C3i Hub, Indian Institute of Technology Kanpur

HCL HACK IITK 2020

Challenge Round 2

Instructions:

1. There are two questions given below, you have to solve any one of them and submit
2. Please read the Questions very carefully
3. You may, if you want, attempt both and submit what you think is a better solution
4. You can make multiple submissions, only the last submission is taken for evaluation, the previous versions are automatically erased

Question 1: Project (Bot-net Detection)

Description: A bot-net is a network of infected hosts (bots) that works independently under the control of a Botmaster (Bot herder), which issues commands to bots using command and control (C&C) servers. Traditionally, bot-nets used a centralized client-server architecture which had a single point of failure but with the advent of peer-to-peer technology, the problem of single point of failure seems to have been resolved. Gaining advantage of the decentralized nature of the P2P architecture, botmasters started using P2P based communication mechanism. P2P bot-nets are highly resilient against detection even after some bots are identified or taken down. P2P bot-nets provide central frameworks for different cyber-crimes which include DDoS (Distributed Denial of Service), email spam, phishing, password sniffing, etc. So, the objective is to develop a tool for identifying P2P bot-nets using network traffic analysis. Also, the developers should detect the hosts involved in P2P traffic and then the detected hosts are further analyzed to detect bot-nets.

Steps to follow:

- **Data collection**: Collect Botnet analysis dataset (PCAP files) provided.
- **Phase 2 (Botnet detection):**
  Follow below steps to detect botnet traffic.
  - **Feature extraction:** Find out the prominent features from the dataset collected.
  - **Feature selection:** Select only important features so that prediction time will be reduced.
  - **Classification:** Use machine learning classifiers to train the classifiers using extracted features.
- **Develop tool:** Develop a command line tool accepts a pcap file as input and outputs whether it corresponds to p2p botnet activities.

Project must fulfill these requirements as mentioned below:

Tool must have good accuracy, precision, recall, and F-score for machine learning model with low false positive and low false negative rate.

Tool will take PCAP file as input and output will tell whether packet capture traffic contains malicious (botnet) traffic or not.

NOTE: To train the model do not use all files as you will need to test the various figures of efficacy of your models. So, keep 25% of malicious and 25% of benign file data for testing purposes.

Deliverable:

Final submission will be

- A botnet detection tool named botnetdetect.py. This tool takes pcap file as an input and output which traffic is malicious (related to botnet) or normal/benign in a text file in the same folder where script executes. For example:

  python botnetdetect.py absolute_path_for_pcap_file

- Trained model that botnetdetect.py loads while testing new pcap file.
- Preprocessed data in a .csv file with machine learning model source code that is used to train your model for validation of your submitted model.
- And also submit the readme file containing all the information about the tool like software, libraries used, requirement for fresh installation and how to use the tool.
- All the above mention details must be submitted in a single folder in zip format.

Dataset for Botnet Detection Problem:

Botnet_Detection_Dataset.7z: 1.4GB
AWS S3 Link https://mettl-miscellaneous-public.s3.ap-south-1.amazonaws.com/client_public_data/369004/botnet/Botnet_Detection_Dataset.7z

Dataset Description:

Botnet_Detection_Dataset: Zip file size is **1.4 GB** and after unzip **5 GB** (Downloading size is equal to ZIP size)

Total botnet dataset = 45 PCAP files
Total Benign dataset = 47 PCAP files
All Packet capture file contains millions of packets for analysis.
1.4 GB: Botnet_Detection_Dataset.7z
After extraction of the Botnet detection dataset folder size will be 5GB.
Total Storage needed for analysis is 5 GB.

Tree Structure of Botnet_Detection_Dataset folder for few samples

```
Botnet_Detection_Dataset
├── Benign
│   ├── ip_details.txt (Text file containing IP address)
│   ├── p2pbox1
│   │   ├── p2pbox1.2011032611.pcap.clean.pcap
│   │   ├── p2pbox1.2011032702.pcap.clean.pcap
│   ├── p2pbox2
│   │   ├── p2pbox2.2011032800.pcap.clean.pcap
│   │   ├── p2pbox2.2011032909.pcap.clean.pcap
│   └── torrent
│       ├── torrent_00009_20180321021851.pcap
│       ├── torrent_00011_20180321041851.pcap
└── Botnet
    ├── storm
    │   ├── gtisc-winobot.20071027.1193443201
    │   ├── gtisc-winobot.20071027.1193446801
    ├── storm-IP (Text file containing IP address)
    ├── vinchuca
    │   ├── abc2.pcap
    │   └── abc.pcap
    ├── vinchuca_IP (Text file containing IP address)
    ├── zeus
    │   ├── zeus1-10.pcap
    │   ├── zeus1-11.pcap
    └── zeus_IP (Text file containing IP address)
```

Here IP addresses are used to identify the network traffic.

