Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Graph Based Role Mining Techniques for Cyber Security

KIRI OLER, SUTANAY CHOUDHURY

Pacific Northwest National Laboratory

Flocon 2015, Portland, OR, USA

U.S. DEPARTMENT OF ENERGY

Motivation

- Analyzing a machine's past behavior in the face of an alert is a common task for security analysts
- Analyzing past history for a machine can be a time consuming task, and slows down response to the alert
- Therefore, techniques that summarize the past behavior of a system in terms of easily understood cyber features described in English can be a powerful capability
- This work presents Role Mining algorithms as an approach towards accomplishing this goal

Preliminaries

- We use the terms graphs and network almost interchangeably in this presentation
- Graph refers to the mathematical model describing links or edges between a set of entities
  - Example [Computer Network] entities: hosts, links: communication between hosts
  - Example [Social Network] entities: people, links: *friend* as in Facebook, *connection* as in LinkedIn etc.
- A graph is directed, if the relationship between two nodes have a directional aspect
  - Example: My laptop (say 130.20.177.117) making a request to www.google.com (130.20.128.83) => “130.20.177.117 -> 130.20.128.83” in the graph
  - Sutanay and Kiri works together => represented as a directed graph

Preliminaries (cont.)

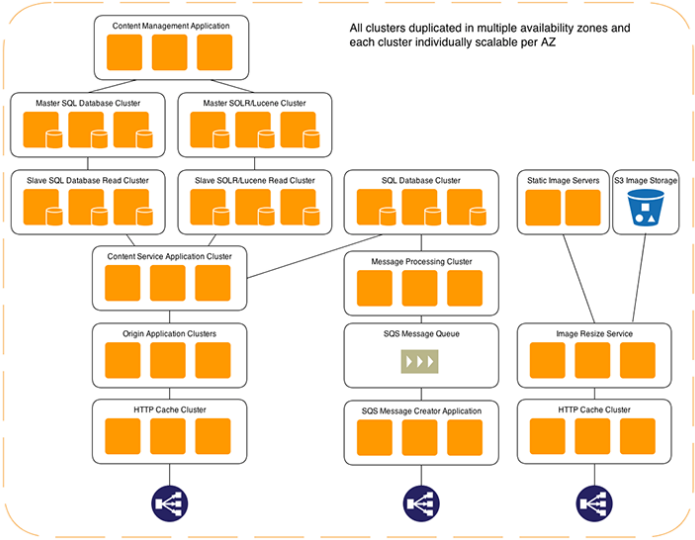
- Formally, a graph is an ordered pair G = (U, E) where U is a set of nodes and E is the set of edges connecting pairs in U
- Weighted graphs can have a weight, or a number associated with each node and/or edge
- We build directed, weighted graphs from flow data
  - A node represents an IP address from network flow data
  - An edge models the summary of communication between two IP addresses
  - Flow attributes such as number of exchanged bytes are aggregated and represented as edge weights
  - Additionally, edges have attributes such as protocol
  - Edge direction indicates directionality of communication

What is Role Mining

- Define meaningful roles for nodes on a network
  - Each node in the network is assigned a set of features
  - Features are based on a node's structural properties and traffic-based behavioral attributes
- Roles are defined based on strength of a subset of features
  - Assume every node is assigned a 8-length feature set (in-degree, out-degree, centrality etc.)
  - Role A may be defined as nodes with "high in-degree and high-centrality"
- Roles have potential to take on different types of meaning, e.g. classifying nodes based on hardware type or criticality to network operations.

Expected Payoffs

- The role distribution can be monitored to detect anomalies or failures in an enterprise
- We are investigating techniques for building multi-scale graph models of an enterprise's cyber behavior
- We foresee "role"-based coarsening as a way to construct multi-scale models of a cyber architecture



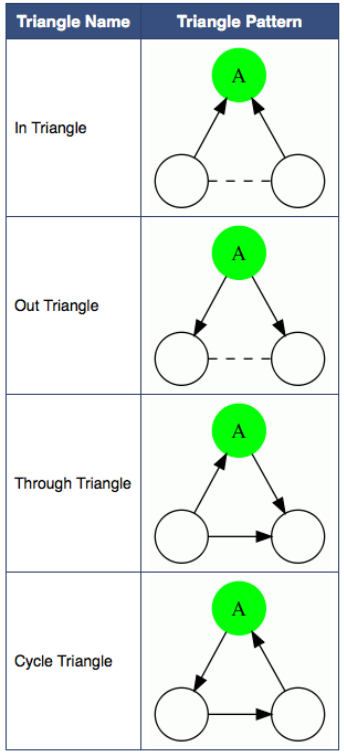
http://aws.amazon.com/solutions/case-studies/discovery-communications/

