Final Capstone Proposal

Kimberly Nowell-Berry

Thinkful Data Science

For my final capstone in the Thinkful Data Science Course, I will be implementing multiple models in attempt estimate the probability of sequences in computer log activity. I will determine the likelihood that a user follows a particular path.

When a malicious user or attacker attacks a network, there are generally traces of this behavior left behind in the form of logs, whether these are endpoint logs, e.g., Windows or Linux Operating system event, access and security logs, network based logs, e.g., Intrusion Detection Systems (IDS), netflow logs, firewall logs, router logs, or application logs, e.g., HypterText Transfer Protocol (HTTP) web server logs, database logs, etc. With the multiple different logs, identification of a malicious user is difficult, and anomaly detection plays a critical role in identifying malicious behavior.

While anomaly detection identifies interesting events in these logs, it is typically not sufficient individually to determine whether or not a particular event is actually malicious. In order to just that, multiple individual anomalous events that can be correlated together provide a strong indication that actual malicious activity is present.

In this Capstone, I will identify anomalous activity in log data using multiple anomaly detection techniques, including but not limited to Kalman Filters and time series anomaly detection, relate those back to tactics, and use a Markov chain to predict the probability that a user performed those activities. Probabilities that are very low would indicate that that sequence of events has happened and would be a strong indication that an attack has occurred. This utility of this model is two-fold:

- Correlating anomalous events together automatically reduces the human requirement to manually pull these events together, freeing security analysts to perform other tasks
- Providing the probability of a particular sequence being followed can aide in attribution of specific attacks, which is real-world, hard problem

Data

For this Capstone, I will be using the sample data available from Carnegie Mellon's Software Engineering Institute. The data consists of email, http, network logon events, and others. The data contains specific attacks that have been identified in separate answer keys so that results can be validated against the true results.

This data will be downloaded and imported into dataframes for analysis.

Challenges

The biggest challenge I anticipate is achieving good results. This is an experient to determine whether the theory of mapping the individual anomaly detection results to a Markov chain for predictive capabilities is viable solution to this problem. If so, this would be significant contribution to Threat Hunting in the CyberSecurity field.