**Author:** Abdul Ghouse

**Email:** aghouse@gmail.com

**Date:** 11/11/2023

**Content:** Data Sheet

**Topic:** Cybersecurity Machine Learning, threat detection sniffing packets on the network

**Version:** 1.0

* 1. **Motivation:**
* This datasheet provides information about the data source for this cyber security dataset.
* Master dataset for this initiative has much larger dataset for IOT, sysems, etc. This content is focused more on network data.
* Master data source (including network) is from University of New South Wales(UNSW): [The TON\_IoT Datasets | UNSW Research](https://research.unsw.edu.au/projects/toniot-datasets)

Subset data under evaluation, cleaned network data (Train\_Test\_Netwrk.csv): Link

UNSW Canberra, Australia. University of New South Whales Australia. This is a dataset that has been collected from the network and IOT (Internet of things) data set both organically and synthetic setup. Simulating normal network traffic and attacks. It consists of 65% normal traffic and 35% attack traffic captured. The target column identifies normal “0” and malicious traffic “1”. The data further captures or subclassified what kind of attack this malicious traffic is. Idea is to advance cyber security interests with ML algorithms and identify malicious traffic with high accuracy and automation.

* 1. **Composition**

Network dataset source is a csv file. It has 46 columns or features and 461043 rows or records.

These are records of packet captures from the network represented as records.

Index(['ts', 'src\_ip', 'src\_port', 'dst\_ip', 'dst\_port', 'proto', 'service', 'duration', 'src\_bytes', 'dst\_bytes', 'conn\_state', 'missed\_bytes', 'src\_pkts', 'src\_ip\_bytes', 'dst\_pkts', 'dst\_ip\_bytes', 'dns\_query', 'dns\_qclass', 'dns\_qtype', 'dns\_rcode', 'dns\_AA', 'dns\_RD', 'dns\_RA', 'dns\_rejected', 'ssl\_version', 'ssl\_cipher', 'ssl\_resumed', 'ssl\_established', 'ssl\_subject', 'ssl\_issuer', 'http\_trans\_depth', 'http\_method', 'http\_uri', 'http\_version', 'http\_request\_body\_len', 'http\_response\_body\_len', 'http\_status\_code', 'http\_user\_agent', 'http\_orig\_mime\_types', 'http\_resp\_mime\_types', 'weird\_name', 'weird\_addl', 'weird\_notice', 'label', 'type'], dtype='object')

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

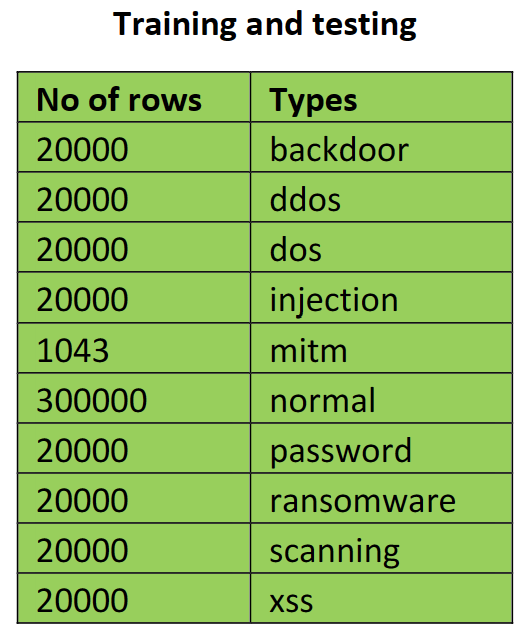
A screenshot of a computer program

Description automatically generated

A close-up of a computer screen

Description automatically generated

Records with network attack types and normal counts/quantity are listed below.



The dataset by nature is sparce and has time stamps. Dataset contains 39.4% malicious traffic records and 65.0% normal traffic.

* 1. **Collection Processes**

To simulate malicious traffic networks, systems (windows and Linux) systems, virtual switches, monitoring agents/sniffing tools etc. are setup to capture traffic to source records. NOTE: Edge layer is also a part of this topology to collect stats on IOT devices. We will be focusing this initiative on the network stats.

A diagram of a computer

Description automatically generated

* 1. **Preprocessing cleaning / labeling**

This research initiative has provided us with a great test-train data set. In our case sourcing the data has been done for us. Our task in this case would be to focus on EDA (exploratory data analysis) massage the data and convert text to workable scaled numbers to use for our machine learning algorithms.

* Network traffic source and destination: Often when there are threats in the network some of the initial thoughts are to identify where the attack vectors started and where is the target data flowing to. Although sophisticated attacks can take multiple hops identifying source and destination locations could be useful to gauge the nature of the threats. So we are using open source IP – location translation database from GeoLite2 to generate this location plot.

[GeoLite2 Free Geolocation Data | MaxMind Developer Portal](https://dev.maxmind.com/geoip/geolite2-free-geolocation-data)

* This dataset is sparce and has several “-“ and “unknown” in the data, we are dropping columns that are largely filled with those.
* We will also use the epoc time stamps to human readable time stamps and make that the index in case we want to review events based on timelines. In our case since its classification we are dropping time stamps at least for this phase of the classification study.
* Data type of the data is reviewed to make sure we can work with the data types.
* We will use OneHotEncoder() and StandardScaler() to encode and scale the data to desired formfactor to feed the data as inputs to various models.
* The dataset is now on a workable format. However the number of fields are large (49) which poses a challenge to computing cost and time.
* This part of the exercise is for us to reduce the features. We are using correlation matrix to identify what features are highly co-related and can be cut to improve model stability. This now reduces the features by just (3) giving us a resulting features of (46)
* We will now focus on model evaluation and improving computing time and costs to get effective results.

1. **Dataset Other factors to consider:** Threat vectorsincreases on a regular basis if this initiative or effort is put to practice or production use capturing packets and retraining the models needs to be done on a regular bases based on the model accuracy drift you are beginning to see.

**Use of this dataset, distribution, and credits:** All use of this data, data capture credit and guidelines can be obtained below.

* UNSW Cranbera Faculty and students. [ADFA | UNSW Canberra](https://www.unsw.edu.au/canberra)
* [Dr Nour Moustafa (unsw.edu.au)](https://www.unsw.edu.au/staff/nour-moustafa) and team.
* [The TON\_IoT Datasets | UNSW Research](https://research.unsw.edu.au/projects/toniot-datasets)