Correlational and Distributional Relationships between Log Alert Types in System Security Monitoring

Mentee: Michael Colomb | Mentor: Uttam Thakore
University of Illinois Department of Computer Science

Purpose
Security monitors can be imperfect and redundant in the information they provide about incidents. How do combinations of alert types provide novel or redundant information, and what are the implications of this information on IDS confidence and computational cost?

Background

Computer Logs → a record of system activity
- Syslog: system information
- IDS (Bro): suspicious activity
- Netflow (Argus): network traffic

Alert type → specific type of log line
- i.e. failed SSH login or HTTP GET request

Log contents vary depending on activity
- Malicious activity should be reflected

Access to known malicious activity from NCSA
- 23 incidents between 2008 and 2010

Methods

Parsing relevant unstructured, textual logs
- Our supervised machine learning analysis requires structured, labeled data
- Store structured logs in elasticsearch, a database search engine, for easy JSON storage and retrieval

Distribution Generation
- Distribution for each combination of event type and alert type

Analysis:

**Pearson Correlation Between Alert Types of Same Event Type**
Linear strength indicates redundancy across alert types

\[ Q_{xy} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \]

**Jensen-Shannon Divergence Between the Frequency Distributions of the Same Alert Type of Different Event Types**
Divergence directly corresponds to the discriminate power

\[ JS \left( P \left| \| Q \right| \right) = \frac{D(P \left| \| M \right|)}{2} + \frac{D(Q \left| \| M \right|)}{2} \]

\[ M = \frac{P + Q}{2} \]

\[ D(P \left| \| Q \right|) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \]

**Jensen-Shannon Divergence Between the Joint Frequency Distributions of Alert Types**
Divergence directly corresponds to the discriminate power of monitoring alert types in combination

**Future Work**

Finish Current Investigation
- Parse all syslog, IDS, and netflow alerts
- Determine which alerts, and in what combinations, provide the least redundancy and greatest discriminate power

Contribute to security monitor for live systems
- Increase confidence through joint alert monitoring
- Decrease computational cost by ignoring redundant alerts

Skills Gained

- Parsing skills necessary for dealing with large amounts of raw data
- Use of tools such as Python, Numpy, and Elasticsearch

**Acknowledgements & References**

Thank you to Uttam Thakore for his consistent and knowledgeable mentorship with this project and to the PURE committee for making this opportunity available.