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

1.1 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
  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.

1.2 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')
```

Description of Network Features

Service profile: Connection activity			
ID	**Feature**	**Type**	**Description**
1	ts	Time	Timestamp of connection between flow identifiers
2	src_ip	String	Source IP addresses which originate endpoints' IP addresses
3	src_port	Number	Source ports which Originate endpoint's TCP/UDP ports
4	dst_ip	String	Destination IP addresses which respond to endpoint's IP addresses
5	dst_port	Number	Destination ports which respond to endpoint's TCP/UDP ports
6	proto	String	Transport layer protocols of flow connections
7	service	String	Dynamically detected protocols, such as DNS, HTTP and SSL
8	duration	Number	The time of the packet connections, which is estimated by subtracting 'time of last packet seen' and 'time of first packet seen'
9	src_bytes	Number	Source bytes which are originated from payload bytes of TCP sequence numbers
10	dst_bytes	Number	Destination bytes which are responded payload bytes from TCP sequence numbers
11	conn_state	String	Various connection states, such as S0 (connection without replay), S1 (connection established), and REJ (connection attempt rejected)
12	missed_bytes	Number	Number of missing bytes in content gaps

Service profile: Statistical activity			
ID	**Feature**	**Type**	**Description**
13	src_pkts	Number	Number of original packets which is estimated from source systems
14	src_ip_bytes	Number	Number of original IP bytes which is the total length of IP header field of source systems
15	dst_pkts	Number	Number of destination packets which is estimated from destination systems
16	dst_ip_bytes	Number	Number of destination IP bytes which is the total length of IP header field of destination systems

Service profile: DNS activity			
ID	**Feature**	**Type**	**Description**
17	dns_query	string	Domain name subjects of the DNS queries
18	dns_qclass	Number	Values which specifies the DNS query classes
19	dns_qtype	Number	Value which specifies the DNS query types
20	dns_rcode	Number	Response code values in the DNS responses
21	dns_AA	Bool	Authoritative answers of DNS, where T denotes server is authoritative for query
22	dns_RD	Bool	Recursion desired of DNS, where T denotes request recursive lookup of query
23	dns_RA	Bool	Recursion available of DNS, where T denotes server supports recursive queries
24	dns_rejected	Bool	DNS rejection, where the DNS queries are rejected by the server

Service profile: SSL activity			
ID	**Feature**	**Type**	**Description**
25	ssl_version	String	SSL version which is offered by the server
26	ssl_cipher	String	SSL cipher suite which the server chose
27	ssl_resumed	Bool	SSL flag indicates the session that can be used to initiate new connections, where T refers to the SSL connection is initiated
28	ssl_established	Bool	SSL flag indicates establishing connections between two parties, where T refers to establishing the connection
29	ssl_subject	String	Subject of the X.509 cert offered by the server
30	ssl_issuer	String	Trusted owner/originator of SLL and digital certificate (certificate authority)

Service profile: HTTP activity			
ID	**Feature**	**Type**	**Description**
31	http_trans_depth	Number	Pipelined depth into the HTTP connection
32	http_method	String	HTTP request methods such as GET, POST and HEAD
33	http_uri	String	URIs used in the HTTP request
35	http_version	String	The HTTP versions utilised such as V1.1
36	http_request_body_len	Number	Actual uncompressed content sizes of the data transferred from the HTTP client
37	http_response_body_len	Number	Actual uncompressed content sizes of the data transferred from the HTTP server
38	http_status_code	Number	Status codes returned by the HTTP server
39	http_user_agent	Number	Values of the User-Agent header in the HTTP protocol
40	http_orig_mime_types	String	Ordered vectors of mime types from source system in the HTTP protocol
41	http_resp_mime_types	String	Ordered vectors of mime types from destination system in the HTTP protocol

Service profile: Violation activity			
ID	**Feature**	**Type**	**Description**
42	weird_name	String	Names of anomalies/violations related to protocols that happened
43	weird_addl	String	Additional information is associated to protocol anomalies/violations
44	weird_notice	bool	It indicates if the violation/anomaly was turned into a notice

Service profile: Data labelling			
ID	**Feature**	**Type**	**Description**
45	label	Number	Tag normal and attack records, where 0 indicates normal and 1 indicates attacks
46	type	String	Tag attack categories, such as normal, DoS, DDoS and backdoor attacks, and normal records

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

Training and testing

No of rows	Types
20000	backdoor
20000	ddos
20000	dos
20000	injection
1043	mitm
300000	normal
20000	password
20000	ransomware
20000	scanning
20000	xss

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

1.3 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.

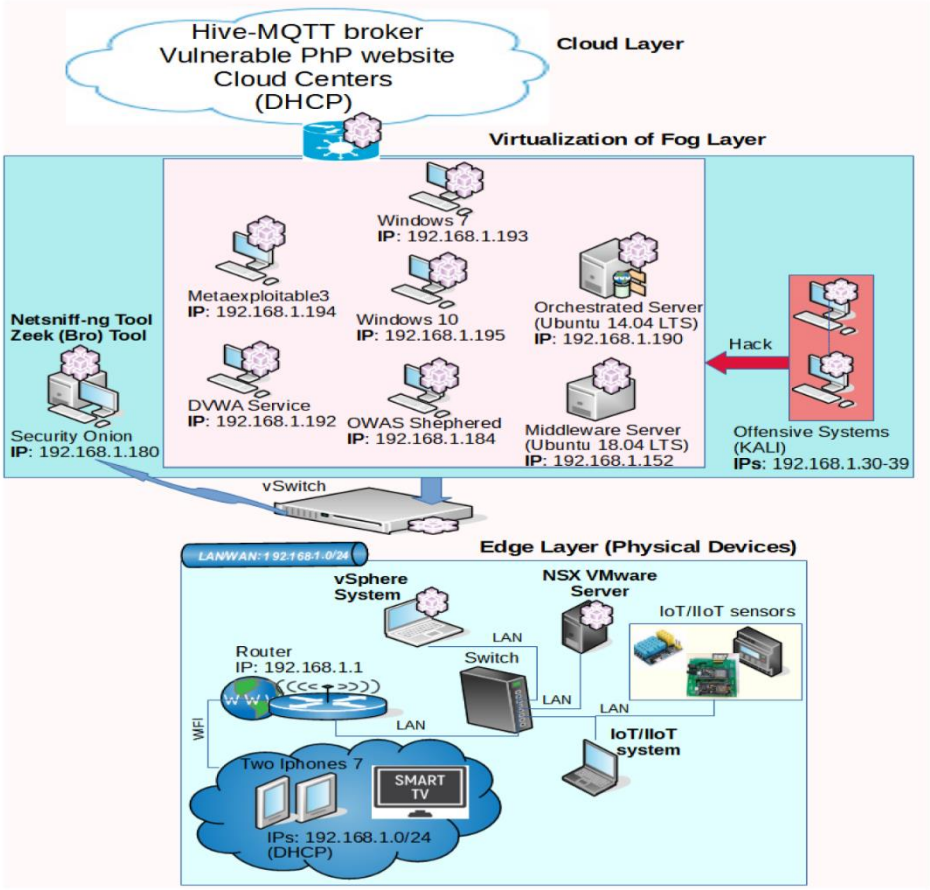

Figure 1: Configured Testbed of TON_IoT datasets for collecting network data

1.4 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]

- 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.

2 **Dataset Other factors to consider:** Threat vectors increases 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
- Dr Nour Moustafa (unsw.edu.au) and team.
- The TON_IoT Datasets | UNSW Research