Project Title:

Analysis and detection of malicious events in Network traffic

(Part 1)

Submitted by: Group #04

1. Tim Coleman (timothy.coleman@colorado.edu)
2. Anusha Basavaraja (Anusha.Basavaraja@colorado.edu)
3. Christopher Bisbee (Chris.Bisbee@colorado.edu)

Description:

- As the number of IoT and network related devices keeps growing at an exponential rate, security is a major concern among network administrators and companies at large trying to secure the networks and devices from malicious attacks such as ransomware, spyware, DDoS, and other types of attacks.
- The goal of this project is to analyze and learn interesting patterns from network traffic that contains known malicious intent in an effort to try and prevent attacks from happening and mitigating loss.
- There are published data sets such as CTU-13 and IoT 23 which contain netflow and pcap network captures that are classified as benign or malicious. Part of the goal for this project would be to build a machine learning model in an attempt to detect malicious events and optimize the balance of false positives to false negative ratios and maximize the true positive detection rate.
- The group would like to explore possibilities to map network data and selected attributes to data warehouse multidimensional databases such as HDF5 to explore other interesting patterns.

Prior Work:

- Botnet detection approaches have evolved extensively and expeditiously to cope with the development in botnet architectures and protocols.
- Botminer, an approach based on group behavior analysis, combines both packet payload and network flow monitors for botnet detection. The model employs clustering approaches to find similar communication behaviors, as well as network activities. Correlations between the formed clusters is used to identify botnets and infected hosts. *(G. Gu, R. Perdisci, J. Zhang and W. Lee, "BotMiner: clustering analysis of network traffic for protocol- and structure-independent botnet detection", Proc. 17th USENIX Security Symposium, pp. 139-154, 2008)*
- Zhao et al. investigated a botnet detection system based on packet header information and time intervals. Decision Tree based machine learning algorithms were utilized to generate detection models using network flow features of traffic packets. On their generated data set focusing on P2P botnets, their method achieved high accuracy with small time windows. *(D. Zhao, I. Traore, B. Sayed, W. Lu, S. Saad, A. Ghorbani, et al., "Botnet detection based on traffic behavior analysis and flow intervals", Computers & Security, vol. 39, pp. 2-16, 2013)*
- Haddadi et al. employed three machine learning algorithms, namely C4.5 Decision tree, Bayesian Networks and Genetic programming-based SBB, for building detection models. They achieved very high detection rates both for HTTP and P2P based botnets. *(F. Haddadi and A. N. Zincir-Heywood, "A Closer Look at the HTTP and P2P Based Botnets from a Detector's Perspective", Proc. 8th InternationalSymposium on Foundations and Practice of Security (FPS 2015), pp. 212-228, 2015.)*

Datasets:

- We are planning to use two Datasets, where one of it is ready to use and another one requires some preprocessing task to transform it. Later, both datasets will be integrated for our model.
- URL to our Datasets:
  https://www.stratosphereips.org/datasets-ctu13
  https://www.stratosphereips.org/datasets-iot23
- The Stratosphere Laboratory is the provider of the above datasets.https://www.stratosphereips.org/team
- As the dataset is readily available and there is no limitation on size of data to be downloaded, we plan to download small sample of data first and test our code on it. Once the code runs well, then we plan to download entire dataset onto cloud and start our testing phase.
- To be precise, Tim has datasets readily available in his system currently.

Proposed Work:

- Major steps in our proposed work include - data preprocessing, feature extraction, and data mining. (i.e., data cleaning, data transformation, data integration, data reduction and data classification task)

- The CTU-13 data is in netflow v5. The IOT23 is raw packet capture in .pcap format. As our group would like to make use of both data sets, the first step would be to transform the IOT23 data into netflow v5 using python based libraries to handle the conversion.

- Once both data sets are in proper format, we could map data attributes to a multidimensional “data cube” such as hdf5 format and use this for data exploration to perhaps trigger further questions (e.g. categorize flows by destination IP range and look at the distribution of flows across by regions).

List of tools we intend to use:

Python (libraries):

- PyPCAPKit for extracting PCAP file: https://pypi.org/project/pypcapkit/
- NetworkX for network analysis in Python: https://networkx.org/
- Netflow: https://pypi.org/project/netflow/
- Scikit-learn for implementing Machine Learning in Python: https://scikit-learn.org/stable/
- Scapy - https://scapy.net/
- Data Visualization: https://plotly.com/python/

DB:

- The hdf5 library: https://www.hdfgroup.org/solutions/hdf5/

Evaluation Metric:

- Evaluating the machine learning model will involve training the model on part of the classified data and testing it on the other half to determine how well the model performs in regards to accuracy, precision and recall, which are common metrics used to determine the overall performance.

Thank you!