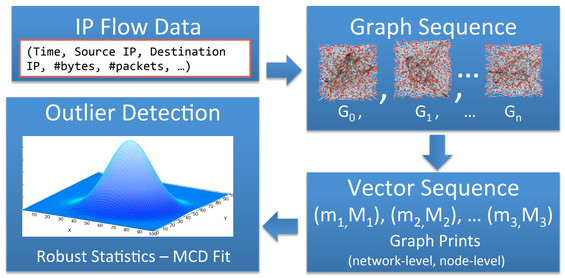
Graph-Prints: A Contextual, Model-free, Multi-Scale Network Analysis Framework for Characterizing Network Flow Data

Christopher Harshaw*, Robert A. Bridges, Michael D. Iannacone, Joel W. Reed , John R. Goodall
*Rice University, Computational and Applied Mathematics
crh7@rice.edu

Anomaly Detection Workflow

Given a sequence of graphs, $\{G_i\}_{i=1}^N$ discover anomalies at multiple related levels



Creating Graphs from Network Flow

- Nodes are IP addresses
- Directed edges are aggregate flows, colored red if both ports >= 1024 black otherwise
- Time windows of 31 seconds with 1 sec overlaps

Test Data Set

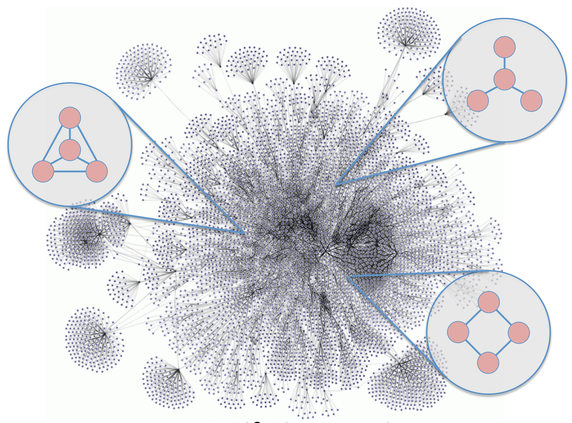
Bit Torrent (BT) Traffic Data Set

- Ambient traffic collected at Oak Ridge National Laboratory using Argus**
  - 5 hours of traffic, 3 million flows
- Anomalous BT traffic collected from personal laptop using Argus
  - 30 min BT traffic, 18k BT flows
- Implanted BT traffic data into ambient data by matching IP, router, and timestamps

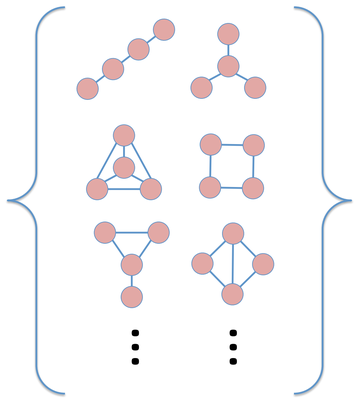
** www.qosient.com

Graph Prints Method

Network Motifs: "Network Building Blocks"



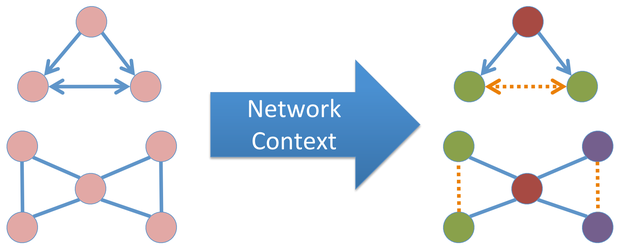
Motif Discovery in Large Network*



Network Motifs

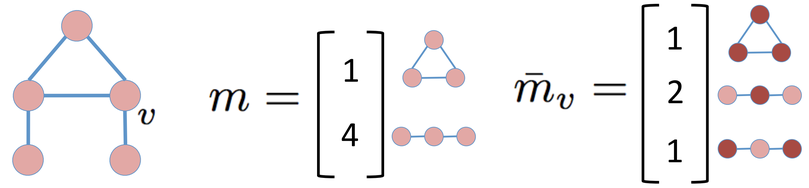
Graph Coloring: "Contextual Information"

- Different Entities
- Different Interactions
- Recognize contextual patterns



Multi-Scale Graph Prints

- Network Level: motif counts, m
- Node Level: motif automorphism orbit counts, $M = \{\bar{m}_{v_1} \cdots \bar{m}_{v_n}\}$
- Graphs to Vectors: $\{G_i\}_{i=1}^N \to \{(m_i, M_i)\}_{i=1}^N$



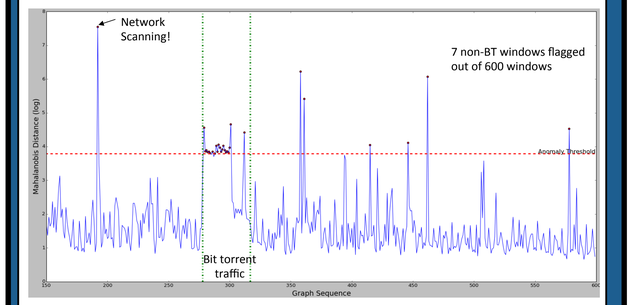
Example Network | Network Motif Count | Motif Automorphism Orbit Count for v

* www.yworks.com

Results

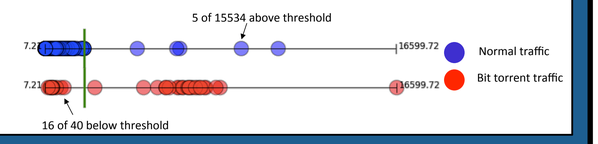
Graph Level Anomaly Detection

- Fit motif vectors to normal distribution using Minimum Covariance Determinant, a robust statistical fitting
- Score vectors by mahalanobis distance and declare anomalous if above a threshold



Node Level Anomaly Detection

- Cluster non-BT node orbit vectors using k-means with gap statistic
- Compute distance to nearest centroid for BT and non-BT node orbit vectors



References

1. Tijana Milenkovic and Natasa Przulj, "Uncovering biological network function via graphlet degree signatures", Cancer Informatics 2008, 6:257-253.
2. Rousseeuw, Peter J., and Katrien Van Driessen. "A fast algorithm for the minimum covariance determinant estimator." *Technometrics* 41.3 (1999): 212-223.

Acknowledgements

This material is based on research sponsored by: the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD) via BAA 11-02; the Department of National Defence of Canada, Defence Research and Development Canada (DRDC); the Kingdom of the Netherlands; and the Department of Energy (DOE). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing official policies or endorsements, either expressed or implied, of: the Department of Homeland Security; the Department of Energy; the U.S. Government; the Department of National Defence of Canada, Defence Research and Development Canada (DRDC); or the Kingdom of the Netherlands