REDJACK

Protographs: Graph-Based Approach to NetFlow Analysis

Jeff Janies

RedJack

FloCon 2011

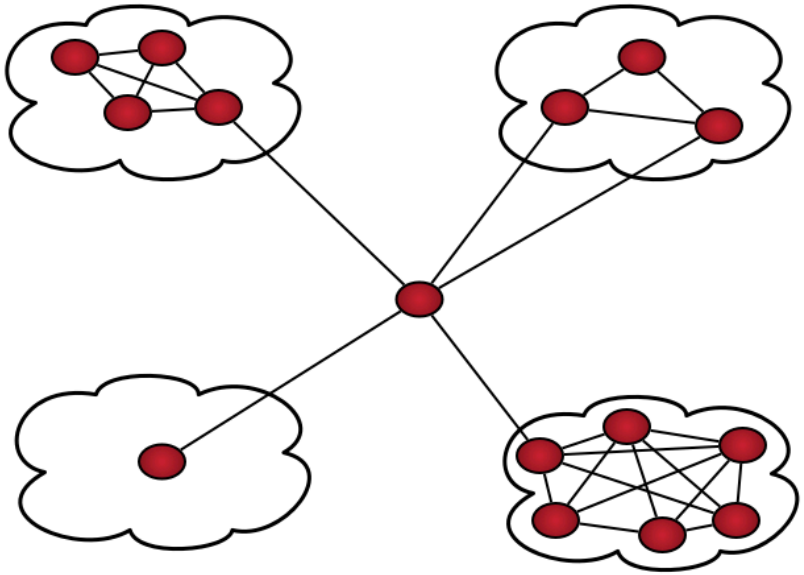
Thesis

REDJACK

- 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

REDJACK



- 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

REDJACK

- 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

REDJACK

- 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

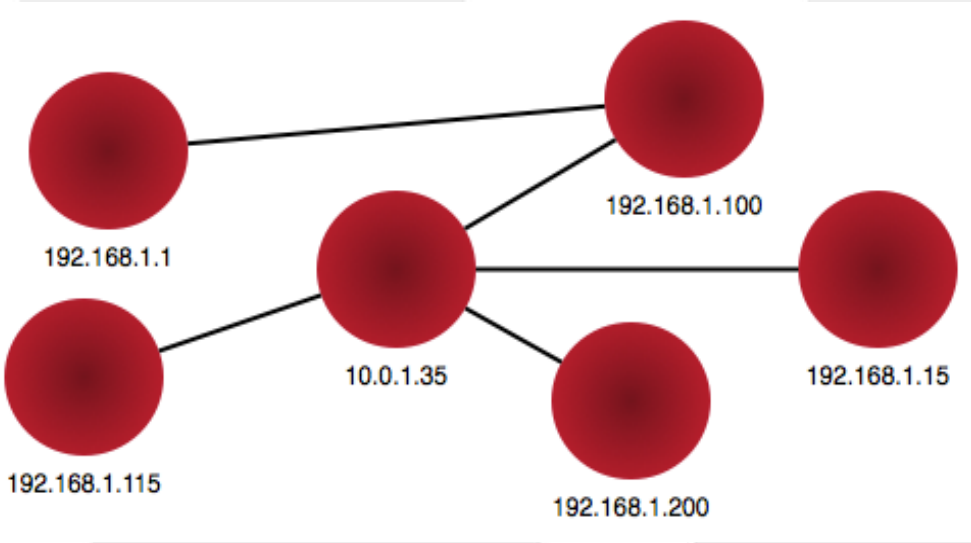
REDJACK

SIP	DIP	Sport	Dport	Flags	Bytes	Pkts	Stime
192.168.1.100	192.168.1.1	21234	80	SAF	220	4	2010/01/01T..
192.168.1.1	192.168.1.100	80	21234	SAF	60035	5	2010/01/01T..
10.0.1.35	192.168.1.15	32143	8080	SAR	180	4	2010/01/01T..
192.168.1.15	10.0.1.35	8080	32143	SAR	502	5	2010/01/01T..