Algorithm

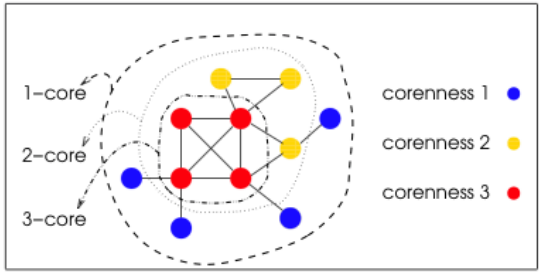
- A matrix of features (V) is generated for each vertex in the graph
  - **Graph theoretic features** In and out degree, number of associated triangle/ triads, k-core ranking, PageRank centrality, clustering coefficient
  - **Flow based features** Median (in and out) flow duration, median (in and out) bytes exchanged, top-k protocols
- The matrix is then factored as $V = G*F$ where:
  - V is the $n\,x\,f$ feature matrix
  - G is the $n\,x\,r$ matrix defining each node's affinity to each role
  - F is the $r\,x\,f$ matrix defining how much each feature impacts each role
- We use the Non-Negative Matrix Factorization algorithm to perform the decomposition
- Once roles are defined F remains constant and new node traffic is classified as follows, $G = V*F\uparrow -1$
- The optimal number of roles is selected by finding the value that yields $G*F$ with Minimum Description Length.

Graph Theoretic Features

- Triangle statistics capture local behavior around a node in the graph
- PageRank based centrality provides a measure of importance of a node in the graph– a higher number indicates higher likelihood of that node being via a random walk on the graph
- K-Core rank provides an indication of the node's position in the graph – a lower rank indicates peripheral presence, whereas a higher number indicates a presence in the "core" network

