Statistics in cyber security
How we can help, with examples
A statistician’s journey in cyber

2000-2007 – Data science for national security
  • Sshh, it’s a secret

2007-2014 – Worked for LANL
  • Focused in enterprise network anomaly detection research
  • Published a lot
  • Made tools that were licensed commercially

2014-2017 – Adventures in deploying commercial solutions
  • Worked for a consulting company
  • Deploying enterprise anomaly detection solutions at large scale

2017-Present – Lead the Data Science team for Microsoft’s Windows Defender Advanced Threat Protection product suite
Windows Defender Advanced Threat Protection

- Measures data from enterprise Windows machines
- Sends data in near real time to Microsoft Azure
- Detection algorithms are applied
- Detections are reported back to the customer security operation
- Automated response to many detections
Some things I learned along the way

- Industry solutions are mostly heuristic/rule based
  - There is much room for data driven methods

- Supervised ML is starting to penetrate the industry
  - Microsoft uses them extensively in Anti-virus
  - But labeled data can be hard to come by, especially outside of the Anti-virus domain

- Data is hard!
  - Quality and sensor gaps are a major issue
  - Volume and velocity is incredible
  - A lot of my work has been in justifying the need for better collection
  - More in Melissa’s talk coming up

- Adversaries constantly innovate
  - That’s their job
  - To counter this, we need agility and general data driven solutions

- We need more statisticians!
  - We think probabilistically instead of deterministically
  - We like interpretability
    - critical to produce useful solutions
    - We don’t use black boxes
    - We try to understand the data
Two basic areas of need in cyber security

- **Detection**
  - Identify the presence of an attacker in near real time
  - Anomaly detection
  - Supervised learning
  - Heavy tails
  - Low SNR
  - Non-stationary null and alternative
  - High dimensional
  - Graph Topology

- **Risk**
  - Establish risk scores for users and machines
  - Statistical process control
  - Multidimensional
  - Millions of users/assets under monitoring
Rules and Supervised ML
The state of the art today

Positives
- Gotta cover the knowns
- Very good TP performance
  - High precision
- Good interpretation
  - We know why they match
- Capture the expert knowledge efficiently

Negatives
- Require labeled data
  - This is hard to get and ever changing
- Not general
  - Missing FNs
  - Especially Advanced Persistent Threat
- Not scalable
  - Require heavy maintenance
  - New ones all the time
  - Performance degradation
- No control over alarm rate
  - If it alarms it alarms
Anomaly detection Positives and Negatives

Positives
- Unknowns/Generality
  - No assumption on attack
  - Just modeling normal
- Scales well
  - Fully automated updating
  - Handles non-stationarity
  - Means less maintenance cost long run
  - Can control alarm rate explicitly
- Scores provide natural prioritization
  - Look at the weirdest thing first
  - Good for hunting
- High resolution
  - One model per asset/user/pair of connected machines, etc

Negatives
- Ease of combination
  - Identify comprehensive kill chain
- Finds more FN

- Reduced FP performance
  - Unless you combine well
- Requires modeling expertise
  - Statisticians can help!
- Can be difficult to interpret
  - Unless you use interpretable models and interpretable, attack-consistent combinations
Rules

Supervised ML

Anomaly Detection

We need good methods to combine

Alerting and Hunting

Raw telemetry
Comprehensive kill chain anomaly detection

Initial penetration
- Deviations in Email behavior due to phishing barrage

Persistence and callback
- Deviations in processes, command lines, registry, etc
- Deviations on network, low reputation, beaconsing, etc
- Credential deviations

C2/Recon
- Deviations in network comms
- Internal Port deviations
- Deviations in HTML/DNS requests for covert channel C2

Lateral Movement
- Network and OS deviations
- Credential anomalies
- Insider/pattern of life anomalies

Staging
- Visible in anomalous volumes and ports focused on one destination host

Exfiltration
- Visible in anomalous volumes leaving the network

* Red indicates deviations the attacker has introduced in the normal behavior of the endpoints and communications
We can detect this object comprehensively!

Overall Score = \( f(\alpha_i) \) e.g. \(-\sum_{i=0}^{n} w_i \log(\alpha_i)\)
Anomaly detection steps

- Modeling
  - Establish stochastic models
  - At high resolution
  - Across many data streams
- Streaming Updating
  - Update each model as it sees data
- Scoring
  - Assign scores to current data with respect to model

- Calibration
  - Correct for poor model fit
  - Under the null, scores should be $U(0,1)$
- Combination
  - Combine scores across the kill chain to maximize power
Anomaly detection flow chart

- **Score data**: $P_0(d_0) = s_0$
  - Calibrate($s_0$)
  - Combine scores $P(\text{attack}) = P_a$
  - Update Model $P_k$
  - Use $P_a$'s for Hunting
  - Threshold $P_a$'s for Alerting

- **Score data**: $P_k(d_k) = s_k$
  - Calibrate($s_k$)
  - Update Model $P_k$

- Time sliced Data $d_0$
- Time sliced Data $d_k$
IoT devices

Third party vendor network

Network Perimeter

IoT devices
Smooth data with zeros

Bursty data with human behavior
User Risk Scoring (URS)

• Based on following Insider Threat attack:
  • For two weeks, act anomalously but not maliciously
    • Strange times of day
    • Connecting to strange servers
    • anomalous social media behavior
  • Attack Phase
    • Late at night
    • Login to multiple unusual servers
    • Copy data to user machine
    • USB event or other Exfil channel

• Method
  • Model Credential behavior per user
  • Use self-updating, lightweight models
  • Score deviations on host and deviations across network
  • Combine atomic scores
  • Accumulate risk scores over longer time periods, eg CUSUM
  • Decay risk scores
  • Deviation from population and self (Bayesian population priors)
Concluding remarks

• The data is enormous, high velocity, and challenging
• The problems are difficult with high noise and low signal
• There is a significant need for more statistical approaches
  • Interpretability
  • Understanding the data
• We need your help!
  • I am seeking collaborations with Gov and Academia
  • We are hiring statisticians!!
  • We love interns!!!
Joshua.Neil@Microsoft.com

selected publications