Log-based Anomaly Detection Through Correlation & Behavior Analysis for Cybersecurity

Tony Hu, Zheng Chen
Drexel University

NEED & INDUSTRIAL RELEVANCE

- Malicious cyber activities, such like price scraping, spam distribution, DOS/DDOS, etc., cause economic loss for enterprises. They are becoming ever more distributed and fail many traditional approaches.
- Logs usually come in a large volume. An efficient approach is needed to discover anomalies from large-volume streaming log data.
- Understanding of the cause of abnormal cyber activities can facilitate management decision.

PROJECT GOALS

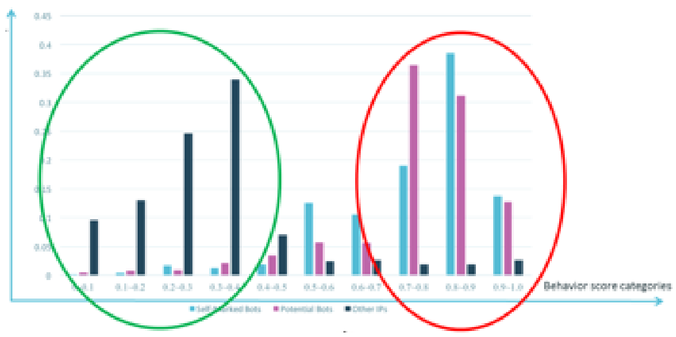
- Design novel efficient methods to discover correlated anomalies from large-volume streaming log data.
- Design methods to understand the purpose of cyber activity anomalies, providing rich information for better management decision.
- Work with IAB members to evaluate the prototype APIs using real-world log dataset and transfer them to the IAB members.

UNIQUENESS/IMPACT

- We focus on application-level log data to avoid sensitive lower-level or hardware-related information that might bring additional security concerns.
- Our methods try to discover anomaly correlations from large-volume streaming log data and aim to provide real-time detection. Existing methods are either batch-based, or not looking for correlations.
- Our methods try to understand the purpose of correlated anomalies by combining contents under a potential probabilistic generative model.

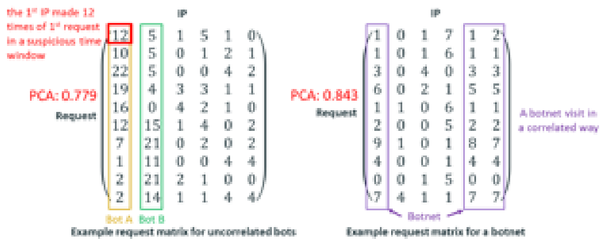
APPROACH (RESEARCH METHODS)

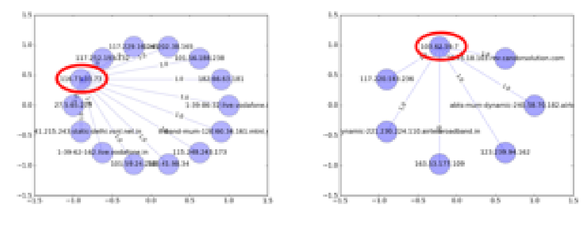
- Model historical field value transitions using a variant of semi-Markov chain: 1) transition time is taken into consideration; 2) primary states are modeled by transition probabilities; 3) secondary states are modeled by density estimation.

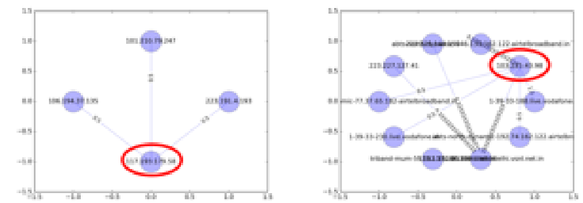


APPROACH (RESEARCH METHODS)

- Use weighted online PCA and spectral clustering to discover correlated anomalies from large-volume streaming log data.







Host	Location	Barracuda Central	WatchGuard
116.73.33.73	Delhi, India	Poor Reputation	Bad
103.62.59.7	Array, India	Poor Reputation	Bad
59.92.161.203	Bangalore, India	Poor Reputation	Bad
122.162.8.30	Delhi, India	Poor Reputation	Bad
117.193.179.58	Bangalore, India	Poor Reputation	Bad
103.225.43.98	Gurgaon, India	Poor Reputation	Bad
115.184.106.205	Mumbai, India	Poor Reputation	Bad

- Use probabilistic generative model to model the logs as being generalized by users when they go through the contents. For example, when a user opens a webpage, there is some probability that one gets focused on certain area of the page; after that one might click on a link in the focused area because of observation of certain keywords.