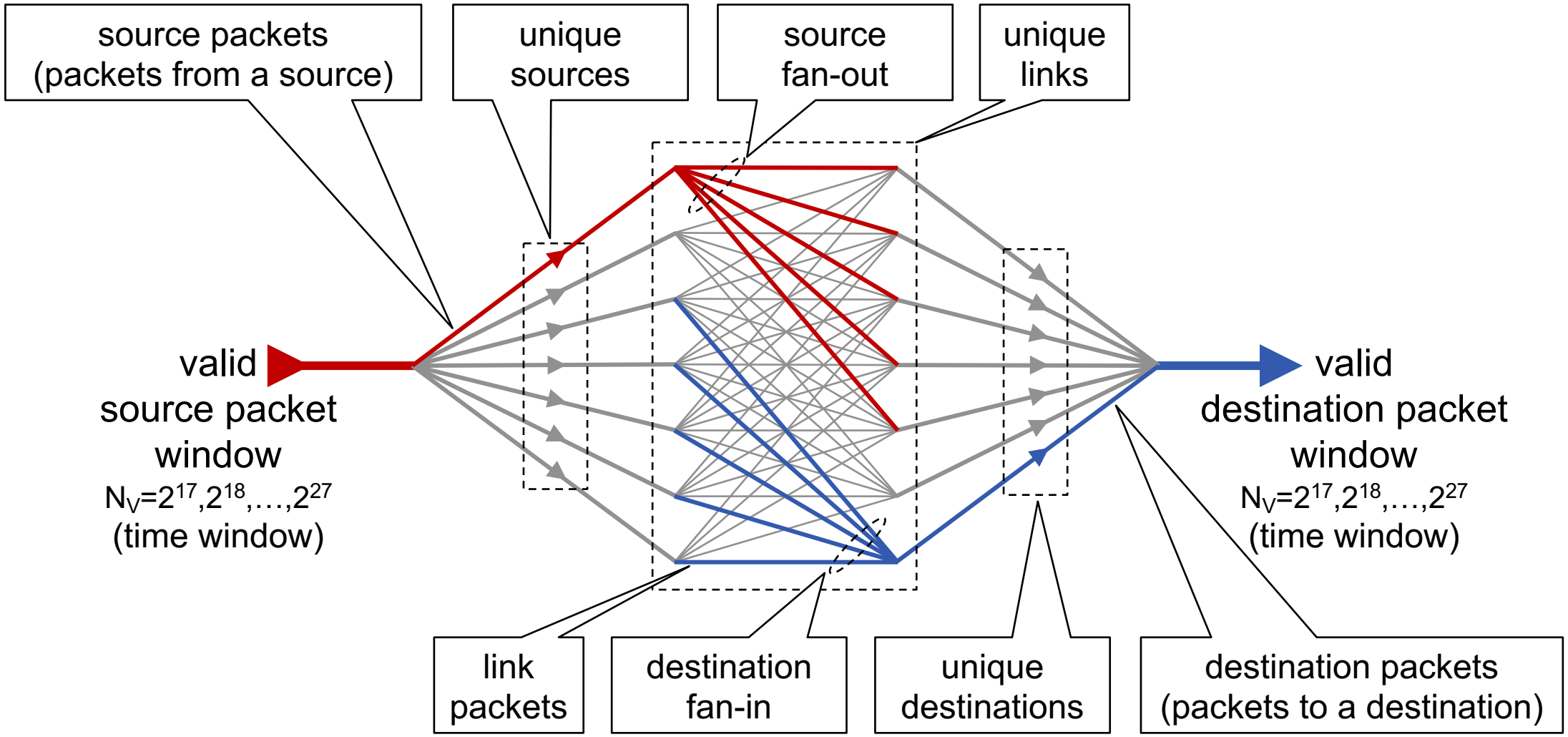
Temporal Cross Correlation of Internet Observatories and Outposts

Jeremy Kepner, Michael Jones, Daniel Andersen, Aydin Buluc, Chansup Byun, K Claffy, Timothy Davis, William Arcand, Jonathan Bernays, David Bestor, William Bergeron, Vijay Gadepally, Daniel Grant, Micheal Houle, Matthew Hubbell, Hayden Jananthan, Anna Klein, Chad Meiners, Lauren Milechin, Andrew Morris, Julie Mullen, Sandeep Pisharody, Andrew Prout, Albert Reuther, Antonio Rosa, Siddharth Samsi, Doug Stetson, Charles Yee, Peter Michaleas

