Project Proposal Name: Meshal Farhan AL-Anazi

# The Efficiency of machine learning Techniques in minimizing Dwell Time of Cyberattacks

Cyberattacks Dwell Time is the time between when a compromise first occurs and when it is detected. CrowdStrike Services Report Shows that Average dwell time grew 10 days to 95 in 2019, up from 85 in 2018. According to attivonetworks.com in 2020 the problem became more worse with notable jump of Cyberattacks Dwell Time to 100 days. Advanced adversaries and state-sponsored threat actors (APT) are constantly modify their tactics and techniques to evade the detection techniques that allow them to remain undetected for a long time. This project tries to answer the below question:

**Can the machine learning Techniques efficiently detect the emergent APT techniques early to minimize the Dwell Time of cyberattacks?**

## Dataset:

advanced persistent threat (APT) able to impact national security and economic stability of any country and they mainly use botnet to perform stealthy cyberattacks. Botnet is considered a multifunctional malware that pretend the normal traffic by leveraging common protocols such as IRC, HTTP, DNS and P2P for command control. All those reasons make the early detection to minimize botnet attack Dwell Time is challenging.

So, in this project the bidirectional NetFlow Botnet datasets will be used. This dataset is created by CTU University, Czech Republic. The datasets consist of thirteen labeled files with fifteen features. each file consists of around two million instances.

The big size of dataset is realized, so datasets of size 50,000 instances will be sampled during preparation phase.

## Tool:

The below python packages will be used:

**Pandas** for data analysis and preparation

**Matplotlib and Seaborn** for data visualization

**scikits.learn and XGBoost** for model fitting and evaluation

**Hyperopt** for Bayesian optimization (if there is enough time to perform wrapping feature selection to find optimal relevant features)

**Note:** the main purpose of hyperopt is for hyperparameter optimization, however in this project it will be used with certain trick to find optimal features.

## Minimum Viable Product (MVP):

The core steps to answer the project question are:

- Dataset cleaning, preparation and exploring
- Two sets of datasets will be created: the first set consists of one training dataset with wider variants of Botnet. On other hand, the second dataset consists of sub-datasets for each day of attack period (the number of days in attack period will be identified during deep dive of dataset exploration). The test

dataset will contain one variant of botnet represents emergent botnet variant.

- The relevant feature will be identified initially using feature-target mutual information (if there is enough time Bayesian optimization will be applied).
- The optimal classification model will be selected by employing cross validation, optimal feature and hyperparameter tuning
- with corresponding test dataset for each day in attack period, The optimal model will be evaluated to get TPR and FPR