Temporal Cross Correlation of Internet Observatories and Outposts


May, 2022
Network Flow Definitions

- Source packets (packets from a source)
- Unique sources
- Source fan-out
- Unique links
- Valid source packet window: $N_v = 2^{17}, 2^{18}, \ldots, 2^{27}$ (time window)
- Link packets
- Destination fan-in
- Unique destinations
- Destination packets (packets to a destination)
- Valid destination packet window: $N_v = 2^{17}, 2^{18}, \ldots, 2^{27}$ (time window)
Multi-Temporal Streaming Traffic Matrices

Stream of traffic matrices

\[ A_{t:t+T} \]

Sparse traffic matrix

\[ N_V = 2^{17} \]

Sources

Destinations

\[ A_{t:t+2T} \]

Sparse traffic matrix

\[ N_V = 2^{18} \]

Sources

Destinations

\[ A_{t:t+4T} \]

Sparse traffic matrix

\[ N_V = 2^{19} \]

Sources

Destinations
Example: Simple Network Property Formulas

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

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**

<table>
<thead>
<tr>
<th>Mitre ATT&amp;CK Matrix</th>
<th>Initial Access … C&amp;C</th>
<th>Exfil Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cybersecurity Response</td>
<td>Cyber Incident activating a Cyber Unit who follow Cyber Procedures</td>
<td></td>
</tr>
<tr>
<td>Cybersecurity Response</td>
<td>Damage Assessment</td>
<td></td>
</tr>
<tr>
<td>Damage Control</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repair</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from DOD Dictionary of Military and Associated Terms, DOD Joint Chiefs of Staff, 2021

ATT&CK = Adversarial Tactics, Techniques, & Common Knowledge
1. Plan
2. Stage
3. Infiltrate
4. Move laterally
5. Detect
6. Cleanse
7. Infiltrate
1. Plan  
2. Stage  
3. Detect
Knowledge Hierarchy Pyramid

- Community
- Social
- Movement
- Transaction
- Flow
- Packet

Protected Cyberspace

Adversary Cyberspace

anonymized sources

Protected Cyberspace

within layer analysis

across layer analysis

anonymized destinations
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

sparse traffic matrix $A$

$\begin{array}{c|c}
\text{external} & \text{external} \\
\rightarrow & \rightarrow \\
\text{internal} & \text{external} \\
\end{array}$

CAIDA Telescope
GreyNoise
Honeyfarm

$\begin{array}{c|c}
\text{internal} & \text{internal} \\
\rightarrow & \rightarrow \\
\text{internal} & \text{external} \\
\end{array}$

GreyNoise
Honeyfarm
## CAIDA & GreyNoise Data

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

![Graph showing light vs heavy tail statistics.](image)

- **Normal “light tail”**
  \[ \langle | \rangle^2 = \sigma^2 \]

- **Power-Law “heavy tail”**
  \[ \langle | \rangle^2 \to \infty \]

\[ p(d) \propto \frac{1}{(d + 3.93)^{1.76}} \]
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