Analyzing Flow Using Encounter Complexes

Leigh Metcalf, netsa-contact AT cert.org

Flocon, Charleston
2014

This material is based upon work funded and supported by Department of Homeland Security under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center sponsored by the United States Department of Defense.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

DM-0000740

Clustering for Data Reduction

Instead of examining 5,000 flows over a time period…

Examine 12 clusters instead.

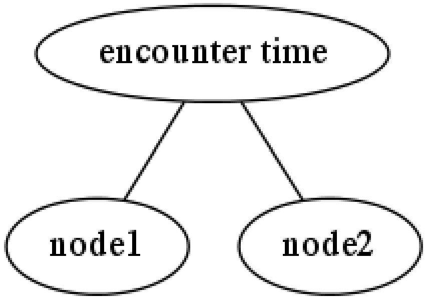
A 99.76% reduction

Previous Work

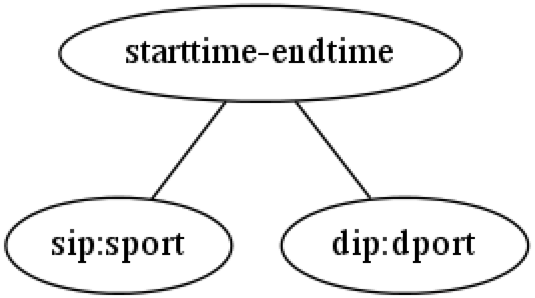
- Network Flow Clustering has been used for:
  - Trojan Detection
    - http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6405737&tag=1
  - Detecting Spoofed Flows
    - http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4239059
  - Finding Botnets
    - http://dl.acm.org/citation.cfm?id=1496721
- Encounter Complexes have been used for:
  - Recovering Spatial Information
    - http://dl.acm.org/citation.cfm?id=1374668

Encounter Trace

Defined as:



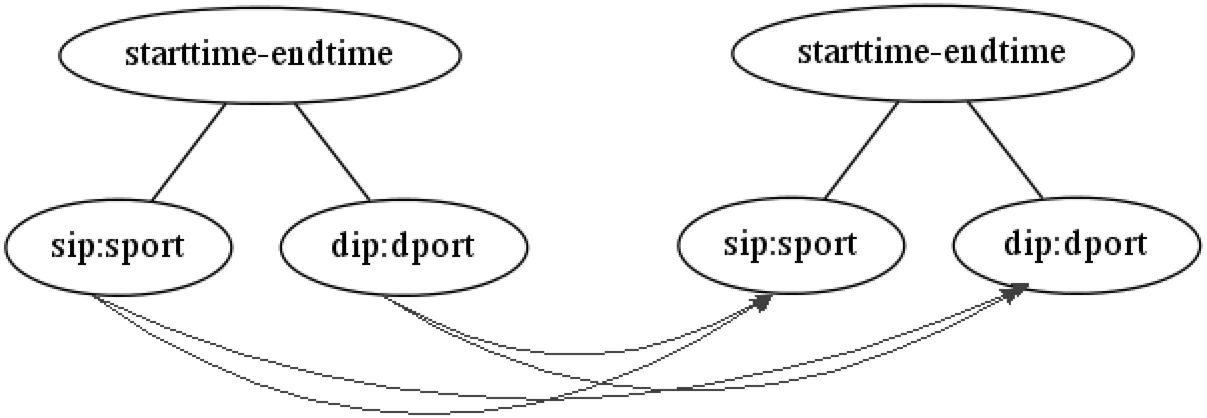
For Flow:



Encounter Complex

Two encounters have an edge between them if:

- They share an endpoint
- The endtime of one is within δ seconds of each other



Encounter Complex

Example of a complex:

```
2009/04/20T11:35:19.529 2009/04/20T11:35:28.935
        10.1.60.203:60515 10.1.60.25:25
2009/04/20T11:36:28.822 2009/04/20T11:36:28.822
        10.1.60.203:51727 10.1.60.25:25
```

Encounter Complex

Example of a complex:

```
2009/04/20T11:35:19.439 2009/04/20T11:35:19.445
          10.1.60.203:50398 10.1.60.187:80
2009/04/20T11:35:19.440 2009/04/20T11:35:19.445
          10.1.60.187:80 10.1.60.203:50398
```