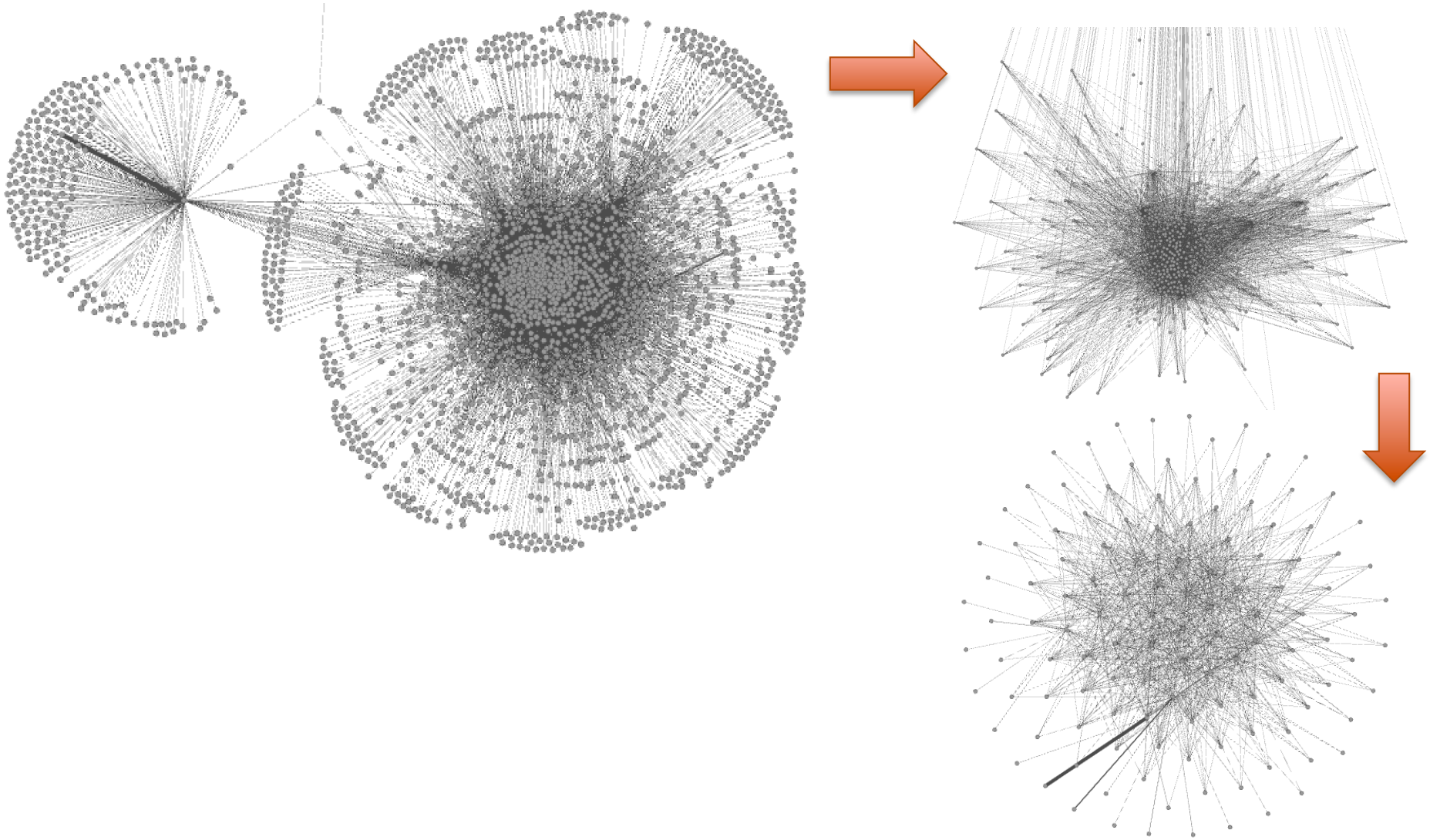


http://docs.graphlab.org/graph_analytics.html



Alvarez-Hamelin, J. Ignacio, et al. "Large scale networks fingerprinting and visualization using the k-core decomposition." *Advances in neural information processing systems*. 2005.

Illustration of graph theoretic features

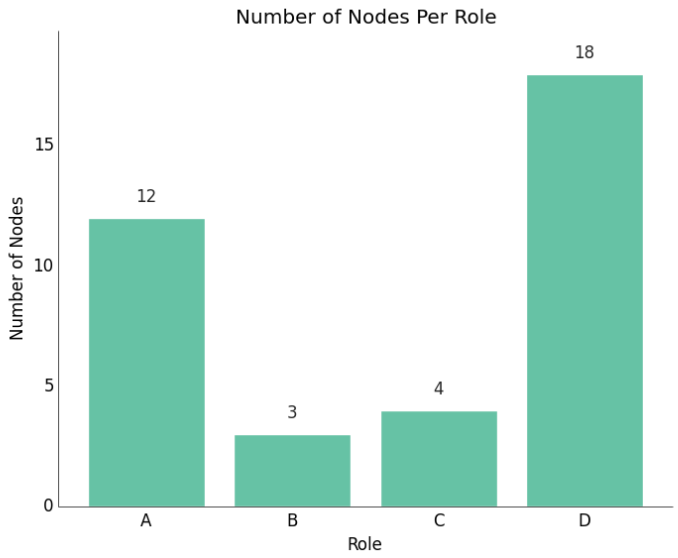


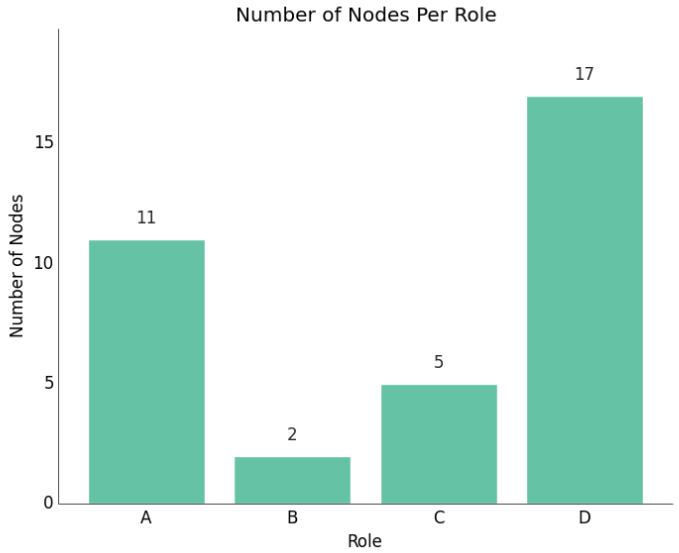
Analysis Objectives

- **Classification Accuracy :**
- We learn the role distribution from a training dataset, say one describing a period of normal activity
- Given a new data set and the previous role definitions, we map each node in the graph to one of the roles
- A node will swap roles only if its behavior has changed significantly
- We test the variation among node classification using the k-nearest neighbors algorithm.

Analysis Objectives

- **Role Distributions** Due to the dynamic nature of networks, the number of nodes identifying as a certain role will vary over time. However, the overall rate of change should hold steady, implying changes to the distribution could be meaningful.