10.0.1.35	192.168.1.100	32144	8080	SAR	180	4	2010/01/01T..
192.168.1.100	10.0.1.35	8080	32144	SAR	502	5	2010/01/01T..
10.0.1.35	192.168.1.115	32145	8080	SAR	180	4	2010/01/01T..
192.168.1.115	10.0.1.35	8080	32145	SAR	502	5	2010/01/01T..
10.0.1.35	192.168.1.200	32146	8080	SAR	180	4	2010/01/01T..
192.168.1.200	10.0.1.35	8080	32146	SAR	502	5	2010/01/01T..

NetFlow as a Protocol Graph

REDJACK

- 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

REDJACK

- 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

REDJACK

- 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

REDJACK

- 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.

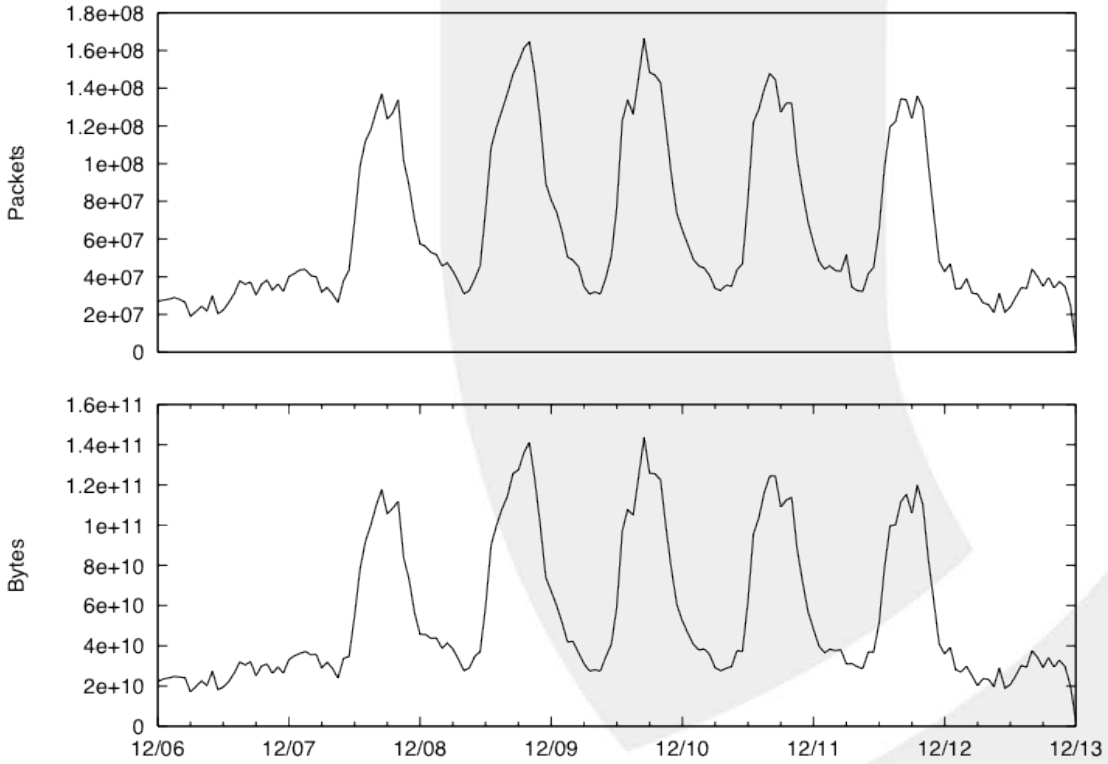
Volume & Betweenness

REDJACK

- 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.