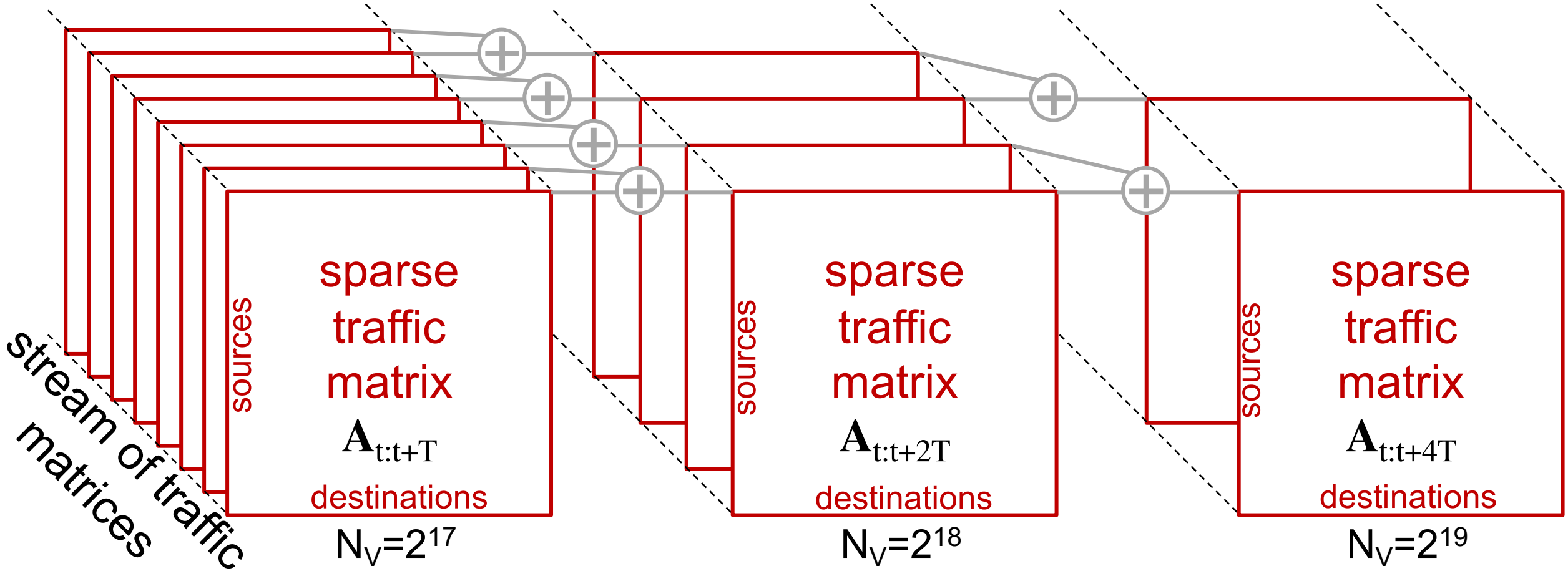
May, 2022

MIT LINCOLN LABORATORY
SUPERCOMPUTING CENTER

Network Flow Definitions

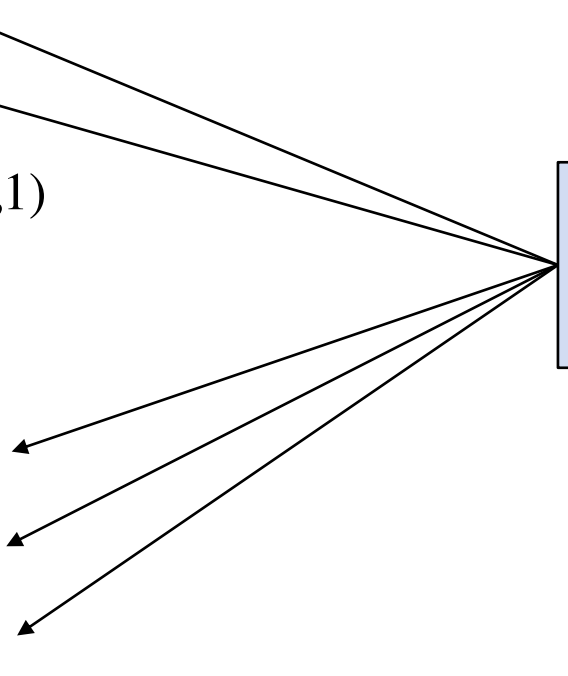


Multi-Temporal Streaming Traffic Matrices



Example: Simple Network Property Formulas

- **Number of valid packets:** $N_{val} = \sum_{ij} \mathbf{A}(i,j) = \mathbb{1}^T \mathbf{A} \mathbb{1}$
- **Source packets:** $\mathbf{A}\mathbb{1}$
- **Destination packets:** $\mathbb{1}^T \mathbf{A}$
- **Unique sources:** $\text{size}(\mathbf{A},1)$
- **Unique destinations:** $\text{size}(\mathbf{A},2)$
- **Number of unique links:** $\text{nnz}(\mathbf{A})$
- **Link packets:** $\mathbf{A}$
- **Source fan-outs:** $|\mathbf{A}|_0 \mathbb{1}$
- **Destination fan-ins:** $\mathbb{1}^T |\mathbf{A}|_0$



Corresponding probability distributions are normalized histograms of these arrays

Cyberspace Security vs Defense vs Deterrence

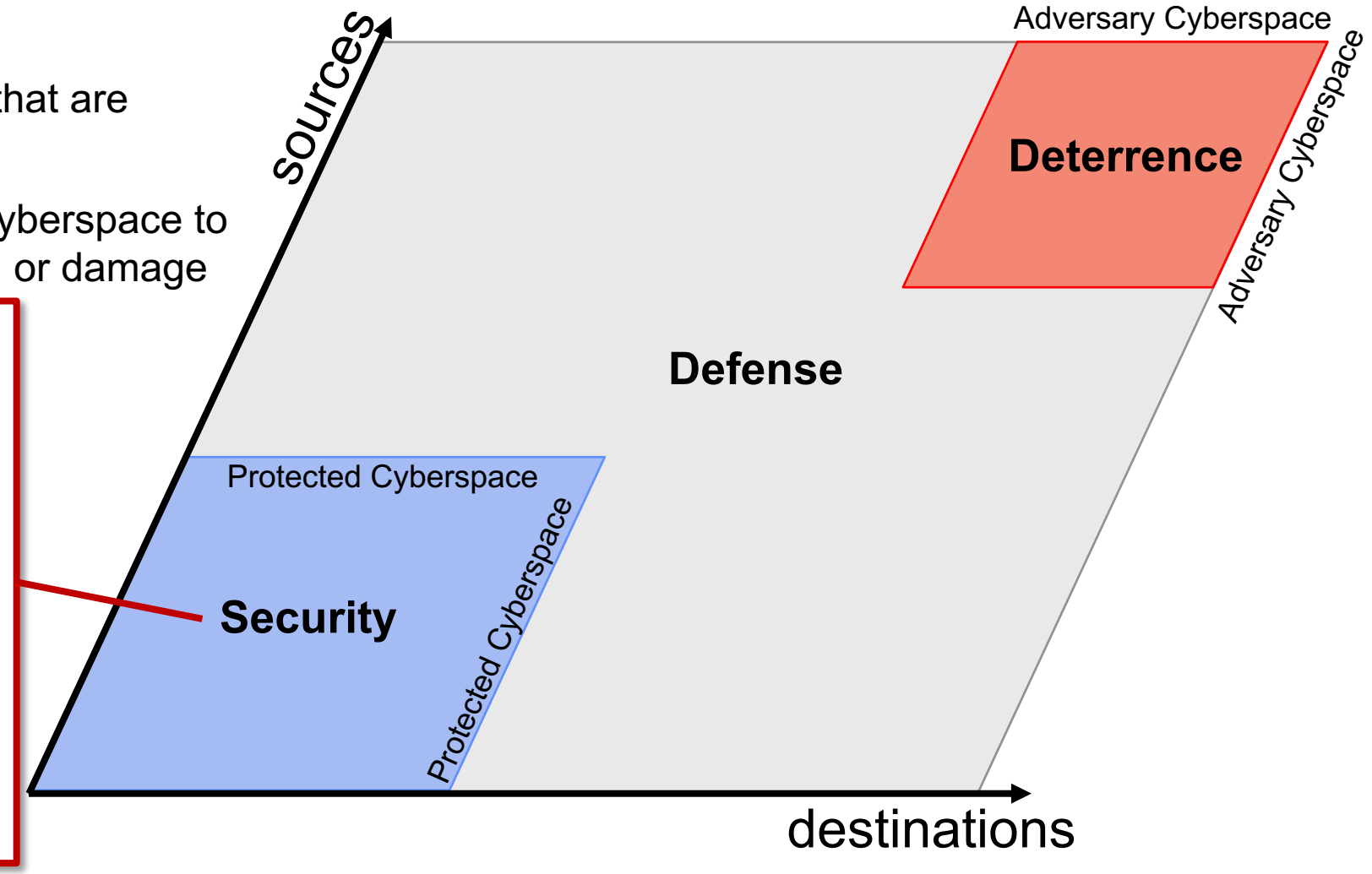
Deterrence - Existence of a credible threat of unacceptable counteraction