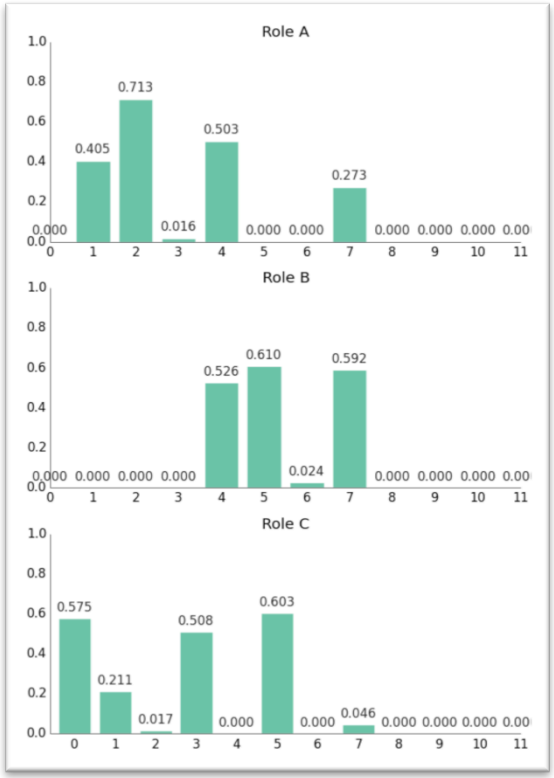


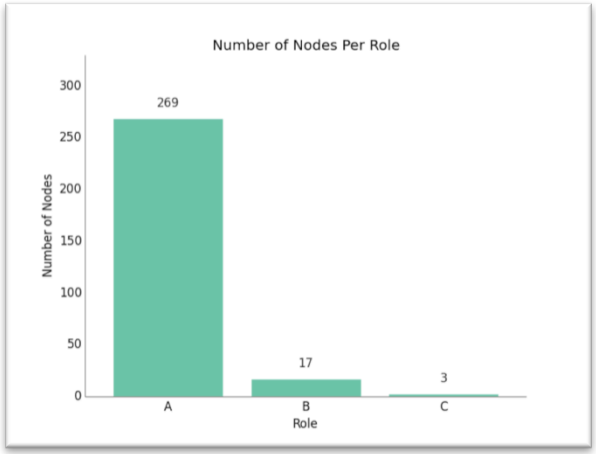


Does the role distribution changes between days?

Algorithm Output

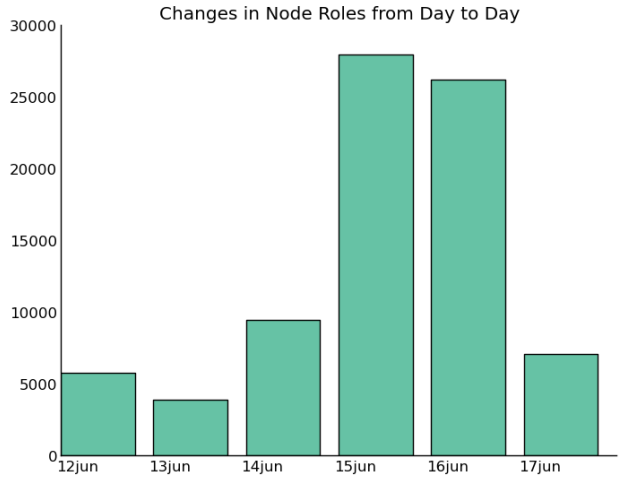
- We get back two things
  - A set of role definitions defined in terms of feature strengths
  - A role assignment for every node





Analysis Objectives

- **Role Changes** Given good role definitions, a node's role is unlikely to vary much over time. The less frequently changes occur, the more likely they are to indicate some anomaly. For instance, if a certain role is heavily influenced by the number of bytes transmitted by a node and a node moves from a role where the number of bytes is relatively small to one where a large number of bytes is large, this could indicate that the node is being exploited to move sensitive data to an undesirable location.



