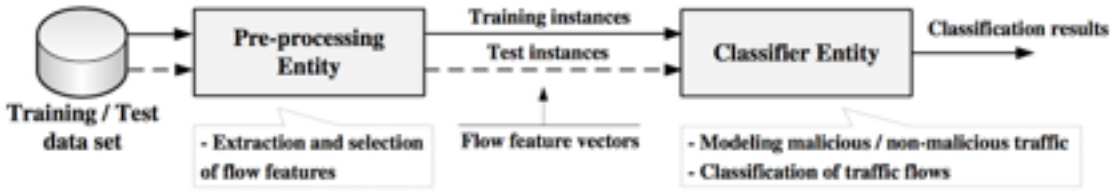
Flow based botnet detection using supervised learning

Problem Statement

Botnets are one of the most serious threats to the Internet security and one of the most challenging topics within the fields of computer and network security today. Botnets rep- resent a usually large collections of computers compromised with a sophisticated bot malware, that puts them under the control of a remote attacker. Botnets deploy C&C channel using a variety of communication protocols, such as: IRC, HTTP / HTTPS and in the most recent botnets, P2P protocols. While a large set of net flow data is available, the information that the net flow data includes is only metadata of the network flow such as flow duration, total bytes, transport header flags. Although this is not ideal data source for botnet detection, because of the large availability of net flow data, it is worth while to build a supervised detection model based on net flow features. The contribution of this project is to evaluate the performance between 2 of the highly regarded supervised machine learning models, Artificial neural network and Random forest, for the task of classifying botnet traffic based on the net flow features.

Proposed approach

At a high level following models the design of the botnet detection system using supervised machine learning algorithm.



The following features shall be extracted from the labelled net flow data set.

Feature	Type
Source port	Numerical
Destination port	Numerical
L4 Protocol identifier	Categorical
Total number of packets	Numerical
Total number of Bytes	Numerical
Mean of number of Bytes per packet	Numerical
Median of number of Bytes per packet	Numerical
Std of number of Bytes per packet	Numerical

Feature	Type
Flow duration	Numerical
Mean of inter-arrival time (IAT)	Numerical
Median of IAT	Numerical
Std of IAT	Numerical
Mean Unmatched flow density	Numerical
Std Unmatched flow density	Numerical
Temporal features	Numerical

To build a detection model, I plan to experiment with random forest algorithms and artificial neural networks.

Milestones

Sprint - 1	Oct 4 - Oct 17	Finalize on the Feature set and the way to extract it from the data set
Sprint - 2	Oct 18 - Oct 31	Implement feature extraction
Sprint - 3	Nov 1 - Nov 14	Implement neural network classifier model
Sprint - 4	Nov 15 - Nov 28	Implement random forest algorithm classifier model
Sprint - 5	Nov 29 - Dec 12	Evaluate the results

Data sources

CTU-13 DATASET: Consists in thirteen captures (called scenarios) of different botnet samples. The netflow data is bidirectional and labelled ~ 20M flows. Nearly 2% of the total traffic are labelled as normal, 0.05% of the total flows are labelled as botnet and the rest as background. The model will be evaluated on the data set by performing 10-fold cross validation.

Tools

Software: Python
Packages: sklearn.ensemble, TensorFlow

Deliverable items

1. A working software with both the classifiers
2. Comparison between both the classifiers based on the metrics Precision/ Recall

References

1. Survey of the P2P botnet detection methods
2. An efficient flow-based botnet detection using supervised machine learning
3. DISCLOSURE: Detecting Botnet Command and Control Servers Through Large-Scale NetFlow Analysis
4. Machine learning based botnet detection with dynamic adaptation
5. Botnet detection techniques- review, future trends, and issues
6. BotFinder- Finding Bots in Network Traffic Without Deep Packet Inspection
7. An-empirical-comparison-of-botnet-detection-methods_2014_Computers-Security
8. An analysis of network traffic classification for botnet detection