Question 2: Project (Ddos attack Detection)

Description: Distributed Denial of Service (DdoS) attack is a menace to network security that aims at exhausting the target networks with malicious traffic. Although many statistical methods have been designed for DDoS attack detection, designing a real-time detector with low computational overhead is still one of the main concerns. On the other hand, the evaluation of new detection algorithms and techniques heavily relies on the existence of well-designed datasets.

Steps to follow:

- **Data collection**: collect Ddos attack detection dataset provided.
- **Feature extraction:** extract features from the collected dataset using a script.
- **Feature selection:** Select only important features so that prediction time will be reduced.
- **Classification:** Use machine learning classifiers to train the classifiers using extracted features.
- **Develop tool:** Develop a command line tool that accepts a pcap file as input and detects if the pcap corresponds to a ddos attack or not.

Project must fulfill these requirements as mentioned below:

Tool must have good accuracy, precision, recall, and F-score for machine learning model with low false positive and low false negative rate.

Tool will take PCAP file as input and output will tell whether packet capture traffic contains malicious (Ddos) traffic or not.

NOTE: To train the model do not use all files as you will need to test the various figures of efficacy of your models. So, keep 25% of malicious and 25% of benign file data for testing purposes.

Deliverable:

Final submission will be

- A Ddos detection tool ddosdetect.py. This tool takes pcap file as an input and output which traffic is malicious (related to Ddos attack) or normal/benign in a text file in the same folder where script executes. For example:

  python ddosdetect.py absolute_path_for_pcap_file

- Trained model that ddosdetect.py loads for testing new pcap file.
- Preprocessed data in a .csv file with machine learning model source code that is used to train your model for validation of your submitted model.
- And also submit the readme file containing all the information about the tool like software, libraries used, requirement for fresh installation and how to use the tool.
- All the above mention details must be submitted in a single folder in zip format.

Dataset for Ddos Detection Problem:

1: Ddos_Detection_Dataset.7z: 1.4GB
AWS S3 Link https://mettl-miscellaneous-public.s3.ap-south-1.amazonaws.com/client_public_data/369004/ddos/Ddos_Detection_Dataset.7z
2: Ddos_Detection_Dataset_Part2.7z: 1.5GB
AWS S3 Link https://mettl-miscellaneous-public.s3.ap-south-1.amazonaws.com/client_public_data/369004/ddos/Ddos_Detection_Dataset_Part2.7z

Dataset Description:

Ddos Attack detection dataset is divided into two parts: candidates can download one part and do training and they can download another part to retrain.

Ddos_Detection_Dataset: Zip file size is **1.4 GB** and after unzip **19.6 GB** (Downloading size is equal to ZIP size)

Total DdosAttack dataset = 87 PCAP files
Total Benign dataset = 61 PCAP files
All Packet capture file contains millions of packets for analysis.
1.4GB: Ddos_Detection_Dataset.7z
After extraction of the Ddos attack detection dataset folder size will be 19.6GB.
Total Storage needed for analysis is 19.6 GB.

Tree Structure of Ddos_Detection_Dataset folder for few samples

```
Ddos_Detection_Dataset
├── Ddos_Attack_data
│   ├── 001
│   ├── 002
└── Ddos_benign
    ├── benign1
    │   ├── 1.pcap
    │   ├── 2.pcap
    ├── benign2
    │   ├── 10.pcap
    │   ├── 11.pcap
    ├── benign3
    │   ├── 10.pcap
    │   ├── 11.pcap
    ├── ip_details.txt (Text file containing IP address)
    └── p2pbox1
        ├── 1.pcap
        ├── 2.pcap
        ├── 3.pcap
```

Ddos_Detection_Dataset.7z:
AWS S3 Link https://mettl-miscellaneous-public.s3.ap-south-1.amazonaws.com/client_public_data/369004/ddos/Ddos_Detection_Dataset.7z

Ddos_Detection_Dataset_Part2: Zip file size is **1.5 GB** and after unzip **31.7 GB** (Downloading size is equal to ZIP size)

Total DdosAttack dataset = 150 PCAP files
Total Benign dataset = 60 PCAP files
All Packet capture file contains millions of packets for analysis.
1.5GB: Ddos_Detection_Dataset_Part2.7z
After extraction of the Ddos attack detection dataset folder size will be 31.7GB.
Total Storage needed for analysis is 31.7GB.

Tree Structure of Ddos_Detection_Dataset_Part2 folder for few samples

```
Ddos_Detection_Dataset_Part2
├── Ddos_benign
│   ├── 10.pcap
│   ├── 11.pcap
│   ├── 12.pcap
│   ├── 13.pcap
│   └── 9.pcap
├── Ddos_Dataset_Part2
│   ├── 01
│   ├── 02
│   ├── 03
│   ├── 98
│   └── 99
└── ip_details.txt
```

Ddos_Detection_Dataset_Part2.7z:
AWS S3 Link https://mettl-miscellaneous-public.s3.ap-south-1.amazonaws.com/client_public_data/369004/ddos/Ddos_Detection_Dataset_Part2.7z