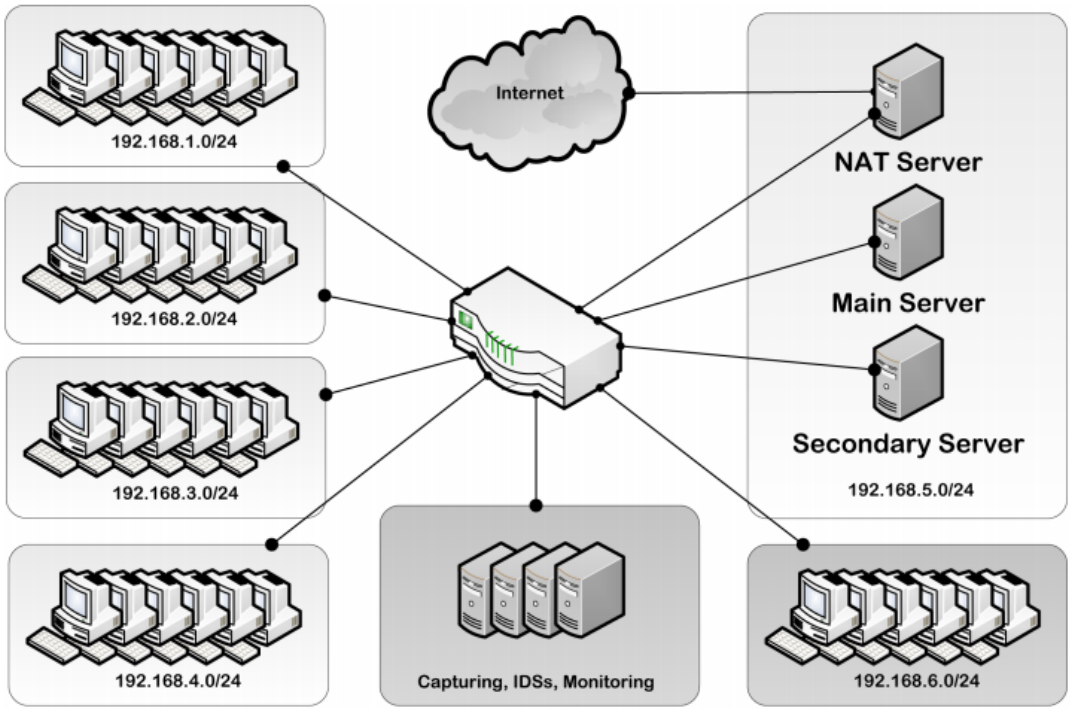
Experimental Analysis – About the Data

- Initial testing made use of traffic captures from an openly available simulated data set* from University of New Brunswick.
- The data was collected from a test bed environment implementing a methodology for generating user profiles and attack scenarios.
- The user profiles are generated based on distributions of packets, flow lengths, requests, endpoints, etc. as observed among users in real traffic flows.
- Attack profiles are generated to simulate buffer overflows, SQL injections, cross-site scripting attacks, DOS attacks and brute force attempts to establish an SSH connection.
- By using simulated data, we have a full knowledge of the anomalous activity within the data set to reinforce the validity of our results.

* Shiravi, Ali, et al. "Toward developing a systematic approach to generate benchmark datasets for intrusion detection." Computers & Security 31.3 (2012): 357-374.

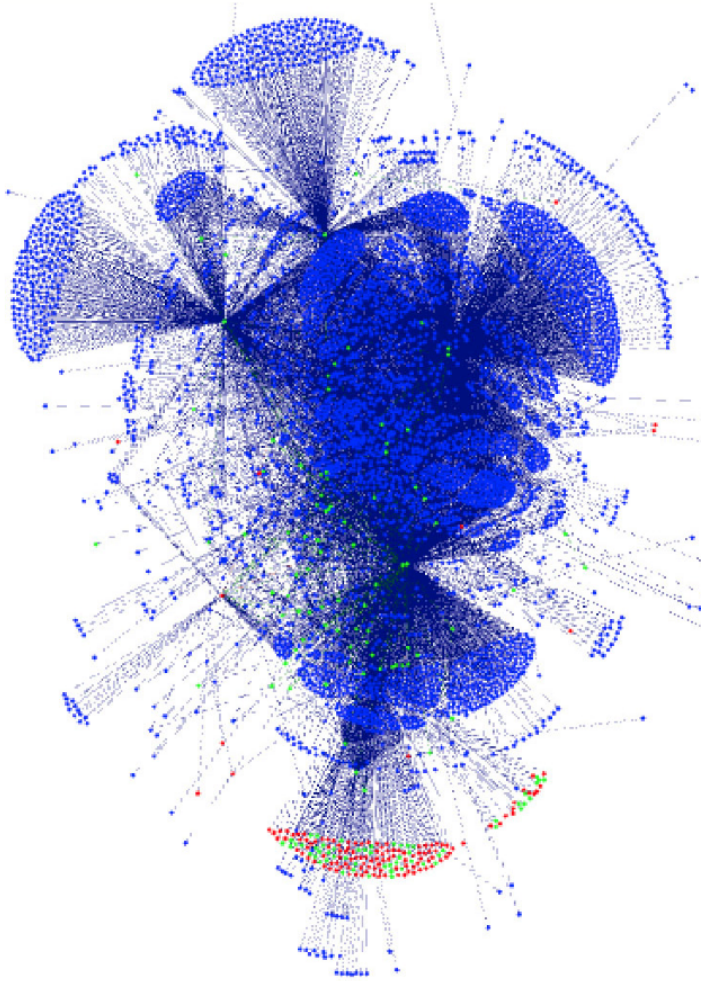
Testbed Topology

- We present our analysis on network traffic flow dataset collected at University of New Brunswick



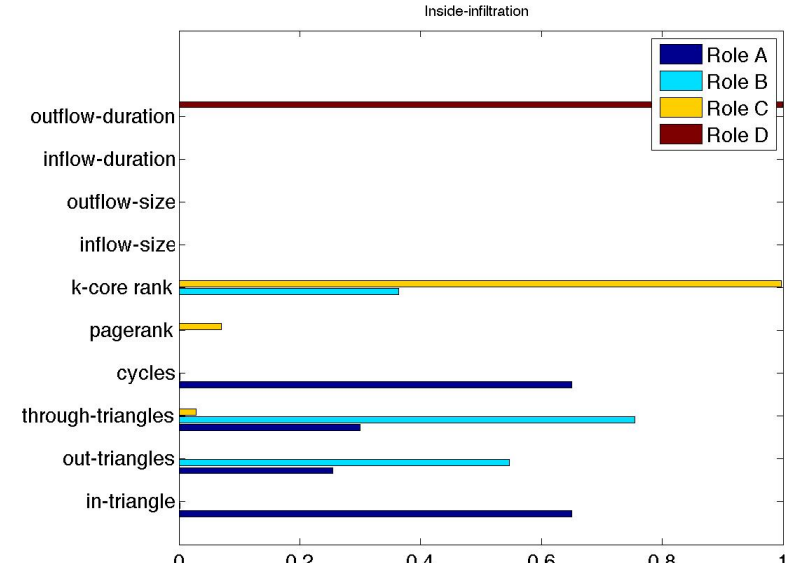
Shiravi, Ali, et al. "Toward developing a systematic approach to generate benchmark datasets for intrusion detection." Computers & Security 31.3 (2012): 357-374.