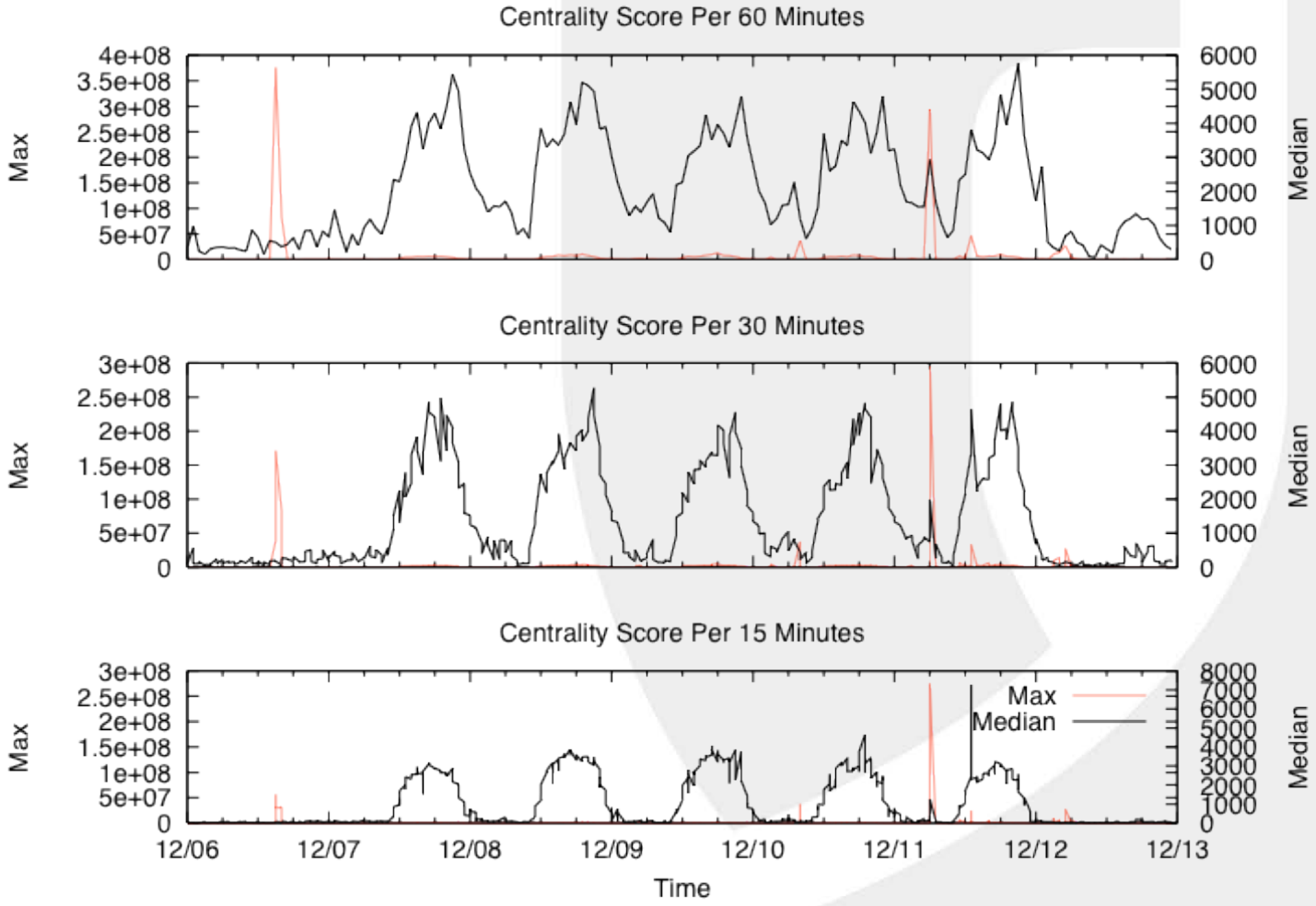
Volume measures

REDJACK

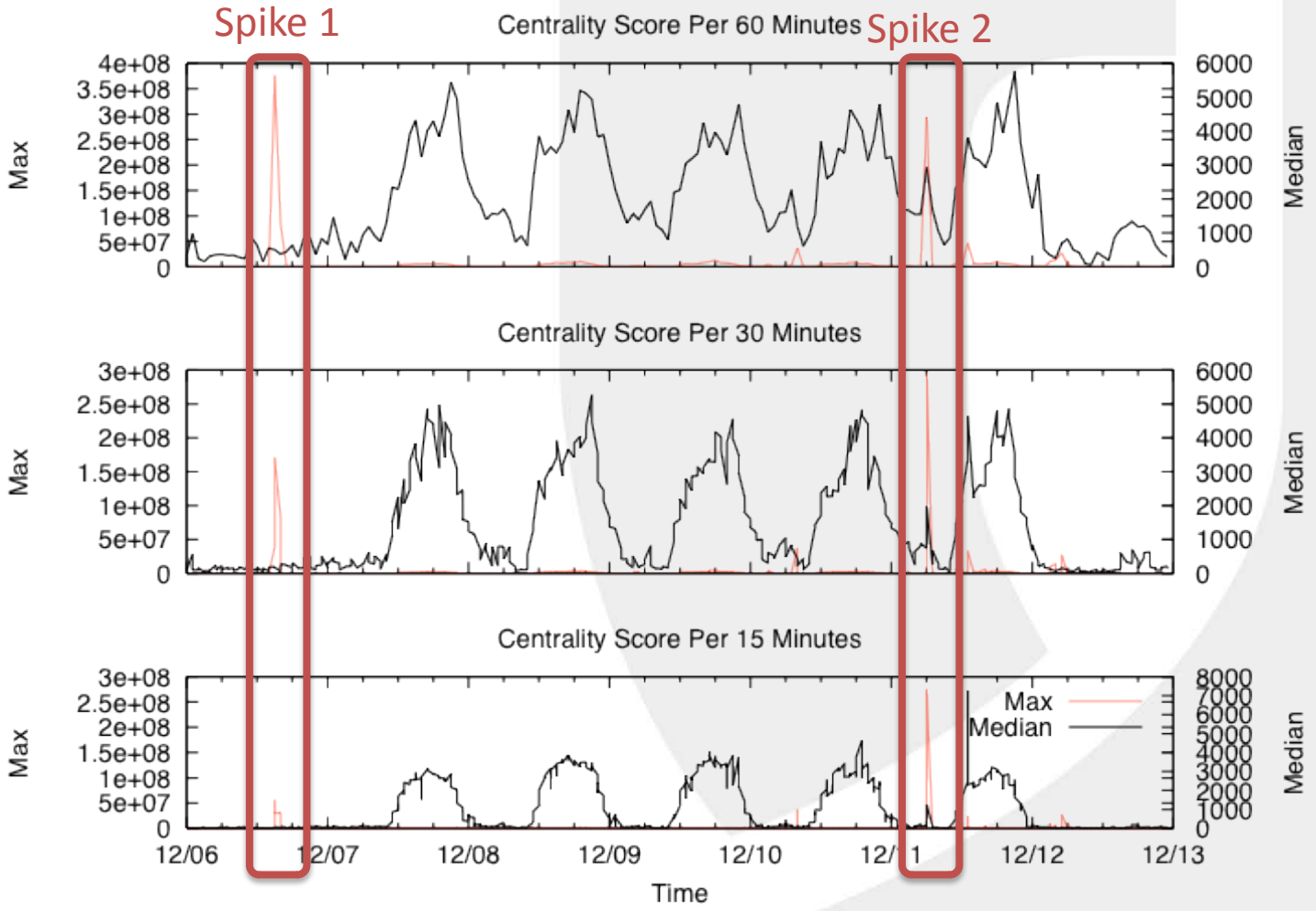


Betweenness Centrality