Defense - Actions taken to defeat threats that are *threatening* to breach cyberspace security

Security - Actions taken *within* protected cyberspace to prevent unauthorized access, exploitation, or damage

MITRE ATT&CK Matrix

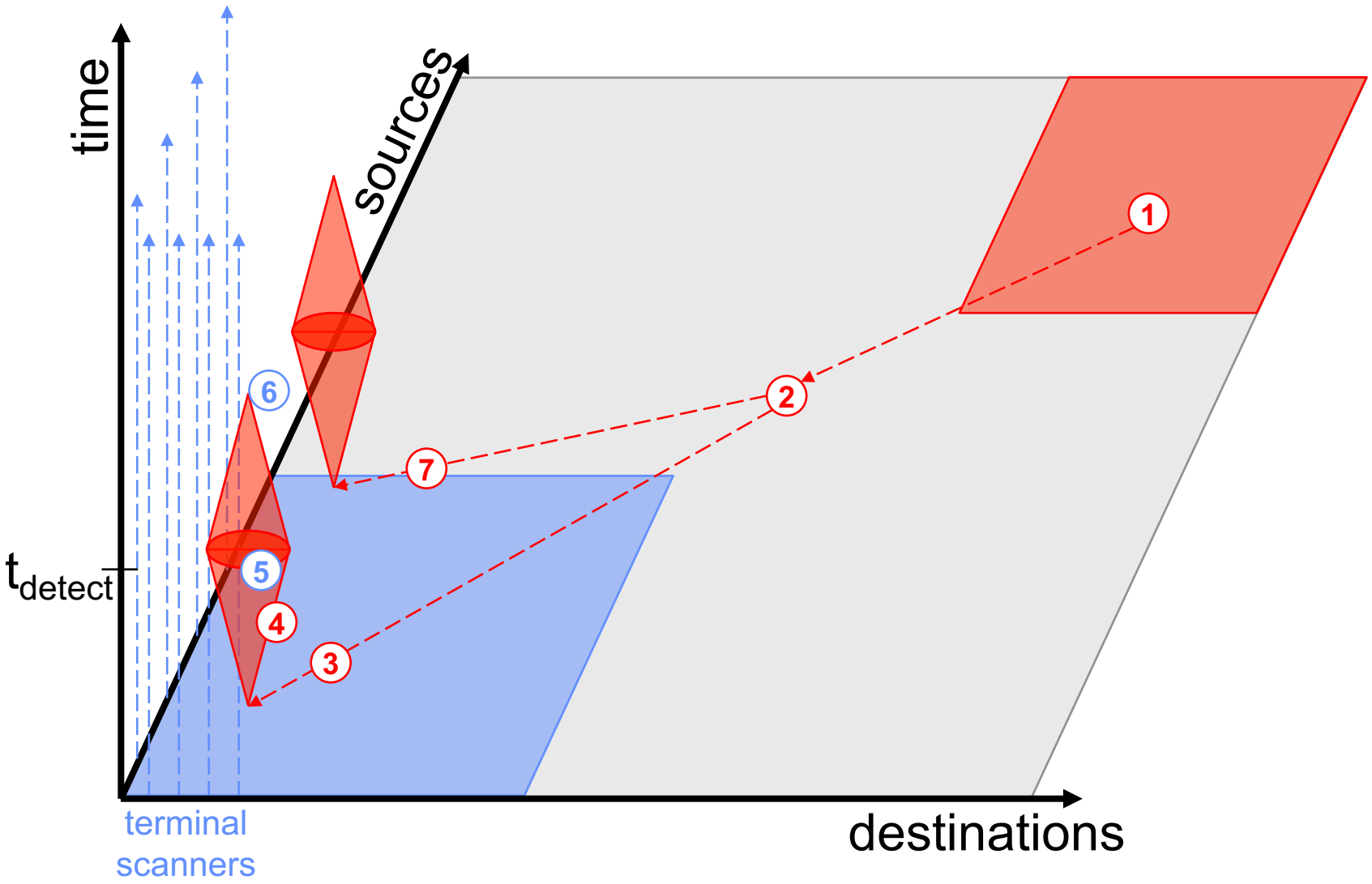
Initial Access … C&C	Exfil Impact
Cybersecurity Response Cyber Incident activating a Cyber Unit who follow Cyber Procedures	Cybersecurity Response Damage Assessment Damage Control Repair



Adapted from DOD Dictionary of Military and Associated Terms, DOD Joint Chiefs of Staff, 2021
ATT&CK = Adversarial Tactics, Techniques, & Common Knowledge