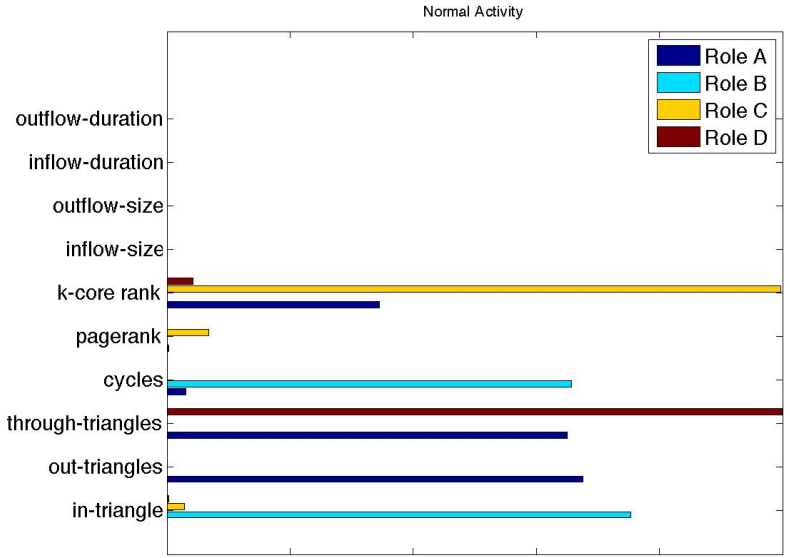
Visualizing IP-graph using Roles

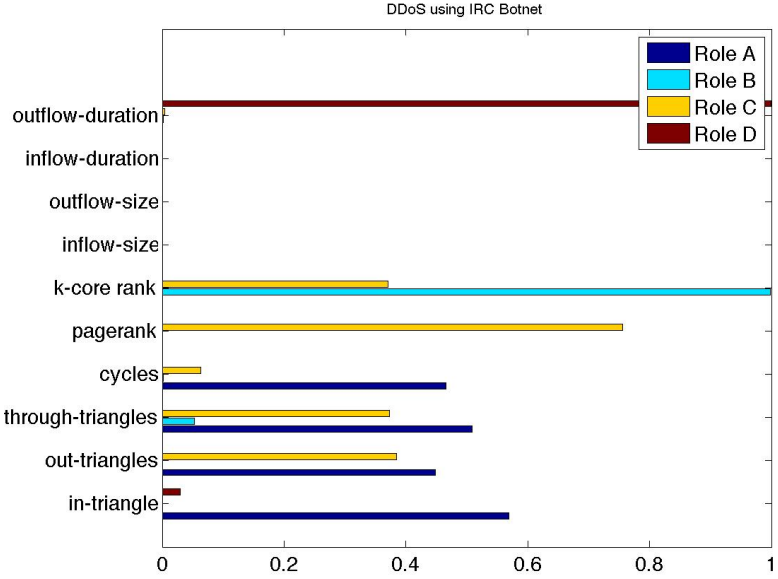


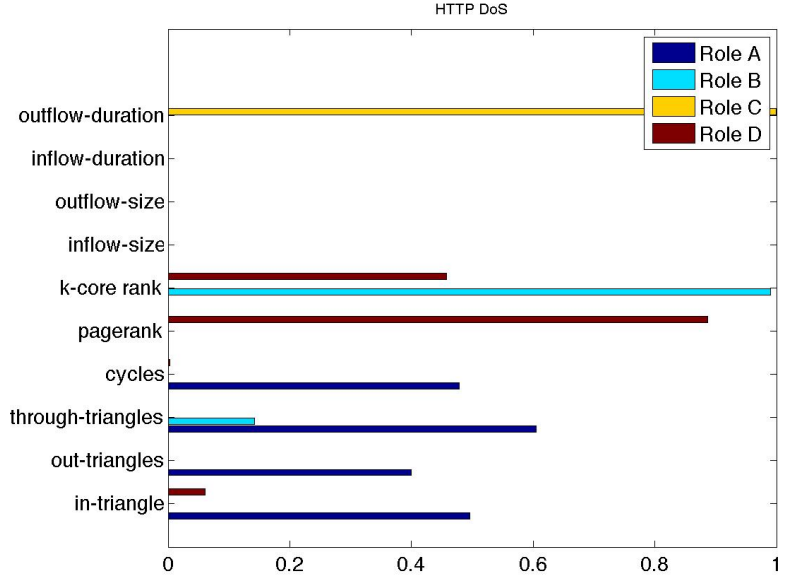
Rendering of Network Traffic Data Showing Communication between IP Addresses. Each IP address is colored by a "behavioral role" learnt using machine learning techniques.

