community-based approaches to protecting networks.

- Specification: slides, paper, example serial parse code, example serial analyze code, example data sets

NVIDIA

North Star Labs

DBOS

OneSparse

PNNL

RPI

UMass

Northeastern University

ILLINOIS INSTITUTE OF TECHNOLOGY

NSF

FABRIC