Protographs: Graph-Based Approach to NetFlow Analysis

Jeff Janies
RedJack
FloCon 2011
Thesis

• Using social networks we can complement our existing volumetric analysis.
  – Identify phenomenon we are missing because they are just not “bandwidth heavy” enough.
  – Relate behaviors in novel ways.
  – What is really the most important host in a collection a network?
Social Network Analysis

- Demonstrates relationships through Graphs
  - Allows us to map out interconnections.
- Objective measure of social importance
  - Who connects the groups together?
  - Who can influence communication?
Protocol Graphs

- Protocol Graphs – Social networks of host communications. *(Who talked to whom)*
  - Undirected Graphs
  - **Vertices** – The hosts that communicated.
  - **Edges** – Connects between hosts that communicated.

- Analyze a specific phenomenon.
  - Ex: BotNet, P2P, Established services
Protograph Tool

• Processes raw SiLK NetFlow data.

• Produces protocol graphs.
  – Only uses IP information.

• Reports **centrality** of hosts.
  – **Centrality** – How integral a host is to the group.
### Example NetFlow

<table>
<thead>
<tr>
<th>SIP</th>
<th>DIP</th>
<th>Sport</th>
<th>Dport</th>
<th>Flags</th>
<th>Bytes</th>
<th>Pkts</th>
<th>Stime</th>
</tr>
</thead>
<tbody>
<tr>
<td>192.168.1.100</td>
<td>192.168.1.1</td>
<td>21234</td>
<td>80</td>
<td>SAF</td>
<td>220</td>
<td>4</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>192.168.1.1</td>
<td>192.168.1.100</td>
<td>80</td>
<td>21234</td>
<td>SAF</td>
<td>60035</td>
<td>5</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>10.0.1.35</td>
<td>192.168.1.15</td>
<td>32143</td>
<td>8080</td>
<td>SAR</td>
<td>180</td>
<td>4</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>192.168.1.15</td>
<td>10.0.1.35</td>
<td>8080</td>
<td>32143</td>
<td>SAR</td>
<td>502</td>
<td>5</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>10.0.1.35</td>
<td>192.168.1.100</td>
<td>32144</td>
<td>8080</td>
<td>SAR</td>
<td>180</td>
<td>4</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>192.168.1.100</td>
<td>10.0.1.35</td>
<td>8080</td>
<td>32144</td>
<td>SAR</td>
<td>502</td>
<td>5</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>10.0.1.35</td>
<td>192.168.1.115</td>
<td>32145</td>
<td>8080</td>
<td>SAR</td>
<td>180</td>
<td>4</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>192.168.1.115</td>
<td>10.0.1.35</td>
<td>8080</td>
<td>32145</td>
<td>SAR</td>
<td>502</td>
<td>5</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>10.0.1.35</td>
<td>192.168.1.200</td>
<td>32146</td>
<td>8080</td>
<td>SAR</td>
<td>180</td>
<td>4</td>
<td>2010/01/01T..</td>
</tr>
<tr>
<td>192.168.1.200</td>
<td>10.0.1.35</td>
<td>8080</td>
<td>32146</td>
<td>SAR</td>
<td>502</td>
<td>5</td>
<td>2010/01/01T..</td>
</tr>
</tbody>
</table>
NetFlow as a Protocol Graph

- That NetFlow Makes this graph.
  - No Volume.
  - No Direction.
  - Just Connections.

- Centrality
  - 10.0.1.35
    - Connects many.
  - 192.168.1.100
    - Connects 192.168.1.1 to the rest of the graph.
  - If either removed, the graph is no longer fully connected.
Centrality

- A measure of social importance.
- **Betweenness** – How efficiently a vertex connects the graph. (protograph)
- **Degree** – How many vertices are connected to the vertex. (SiLK’ rwuniq)
- **Closeness** – How close a vertex is to other vertices.
- **Eigenvector** – How “important” a vertex is.
Betweenness

- Which hosts provide the most shortest paths through the network?

\[ \sum_i \sum_j \frac{g_{ikj}}{g_{ij}}, \quad i \neq j \neq k \]

- \( g_{ij} \) — Geodesic paths through host \( i \) and \( j \).
- \( G_{ikj} \) — Geodesic paths through host \( k \) for \( i \) and \( j \).
Interpretation

• The higher the centrality value the more "important" a host is to the graph.
  – Without a central node the graph will break down into unconnected groups. (*The protocol is effected*)
  – Example:
    • If we have all a sample of P2P traffic, centrality tells us which host to remove to cause the most damage to the overlay’s QoS.
  – **Not** necessarily which host is the most talkative.
Volume & Betweenness

• Spikes in centrality may exist without spikes in bandwidth.
  – Centrality measures something not tied to volume.

• Sample data:
  – One week long sample of TCP/IP traffic.
  – Ephemeral port to ephemeral port.
  – >1K bytes, >4 packets.
  – Divided into intervals of 60, 30, and 15 minutes.
Volume measures

![Graph showing volume measures over time](image-url)
Betweenness Centrality

Centrality Score Per 60 Minutes

Centrality Score Per 30 Minutes

Centrality Score Per 15 Minutes

Time
Betweenness Centrality

Spike 1

Centrality Score Per 60 Minutes

Spike 2

Centrality Score Per 30 Minutes

Centrality Score Per 15 Minutes
Volume measures

Spike 1

Spike 2

![Graphs showing volume measures with spikes labeled as Spike 1 and Spike 2.](image)
Spike 1

• 3 hosts have 4x the centrality measure of any host measured at any other time.
  – all three part of same phenomenon.
  – One host was a scan victim of two unrelated hosts.
    • The only overlap in scan victims was this host.

• One scanned ~37,000 destinations on port 20,000. \textit{(usermin exploit)}

• One SA scanned ~3,500 destinations. (various ports)
Spike 2

• 1 host has 3x the centrality of any other host measured at any other time.
  – Contacts 20,000 hosts that connect a graph of 31,000 hosts.

• Active for 6 minutes and sent out 17 million packets.

• Scanner.
Second Data Sample

• Increased resolution to one minute intervals.

• One Week of TCP/IP ephemeral port to ephemeral port traffic:
  – >120 bytes per direction.
  – >3 packets.
  – Contains at least a SYN and ACK flag in the OR of observed Flags.
Betweenness and Degree

• Comparing centralities gives richer understanding of hosts’ relationships.

• Examine hosts that have high Betweenness with modest Degree.
  – Hosts that are important without being directly connected to many other hosts.
Volume Vs. Centralities

Volume in Log Scale

Bytes

Packets

Flows

Degree

Betweenness

Time (Day)
Only Betweenness Spikes

• Recorded each IP address’ max Degree and Betweenness values.
• Divided spikes, or exceedingly high Betweenness centralities into strata.
  – High (>10,000) - All IP addresses also had comparatively high Degree centrality.
  – Low (>1,000 and <10,000) - We investigated 11 IP addresses that had spikes in Betweenness without comparatively high Degree.
High Betweenness
Low Degree

- 9 victims of vulnerability scans.
  - Vulnerability scans requiring full connections.
  - Scanner connects them to a lot of hosts.
- 1 contacted a host that contacted everything.
  - It provides a service for a promiscuous host.
- 1 connected several of the hosts with high Degree and Betweenness centrality.
  - Connecting segments of a P2P network.
    - Easily identified high value asset to the P2P network.
Summary

• Social network analysis:
  – Identifying components of a behavior.
  – Complementary tool to volumetric measures.
    • It does not consider direction or volume.

• Still a great deal of tuning required to make this into an actionable utility.
References