Experimental Analysis – Role Definitions









“Making this work” - Challenges

- How do we make this work at the production environment?
- Remember that we are aggregating traffic within a certain interval and building a graph based model, so what is the right time resolution for aggregation?
- We are looking at netflow and aggregating all traffic. Interesting signals can be drowned in the aggregation process. Which protocols/traffic class is interesting to extract?
- What other features should we consider?
- How can we benchmark the performance?

Conclusions and Future Work

- Given a target dataset, our end goal is to identify the minimal set of features and algorithmic constraints that lead to alignment of roles learnt from the data with real-world roles taken by the machines.
- Discover the types of meaning inherent in the extracted roles and trace how the meaning differs based on the features selected.
- We are currently focusing on testing the algorithm with different combinations of features.
- Experimenting with various constraints to guide the role mining process.
  - **Diversity Constraints** to ensure less overlap between role definitions and stronger role assignments for nodes.
  - **Sparsity Constraints** so the role definitions will gravitate toward a small number of impactful features, while nodes will only be assigned to roles with which they most strongly identify.

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Sutanay Choudhury
Principal Investigator
M&Ms4Graphs

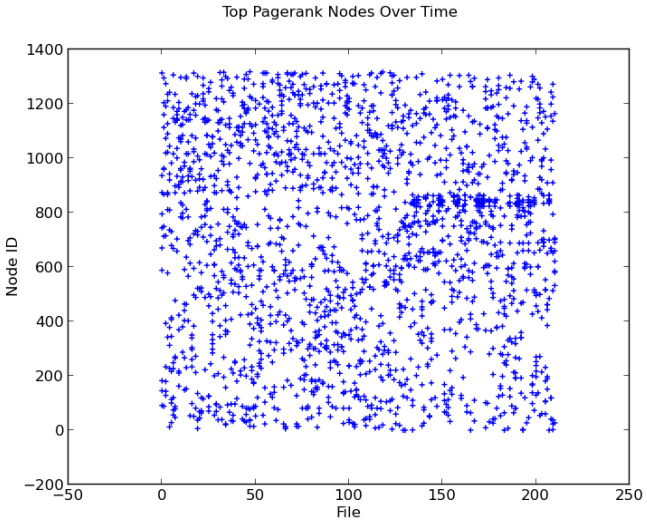
sutanay.choudhury@pnnl.gov

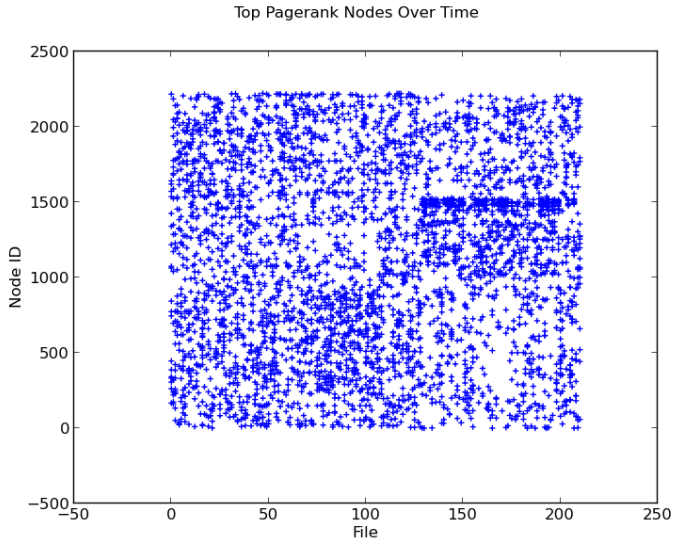
Asymmetric Resilient
Cybersecurity Initiative