Not Encounter Complexes

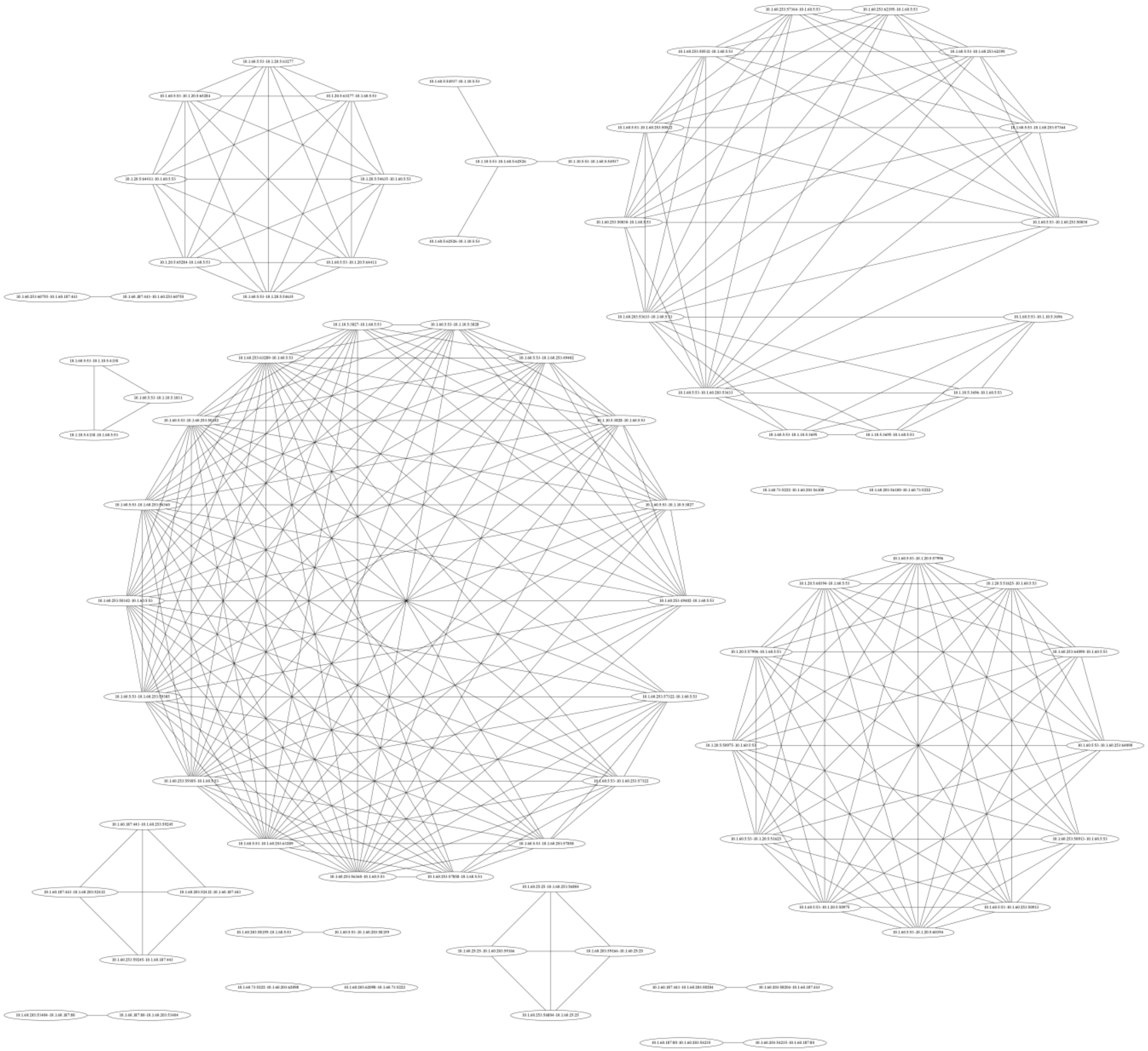
Example 1: The time is too far apart

```
2009/04/20T11:35:19.463 2009/04/20T11:35:19.519
        10.1.60.203:49592 10.1.60.187:443
2009/04/20T13:00:13.738 2009/04/20T13:00:13.738
        10.1.60.187:443 10.1.60.253:56074
```

Example 2: No matching trace found

```
2009/04/20T11:35:19.529 2009/04/20T11:35:28.935
        10.1.60.203:60515 10.1.60.25:25
```

Visualizing This – Is Useless



Analysis

- Graph clusters
  - Each component within the complex is a set of related encounter traces
  - For example:
    - 12 components within one flow when δ=8
  - Degree Analysis
    - What is the encounter with the most connections?
  - Local Clustering Coefficient
    - How dense is my graph

