Semantic Flow Augmentation for the Automated Discovery of Organizational Relationships

Chris Strasburg*, Harris T Lin, Nikolas Kinkel
The Ames Laboratory
{cstras,htlin,nskinkel}@ameslab.gov
* - Presenting
Relationship Discovery – Why does it matter?

- What is the impact of disrupting communication associated with flow set ‘F’?
Relationship Discovery – Why does it matter?

- What is the impact of disrupting communication associated with flow set ‘F’?
Relationship Discovery – Why does it matter?

- Which alarms are most critical to manually investigate?
What is Semantic Flow Augmentation

- # of packets
- # of bytes
- protocol / port numbers
- duration
- ASN numbers
- ...

Alice ———— Bob
What is Semantic Flow Augmentation

* # of packets
* # of bytes
* protocol / port numbers
* duration
* ASN numbers
* ...

* Time Series Analysis
* Service Type
* Graph Analysis
What is Semantic Flow Augmentation

• Semantic – *Of or relating to meaning...*

**Motivation / Intent**
- Infrastructure Use
- Collaboration
- Resource Needs

**Resource Nature**
- Shopping
- Entertainment
- Communication
- Education

**Analysis Techniques**
- *Time Series Analysis*
- *Service Type*
- *Graph Analysis*
- *Lexical Analysis*
- *Domain Clustering*
- *Session metadata*
Why Semantic Augmentation
Why Semantic Augmentation

Is it mission related?

Strength of Relationship
- # of packets
- # of bytes
- protocol / port numbers
- duration
- ASN numbers
- ...

Class of Relationship
- Motivation / Intent
- Resource Nature

Alice  ---  Bob
Statistical Features

• Flow Statistics
  – # of Flows
  – # of Bytes
  – Peer count

• Timeseries Analysis
  – First seen
  – Last seen
  – Fourier Transform Coefficient
Semantic Features

• Lexical Analysis (Mallet)
  – Cluster according to web page contents from:
    • Reverse DNS Lookups
    • WHOIS Org Searches

• Session Metadata
  – Requested URLs

• Service Distribution
  – Interactive / Authenticated (SSH, IMAP, POP)
  – Interactive / Non-Authenticated (STMP, HTTP/S)
  – Non-Interactive (NTP, DNS)
Semantic Features (2)

- Bi-clique Grouping
  - Red = Internal
  - Green = External
  - Edges pruned
  - LP & BRIM Algorithm**


*Gephi http://gephi.org/
Architecture Overview

- "Raw" Data Repository
  - SiLK
  - Mallet
  - LP & BRIM

- Service Distribution
- Netflow Statistics
- Timeseries Coefficient
- Lexical Clusters
- Biclique Clusters

- Model

- Database

- Labels

- Interactive Resource Evaluation
How to Label / Train

Anecdotal Human Process


Not Mission Related

Time consuming!
Kick Start Labeling

1. Feature Labels
2. Initial rank
3. Assign labels
4. New rank
5. Classifier
6. Train
7. New rank
8. Assign labels
9. New rank
10. Classifier
11. Train

Iteration 1

Iteration 2
Anecdotal Validation – Ames Data

• Gathering Data
  – One month of NetFlow data in Ames Lab

• Preprocessing
  – 4 sets of features: simple NetFlow statistics, time series features, lexical analysis features (document topic distributions), biclique community features

• Labeling
  – 4242 IPs (801 white / 3441 black)

• Testing / verifying classifier
  – Weka (Logistic Regression, SVM, Bayesian Network, Decision Tree)
  – 10 cross-fold validation
Performance Results

Logistic Regression

Decision Tree (C4.5)

- Precision
- Recall
- AUC

Legend:
- Lexical
- CC, Service, Biclique
- Netflow
Info Gain by Features

Lexical Topic: 0.15
Country Code: 0.13
Lexical Topic Conf: 0.10
Total Bytes: 0.08
Total Records: 0.07
Total Dest Port: 0.05
Total Source Port: 0.04
Community Focus: 0.03
Community Ext/Int Size: 0.02
Latest Endtime: 0.02
Access Hours: 0.02
Workhour Ratio: 0.02
Service: 0.02
Access Days: 0.01
Earliest Starttime: 0.01
Peer Count: 0.01
Community Size: 0.00
Fourier Weekly: 0.00
Fourier Daily: 0.00
Implementation at Ames Laboratory

- Alerts
- DWL Client
- Database
- Feature Building / Training
- Automated Response
- Labels
- Alice
Challenges / Future Work

• Majority of IPs don’t have a web page
  – Automated query for WHOIS Organization
  – Use of AMP data; actual HTTP resources

• Speed / Streaming
  – Slow to gather features; currently batched daily.

• Searching
  – Search engines w/ free API (Faroo?)

• Production ‘burn-in’
  – Feedback from analysts into a growing set of labels

• Integration with other systems
  – BroIDS Module?

• Mining of graphical data
  – Second derivative clusters (clusters of clusters)
  – Internal resource categorization
Summary

• Flow provides ‘how much’; a bit of semantics is required for mission relevance.

• Public tools:
  – SiLK – Flow Statistics
  – Crawler4J + Mallet – Lexical Analysis
  – Weka – Machine Learning SAK
  – Apache Commons Math – (Timeseries transforms)
  – A sprinkle of Java and a dash of Python