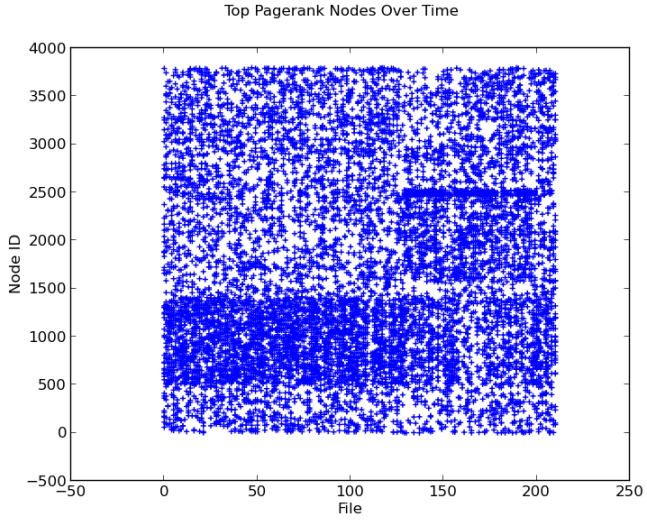
cybersecurity.pnnl.gov

Backup

Evolution of Centrality Distribution







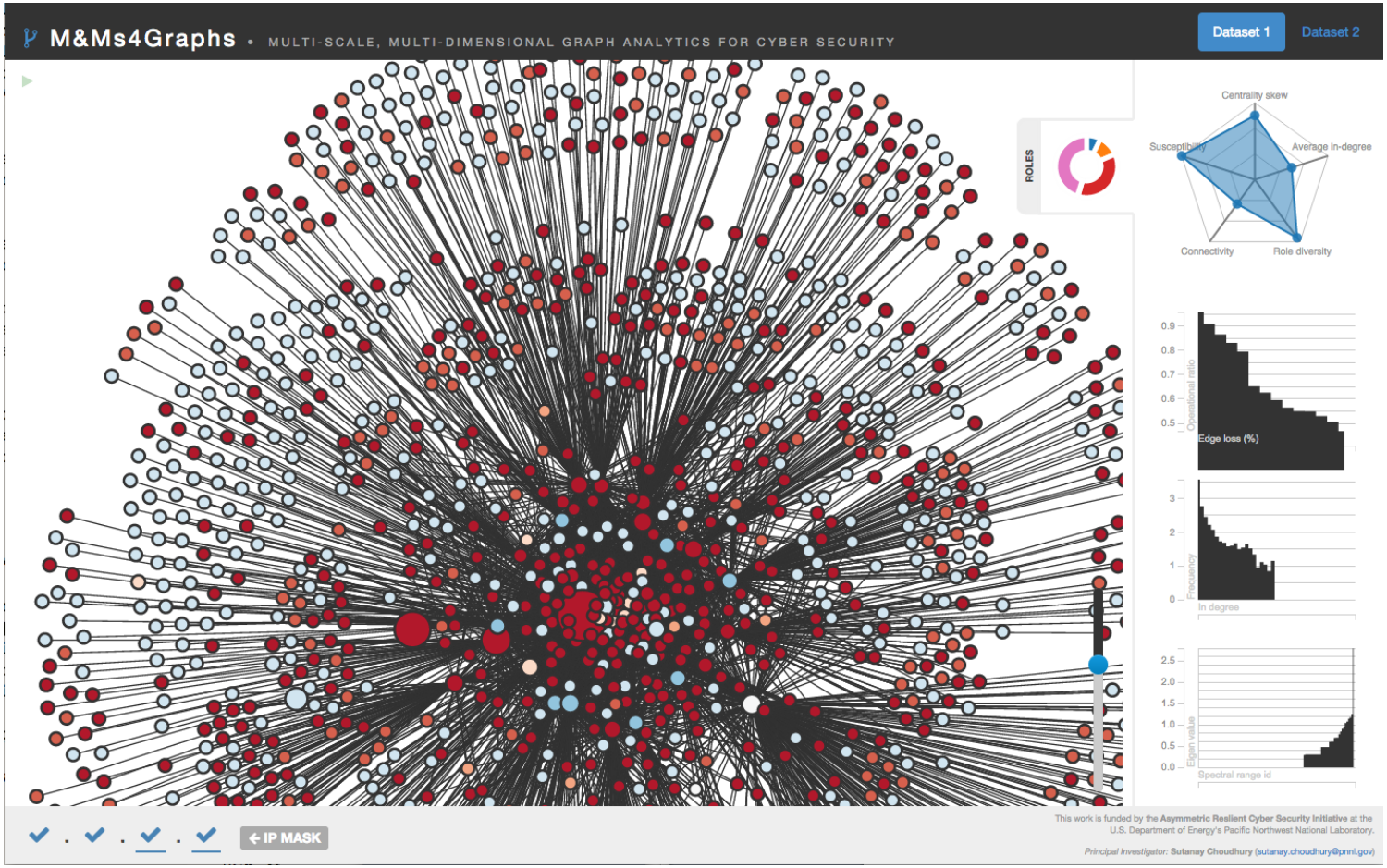
The graph is dynamic, and services become critical depending on the context, time of day/week/year

Plots showing the evolution of the top-10, 30 and 50 node membership from another enterprise network

X-axis showing batch file indices and Y-axis displays The id of the nodes belong in the top-k class

Web based interface

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



http://goo.gl/1iiqc6