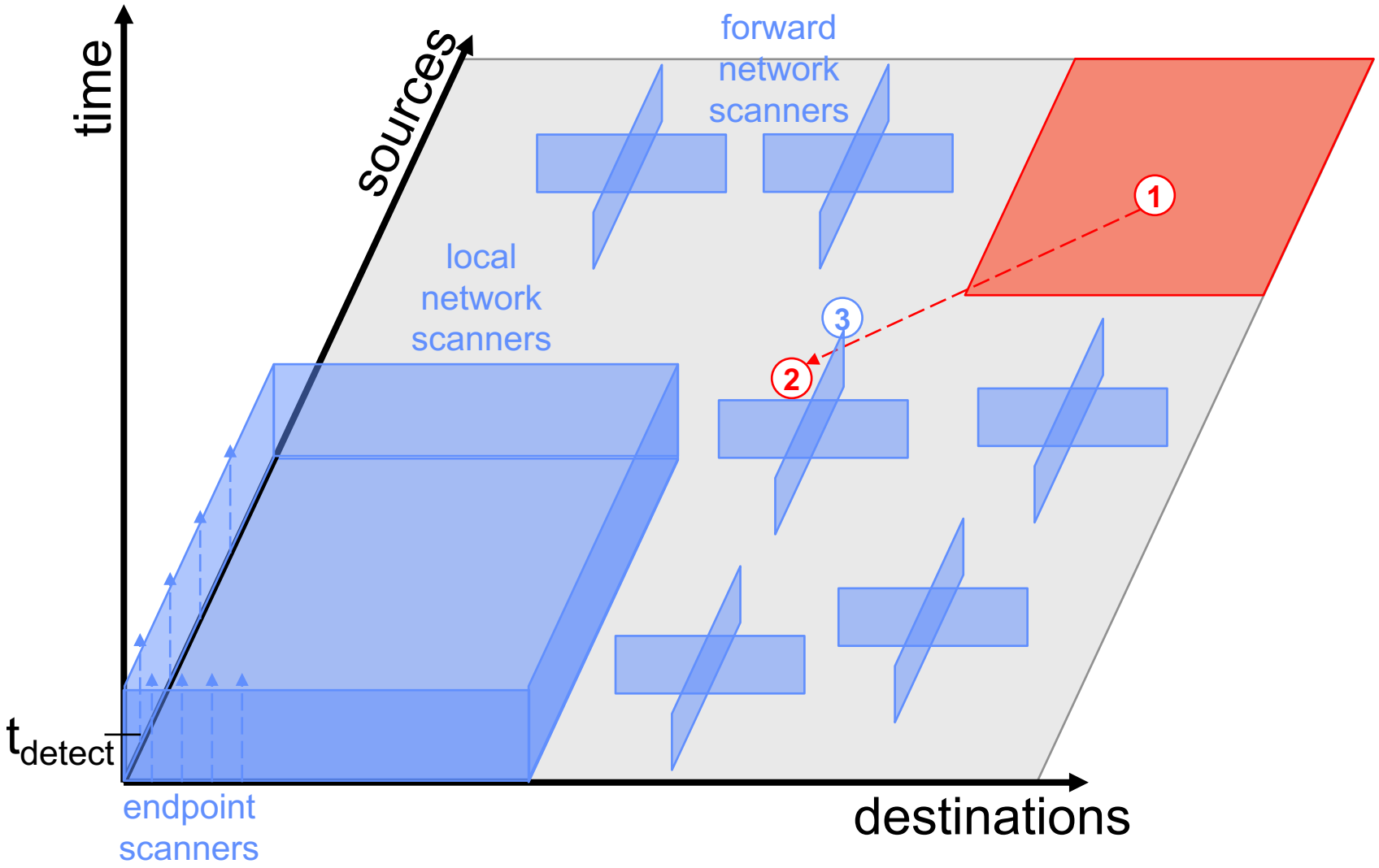
Notional Attack
- Current -

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

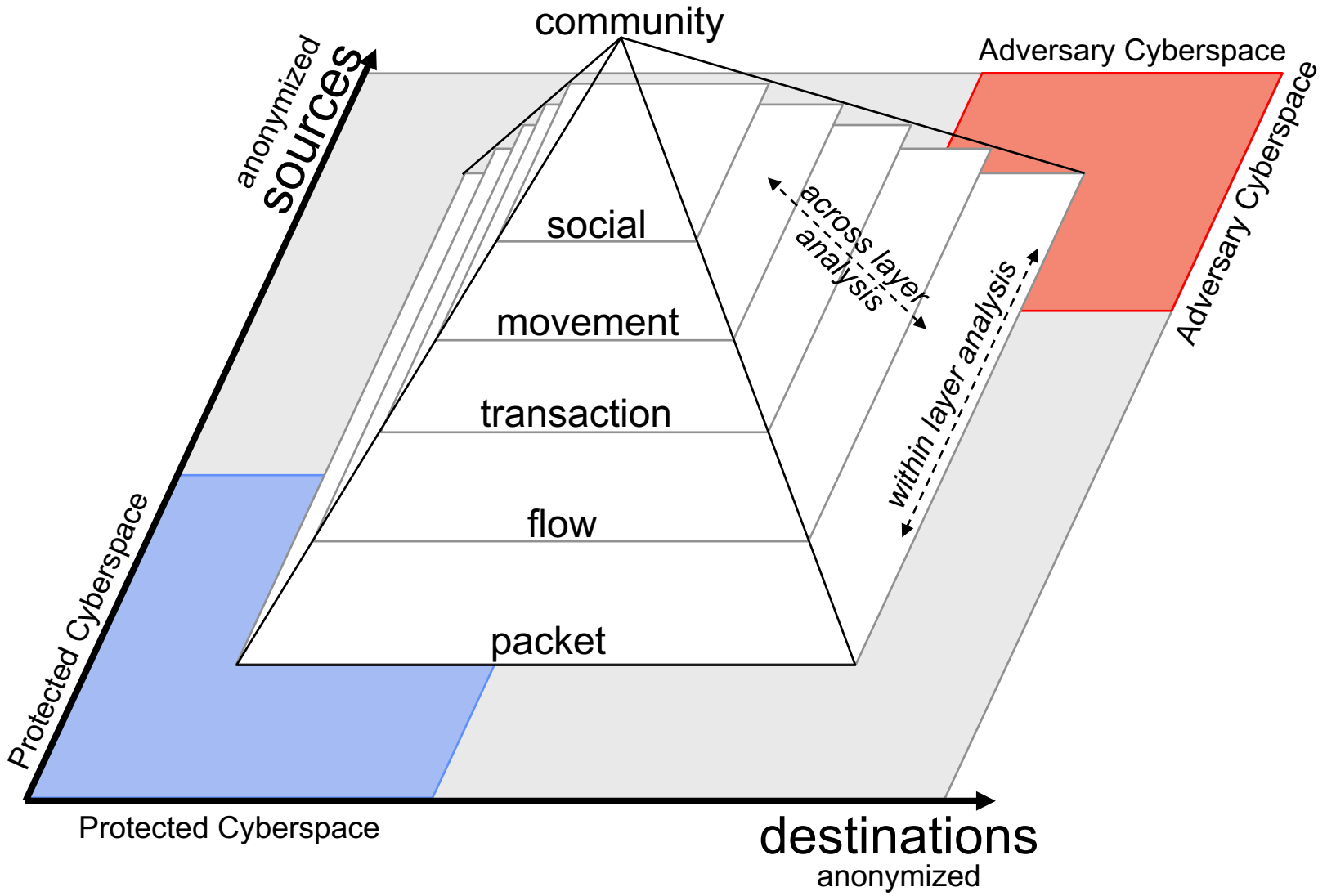


Notional Attack
- Desired -

1. Plan
2. Stage
3. Detect

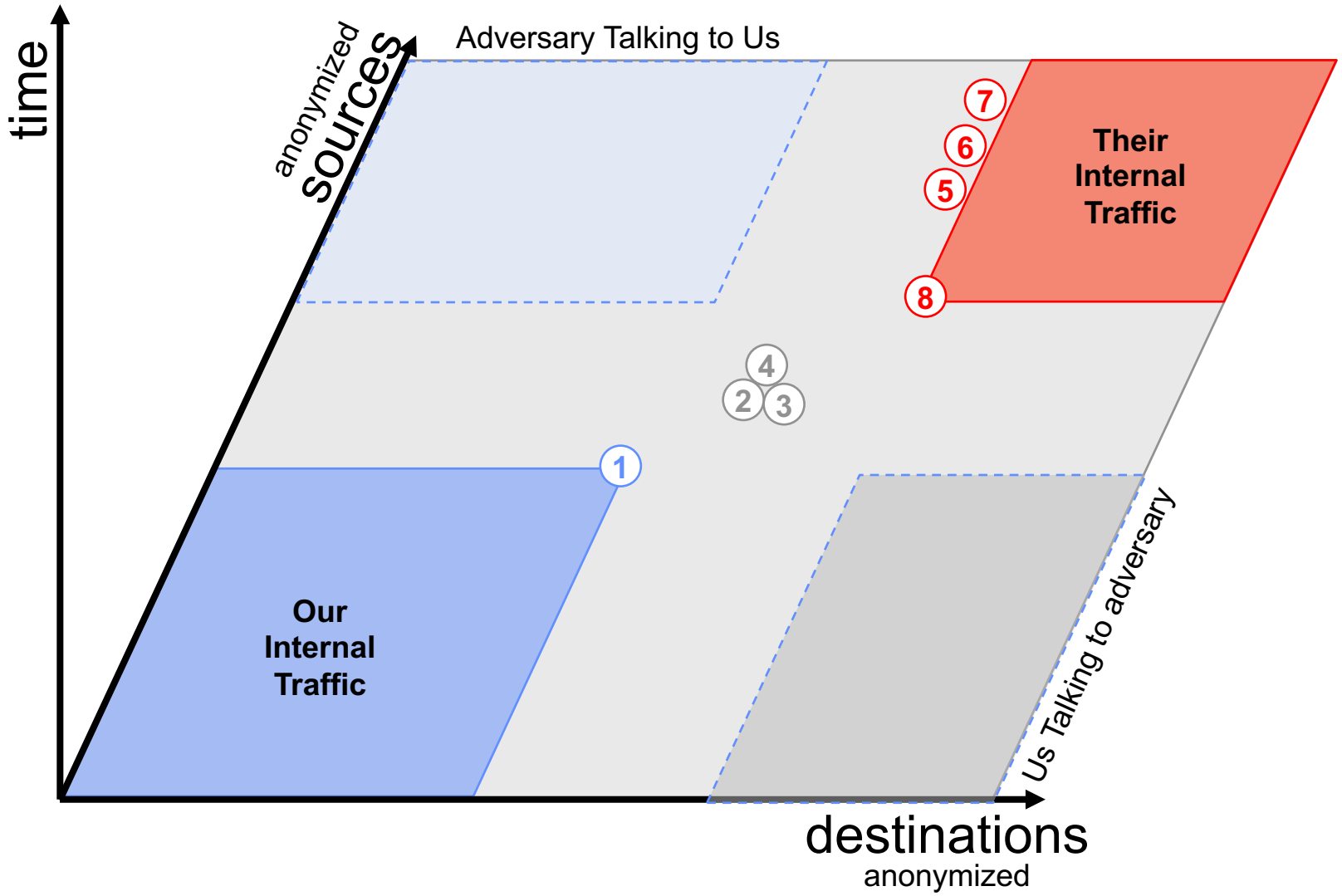