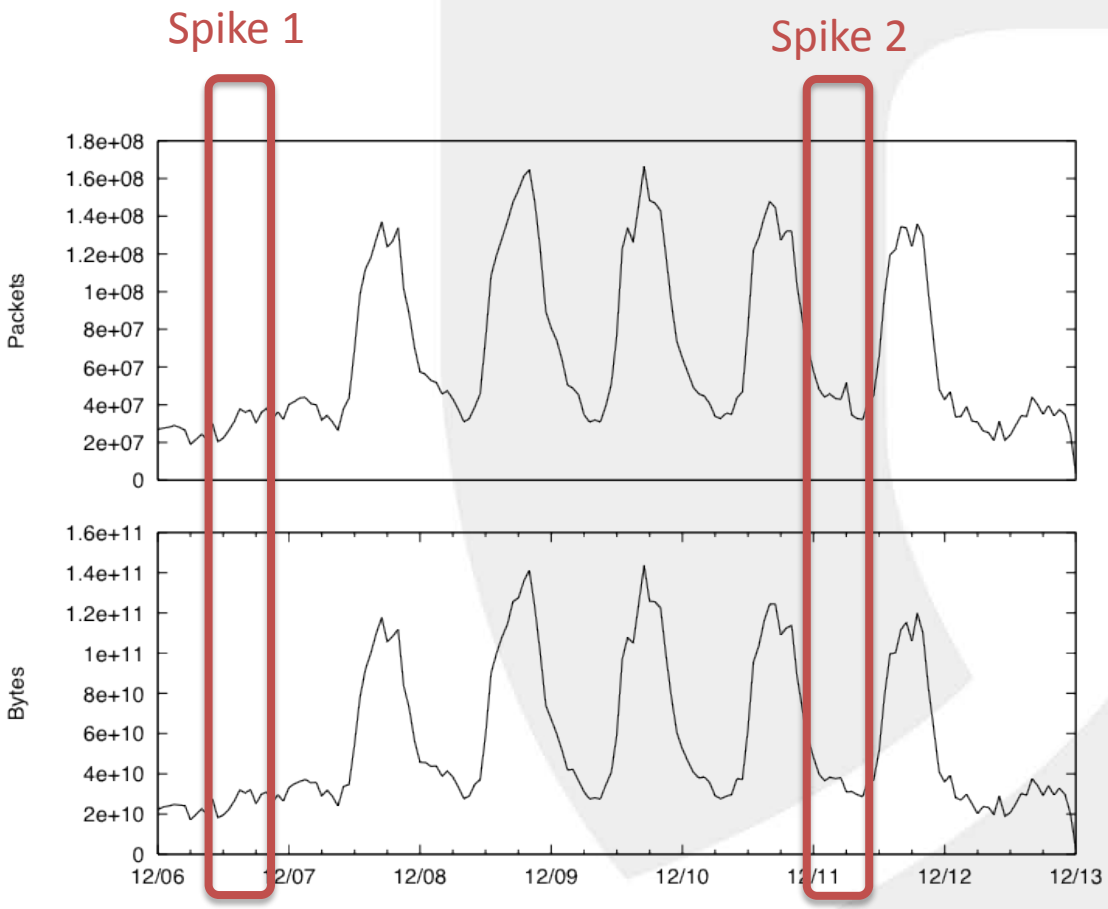
REDJACK



Betweenness Centrality



Volume measures



Spike 1

REDJACK

- 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. *(usermin exploit)*
- One SA scanned ~3,500 destinations. (various ports)

