TON_IoT Datasets for Cybersecurity Applications based Artificial Intelligence

The TON_IoT datasets are new generations of Internet of Things (IoT) and Industrial IoT (IIoT) datasets for evaluating the fidelity and efficiency of different cybersecurity applications based on Artificial Intelligence (AI). The datasets have been called 'ToN_IoT' as they include heterogeneous data sources collected from **T**elemetry datasets of IoT and IIoT sensors, **O**perating systems datasets of Windows 7 and 10 as well as Ubuntu 14 and 18 TLS and **N**etwork traffic datasets. The datasets were collected from a realistic and large-scale network designed at the IoT Lab of the UNSW Canberra Cyber, the School of Engineering and Information technology (SEIT), UNSW Canberra @ the Australian Defence Force Academy (ADFA). The datasets were gathered in a parallel processing to collect several normal and cyber-attack events from IoT networks. A new testbed was developed at the IoT lab to connect many virtual machine, physical systems, hacking platforms, cloud and fog platforms, IoT and IIoT sensors to mimic the complexity and scalability of industrial IoT and Industry 4.0 networks.

Different hacking techniques, such as DoS, DDoS and ransomware against, were launched against web applications, IoT gateways and computer systems across the IIoT network. The directories of the TON_IoT datasets include the following:

1. Raw datasets

1. **IoT/IIoT datasets** were logged in **log** and **CSV** files, where more than 10 IoT and IIoT sensors such as weather and Modbus sensors were used to capture their telemetry data.

   Links of open source tools used:
   - Node Red: https://nodered.org/
   - Modbus of Node Red: https://flows.nodered.org/node/node-red-contrib-modbus

2. **Network datasets** were collected in the packet capture (**pcap**) formats, **log** files and **CSV** files of the Bro tool.

   Links of open source tools used:
   - Security Onion: https://securityonion.net/
   - Kali Linux: https://www.kali.org/
   - Bro (recently named ZEEK): https://www.zeek.org/
   - Wireshark: https://www.wireshark.org/

3. **Linux datasets** were collected by running a tracing tool on Ubuntu 14 and 18 systems, especially atop, for logging desk, process, processor, memory and network activities. The data were logged in **TXT** and **CSV** files.

   Links of open source tools used:
   - netsniff-ng: http://netsniff-ng.org/

- atop: https://linux.die.net/man/1/atop

4. **Windows datasets** were captured by executing dataset collectors of the Performance Monitor Tool on Windows 7 and 10 systems. The raw datasets were collected in a **blg** format opened by Performance Monitor Tool to collect activities of desk, process, processor, memory and network activities in a **CSV** format.

   Link of open source tool used:
   - Windows Performance Monitor: https://techcommunity.microsoft.com/t5/Ask-The-Performance-Team/Windows-Performance-Monitor-Overview/ba-p/375481

2. Processed datasets

- The four datasets were filtered to generate standard features and their label. The entire datasets were processed and filtered in the format of **CSV** files to be used at any platform. The new generated features of the four datasets were described in the '**Description_stats_datasets**' folder, and the number of records including normal and attack types is also demonstrated in this folder.

3. Train_Test_datasets

- This folder involves samples of the four datasets in a CSV format that were selected for evaluating the fidelity and efficiency of new cyber security application-based AI and machine learning algorithms. The number of records including normal and attack types for training and testing the algorithms are listed in the '**Description_stats_datasets**' folder.

4. Description_stats_datasets

- This folder includes the description of the features of the four processed dataset (the folder of processed datasets) and the statistics (i.e., the number of rows of normal and attack types).

5. SecurityEvents_GroundTruth_datasets

- This folder includes the security events of hacking happened in the four datasets and their timestamp (ts). The datasets were labelled based on tagging IP addresses (192.168.159.30-39) and their timestamps in the four datasets. These IP addressed were used for Kali Linux systems to launch and exploit the systems of the four environments of IoT/IoT systems such as Cloud gateways, MQTT protocols, web applications of Node Red, Linux, Windows and network services.

The datasets can be used for validating and testing various Cybersecurity applications-based AI such as intrusion detection systems, threat intelligence,

malware detection, fraud detection, privacy-preservation, digital forensics, adversarial machine learning, threat hunting. The dataset was sponsored by the Australian Research Data Commons (ARDC) and UNSW Canberra.

Free use of the TON datasets for academic research purposes is hereby granted in perpetuity. Use for commercial purposes is allowable after asking the author, Dr Nour Moustafa, who has asserted his right under the Copyright.

For more information about the datasets, please contact the author, Dr Nour Moustafa, on his email: nour.moustafa@unsw.edu.au or eng.nourmosuatafa@hotmail.com.

More information about Dr Nour Moustafa is available at:

- https://www.unsw.adfa.edu.au/our-people/dr-nour-moustafa
- https://research.unsw.edu.au/people/dr-nour-moustafa-abdelhameed-moustafa
- https://www.linkedin.com/in/nour-moustafa-0a7a7859/