Knowledge Hierarchy Pyramid

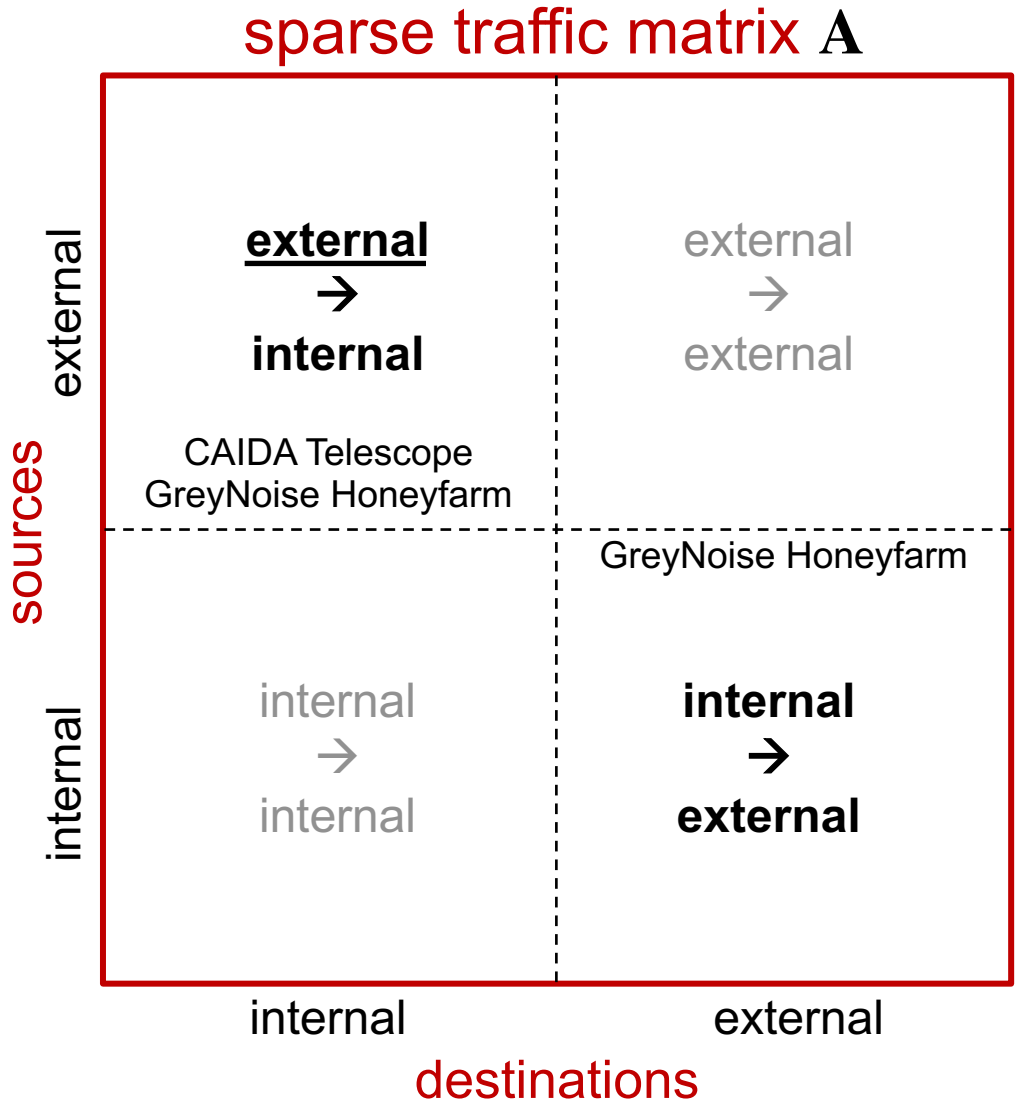


Potential Data Sources: Observatories & Outposts

1. Gov't dark(class B)/blue gateway (~5 years, ~10T packets)
2. MAWI gray trunk (~5 years, ~50B packets)
3. CAIDA gray trunk (~5 years, ~50B packets
4. CAIDA Equinox gray trunk (~100 GigE)
5. CAIDA dark(class A) gateway (5+ years, ~100T packets)
6. Greynoise gateway (~400 active honeypots)
7. Global Cyber Alliance gateway (IoT honeypot farm)
8. Shadowserver gateway (~100M sinkholed botnets)