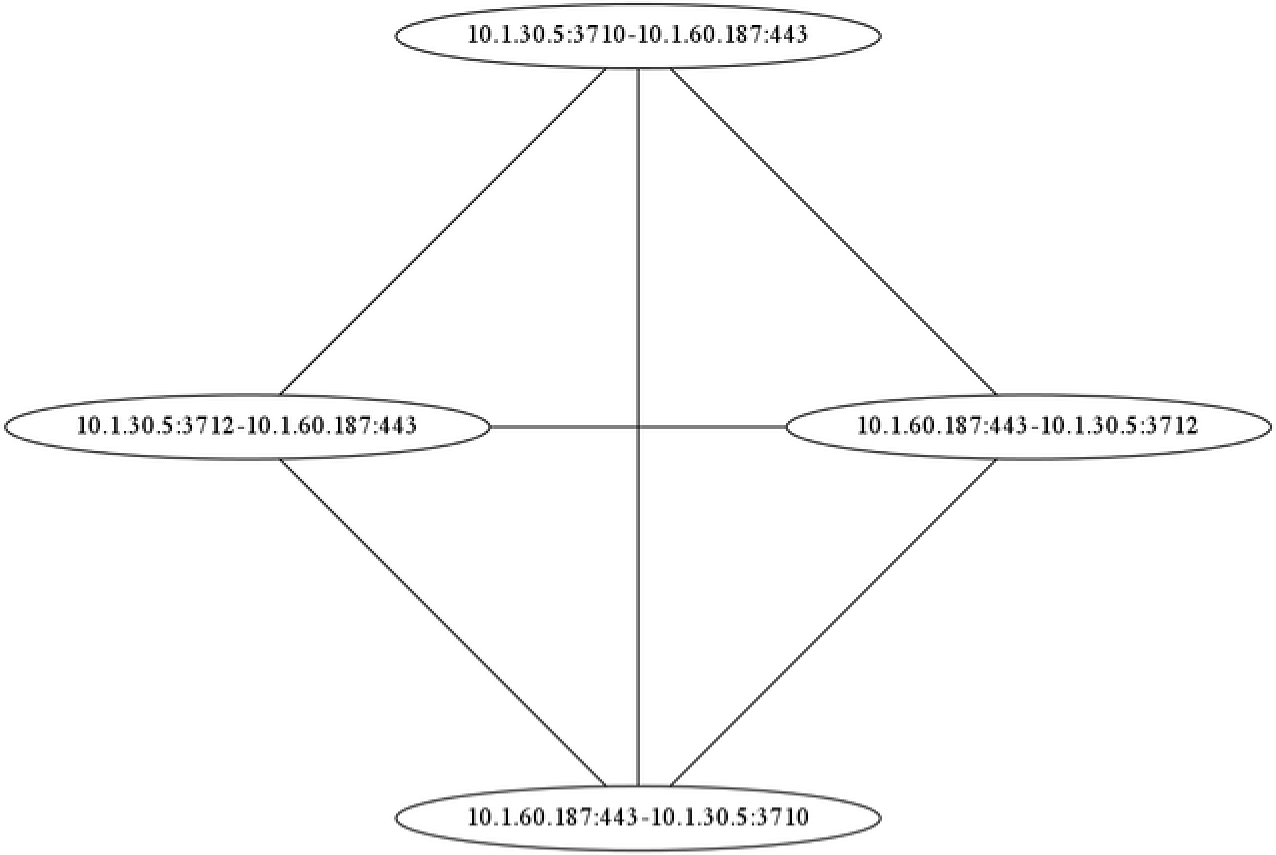
Analysis

Analyzing the largest component created from flow:

- 4,313 vertices, 636,761 edges, 635,631 cycles
- This means we have 4,313 encounter traces, or at worst case, 8,616 flows
- Vertex with highest degree (1,130) is found at:
  - 10.2.195.248:48776 10.1.60.25:25
- Local Clustering Coefficient:
  - Tightly clustered: 0.999994
- Examining the neighbors, it looks like 10.2.195.248:48776 is being very friendly

Analysis

This is what an HTTPS session can look like:



Analysis

This is some of the data from the previous graph:

10.1.60.187:443-10.1.30.5:3710 10.1.60.187:443-10.1.30.5:3712

10.1.60.187:443-10.1.30.5:3710 10.1.30.5:3710-10.1.60.187:443

10.1.60.187:443-10.1.30.5:3710 10.1.30.5:3712-10.1.60.187:443

10.1.60.187:443-10.1.30.5:3712 10.1.60.187:443-10.1.30.5:3710

Analysis

δ makes a difference

- When δ = 1, the graph has 60 components
- When δ = 8, the graph has 12 components
- $G_1 \subseteq G_8$

Increasing δ pulls in more edges

Analysis

Alternative Analysis:

What about those edges which didn’t get an edge in the graph?

In general, these are one-sided conversations.

For example, this appears to be an unanswered ping:

2009/04/21T13:57:21.541 2009/04/21T13:58:02.968 10.1.60.187:0 10.1.100.8:0

Future Work

- Graph fingerprinting
  - Create a graph that looks like a connection with http://www.cnn.com
  - Allow graph edit distance or Jaccard distance to determine similarity with the fingerprint
- Add size of flow as an edge weight to the graphs
- Multigraphs
- Time Series Analysis with Graphs
- A SiLK plugin

CERT

Questions/comments?