Spike 2

REDJACK

- 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

REDJACK

- 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

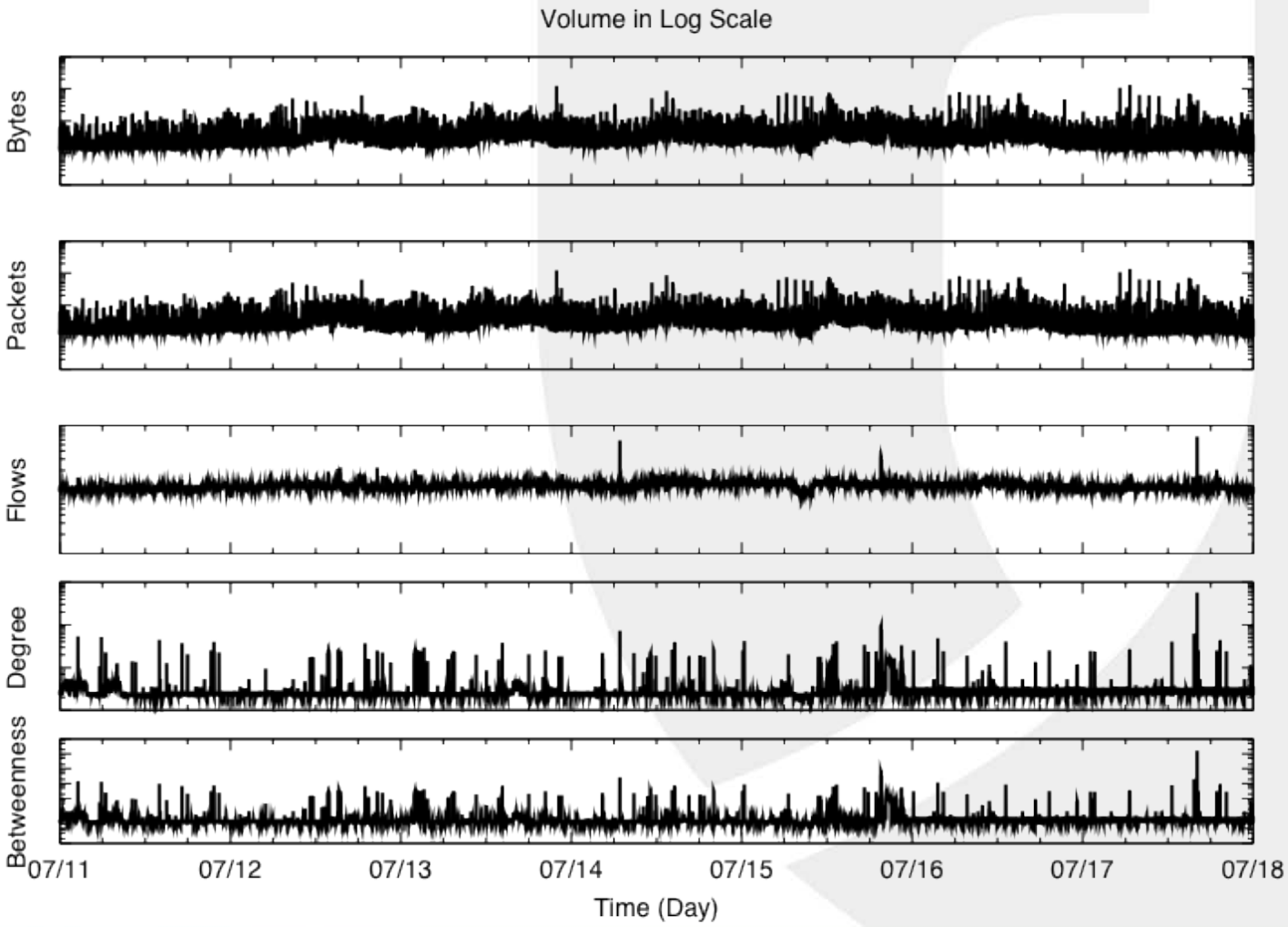
REDJACK

- 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

REDJACK



Only Betweenness Spikes

REDJACK

- 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

REDJACK

- 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

REDJACK

- 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

- Stephen P. Borgatti, “Centrality and Network flow”, Social Neworks, Vol. 27, No. 1. 2005.