Data Sources for Cyber Security Research

Melissa Turcotte
mturcotte@lanl.gov

Advanced Research in Cyber Systems,
Los Alamos National Laboratory

14 June 2018
Background
Advanced Research in Cyber Systems, LANL

Cyber research group at large DOE national research laboratory.

- Identify gaps, emerging threats
- Data collection and sensors
- Modelling and simulation
- Tool/product development

Our perspective

Assume the initial compromise will happen; consequently largely focused on inside-the-perimeter detection.
Cyber Security and Data Science
Attacking have a significant advantage

- Defenders aim to protect networks from both known and unknown threats
- Networks are becoming larger, more complex and permeable
  - Contractor networks
  - Wireless
  - Mobiles and BYOD
- Attackers need only identify and exploit a single weakness
Applicability of Data Science

- Signatures are fragile and reactive
- Leverage anomaly detection and classification to proactively identify new threats
- Robust approaches must focus on *fundamental* behaviours rather than specific features that are easily modified
Data Sources
Sensors and Collection

- Perimeter network data
  - Network flow events, proxy logs, firewall logs
  - Full packet capture

- Internal network data
  - Network flow events, full packet capture
  - DNS data, DHCP logs, e-mail metadata, other central server logs

- Distributed, comprehensive computer data
  - Event/system logs
  - Processes, application, I/O, network, user authentication activity

End Goal: Comprehensive modelling through data fusion.
Cyber Data Challenges

- High volume, can be difficult to collect and process
- Where collected, not necessarily for cybersecurity purposes
  - Debugging
- Where collected, not intended for analytics
  - Primary purpose is to support human operations
  - Variety of formats, fields, etc..
- Data consistency → Not a priority for operational staff
  - Sensors break
  - Collection gets “forgotten” in configuration changes
  - Collection dumps overflow
  - Collection gets turned off
  - “stuff happens”
• Less (or not) useful in its raw form and requires transformation or combination with other data sets prior to analysis
  ▶ IP vs hostname (or other)
  ▶ Inconsistent username identifiers
• Periodicity → user driven actions vs background noise
  ▶ Important to be aware of when modelling
• Low signal to noise ratio
• Rarity of labelled data presents a challenge for supervised learning. No ground truth
• Collection, storage, and use has security, privacy, and human subject research implications

Get involved with the operational staff collecting data and/or data engineers.
Research Opportunities
Public Data Sets

- To motivate a larger research effort focused on operational cyber data LANL have released three publicly available datasets:
  - 9 month time-series user/computer bipartite, 2014
  - 58 day comprehensive, 2015
  - 90 day Netflow and Window Event Logs, 2017
- Internally collected from the LANL corporate network:
  - No outside (Internet) associations
  - Parsed and normalised (to some extent)

https://csr.lanl.gov/data/
Research Opportunities

- Per entity (computer, user, edges, etc..) models that enable anomaly detection
  - Changepoint detection methods
  - Streaming methods for “online detection”
  - Extreme value theory for identifying surprising spikes in non-standard, heavy-tailed distributions

- Score “new” and “rare” activity
  - Latent factor models
  - Community detection approaches

![Graph showing unique edges and new activity over days]
For a given entity, extrapolate to higher-level more interpretable actions.

- Remove or separate user-driven events and correlations

Event times for a user in the LANL network
• Data fusion and meta analysis to increase signal to noise ratio
  ▶ Individual detectors can generate many false positives, combining multiple weak indicators will provide strength to detections.
  ▶ P-value combiners
• Characterise risk within the network
  ▶ Host-user risk analysis.
  ▶ Time-delay metrics on network penetration.
  ▶ Network segmentation.
How do we make anomaly detection more practical?

Data analytics and anomaly detection is very immature (and uncommon) in actual cybersecurity operations.

- Cyber defenders generally discount these approaches in favour of signature detection and intuition

Anomaly detection approaches suffer from:

- High false positives
- Lack of interpretability
- Inability to triage

Three common characteristics of good leads: relevant, detailed, and actionable.

"Incident Response & Computer Forensics", Luttgens et al.
Final Remarks

- Traditional, signature-based methods are insufficient
- Too much attention is given to perimeter prevention
- Internal data collection and analysis is critical
  - Historically limited to governments
  - Commercial product offerings are beginning to appear
- Increasing demand for “data scientists”
  - Data is growing in volume and variety
  - Demonstrated success using statistics
Questions?

mturcotte@lanl.gov