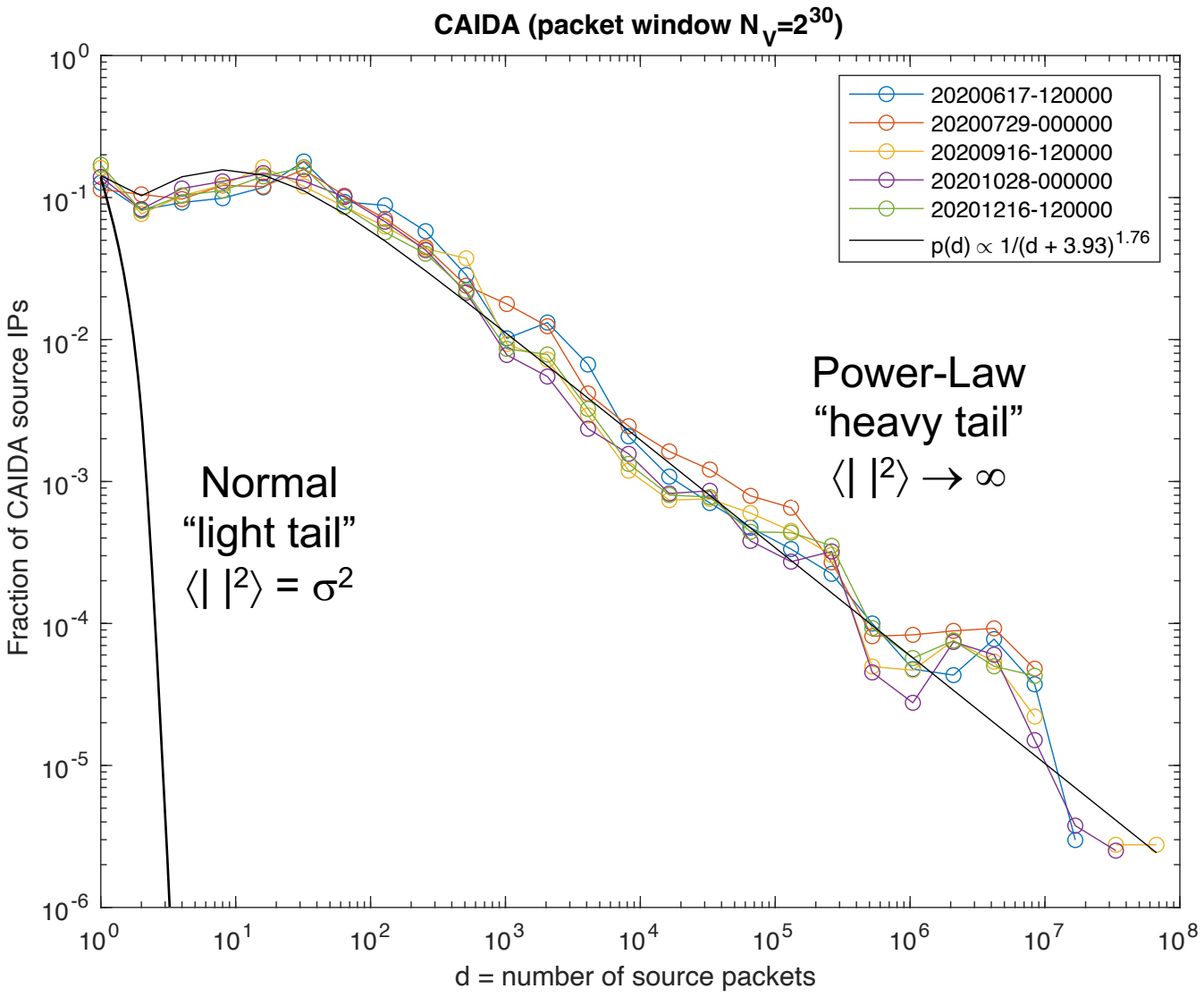
Gateway Internet Traffic Matrices



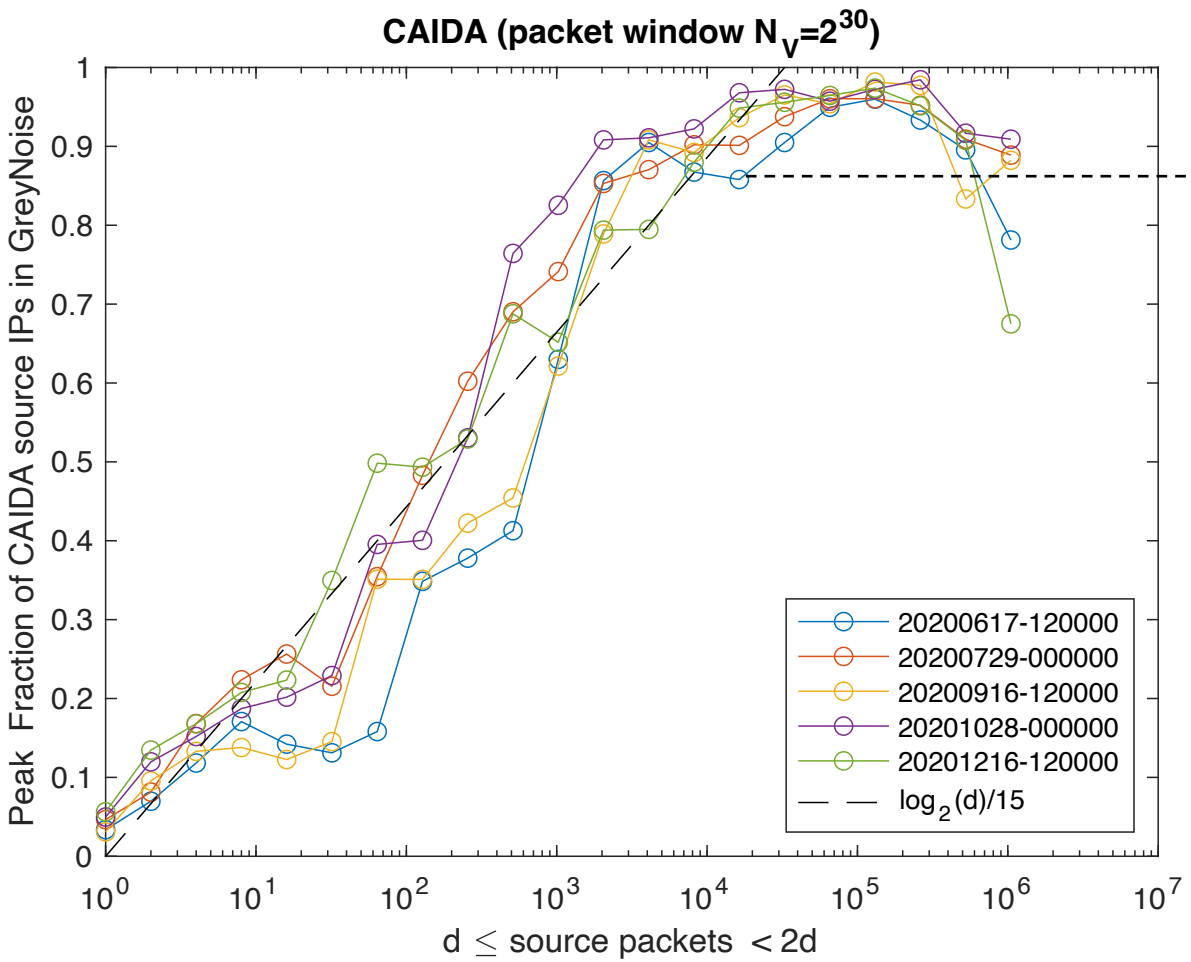
CAIDA & GreyNoise Data

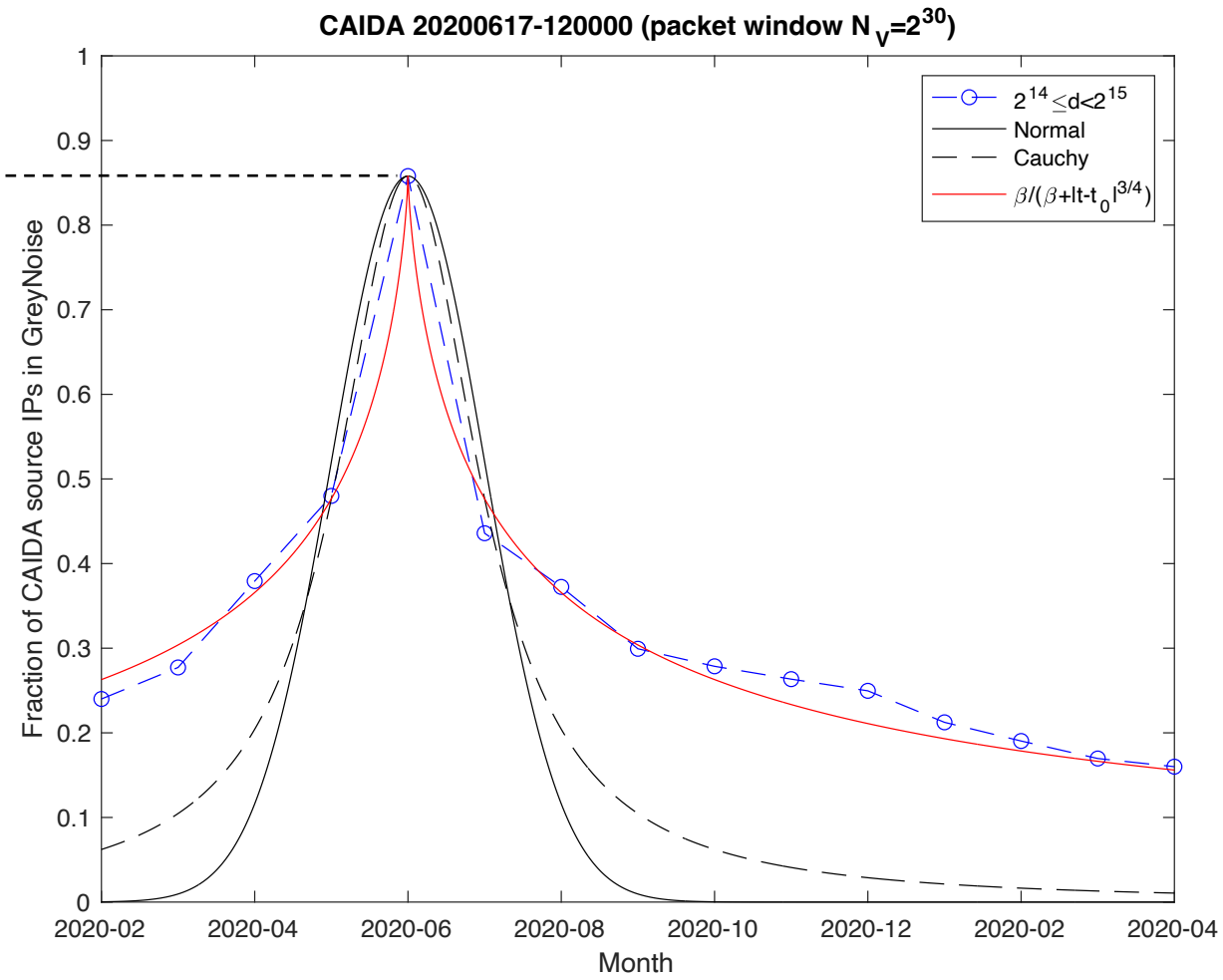
GreyNoise Start Time	GreyNoise Duration	GreyNoise Sources	CAIDA Start Time	CAIDA Duration	CAIDA Packets	CAIDA Sources
2020-02-01	29 days	2,752,690				
2020-03-01	31 days	13,849,634				
2020-04-01	30 days	1,060,905				
2020-05-01	31 days	1,825,351				
2020-06-01	30 days	1,111,458	2020-06-17-12:00:00	1594 sec	2^{30}	670,304
2020-07-01	31 days	1,438,698	2020-07-29-00:00:00	1312 sec	2^{30}	541,300
2020-08-01	31 days	1,367,008				
2020-09-01	30 days	1,245,194	2020-09-16-12:00:00	997 sec	2^{30}	723,991
2020-10-01	31 days	1,997,782	2020-10-28-00:00:00	1068 sec	2^{30}	796,327
2020-11-01	30 days	2,850,037				
2020-12-01	31 days	7,605,790	2020-12-16-12:00:00	1204 sec	2^{30}	701,059
2021-01-01	31 days	2,879,079				
2021-02-01	28 days	2,583,316				
2021-03-01	31 days	3,308,466				
2021-04-01	30 days	11,507,324				

Light vs Heavy Tail Statistics



CAIDA GreyNoise Cross Correlations





Cross correlations well-modeled by a logarithmic spectrum that decays with time

Some Internet Science Results

- Standard data collection sites: endpoints, taps, crawls[1]
  - Each sees different phenomena in the global traffic matrix[5]
- Ubiquitous heavy tail distributions are a challenge for simple statistics
  - Bin by event count (not time)[1,2,5]
- Universal streaming quantities: sources, fan-outs, links, fan-ins, destinations[1,2]
  - Easily computable from anonymized traffic matrices (with the right hardware and software)[5,6,8]
- Scaling relations as a function of bin size abound
  - Parameters stable at a given site; differ site-to-site[3,7]
- Power-law distributions abound; parameters stable at a given site; differ site-to-site
  - High-precision Zipf-Mandelbrot parameters be can found using simple neural networks[1,2]
  - Modeled with preferential attachment with leaf-nodes and isolated links[4]
  - Small deviations from background are indicative of anomalous behavior[5]
- Coeval source correlations are high (low-otherwise) and fit by modified Cauchy distribution[9]
  - Suggests a correlated high frequency “beam” of traffic drifting over time

[1]*New phenomena in large-scale internet traffic*, Kepner et al, 2019; [2]*Hypersparse Neural Network Analysis of Large-Scale Internet Traffic*, Kepner et al, IEEE HPEC 2020; [3]*Multi-temporal analysis and scaling relations of 100,000,000,000 network packets*, Kepner et al, IEEE HPEC 2020; [4]*Hybrid Power-Law Models of Network Traffic*, Devlin et al, GrAPL 2021; [5]*Zero Botnets: An Observe-Pursue-Counter Approach*, Kepner et al, Belfer Center 2021; [6]*Vertical, Temporal, and Horizontal Scaling of Hierarchical Hypersparse GraphBLAS Matrices*, Kepner et al, IEEE HPEC 2021; [7]*Spatial Temporal Analysis of 40,000,000,000,000 Internet Darkspace Packets*, Kepner et al, IEEE HPEC 2021; [8]*Realizing Forward Defense in the Cyber Domain*, Pisharody et al, IEEE HPEC 2021; [9]*Temporal Cross Correlation of Internet Observatories and Outposts,* Kepner et al, GrAPL 2022