New Generations of Internet of Things Datasets for Cybersecurity Applications based Machine Learning: TON_IoT Datasets

Nour Moustafa

University of New South Wales, Canberra, Australia, nour.moustafa@unsw.edu.au

EXECUTIVE SUMMARY

Collecting and analysing heterogeneous data sources from the Internet of Things (IoT) and Industrial IoT (IIoT) are essential for training and validating the fidelity of cybersecurity applications-based machine learning. However, the analysis of those data sources is still a big challenge for reducing high dimensional space and selecting important features and observations from different data sources. The study proposes a new testbed for an IIoT network that was utilised for creating new datasets called TON_IoT that collected Telemetry data, Operating systems data and Network data. The testbed is deployed using multiple virtual machines including hosts of windows, Linux and Kali Linux operating systems to manage the interconnections between the three layers of IIoT, Cloud and Edge/Fog systems. The initial statistical evaluation of the datasets reveals their capability for evaluating cybersecurity applications such as intrusion detection, threat intelligence, adversarial machine learning and privacy-preserving models.

RESEARCH OBJECTIVES

The research attempts to provide a solution to the question, what role does standardisation play in a transformative data collection?. To address this question, there is another question should be solved, what are standard platforms and methods could be utilised to collect and analyse heterogeneous IoT data?, which is urgent to address for enhancing the sustainability of significant data collections that will drive innovation in the Cybersecurity research in Australia. The project aims at proposing standard methods and IoT and IIoT testbeds that will be developed at the cyber range labs, UNSW Canberra for collecting and analysing heterogeneous IoT datasets. The methods and datasets will be publicly published at the UNSW web portal and its cloud storage, as in our benchmark datasets, the UNSW-NB15 [1] and the Bot-IoT [2].

The new testbeds include a broad range of vulnerable and exploited platforms, industrial IoT sensors and services connected to public IoT hubs to ensure mimicking realistic IIoT networks. Four heterogonous data sources are in-parallel collected from telemetry data of IoT systems, data of Windows and Linux systems and their network traffic. The datasets contain a wide range of new attack surfaces and vectors, as well as legitimate events. The datasets and their analysis would improve the validations of different cybersecurity applications based on statistical models, machine/deep learning models, and riching data assets of cybersecurity and IoT applications in Australia and at the globe.

RESEARCH METHODOLOGY

The research methodology includes three main phases: 1) extending the testbed of IoT network at the Cyber Range labs at UNSW Canberra; 2) collecting and filtering heterogeneous datasets; and 3) initial evaluation of datasets using deep learning models, as briefly explained below.

1) Extending the testbed of IoT network at the Cyber Range labs at UNSW Canberra

In the Cyber Range Labs of UNSW Canberra, a testbed network for the industry 4.0 network that includes IoT and IIoT devices and services has been designed, as published in the papers [1-4]. The testbed will be extended to generate a new systematic testbed of IIoT networks for creating new realistic datasets, as presented in Figure 1. The testbed is deployed using multiple virtual machines and hosts of windows, Linux and Kali Linux operating systems to manage the interconnection between the three layers of IoT, Cloud and Edge/Fog systems. A set of IoT devices and sensors, such as green gas IoT and industrial IoT actuators, is connected to MQTT gateways to publish and subscribe to various topics, such as measuring temperature and humidity.

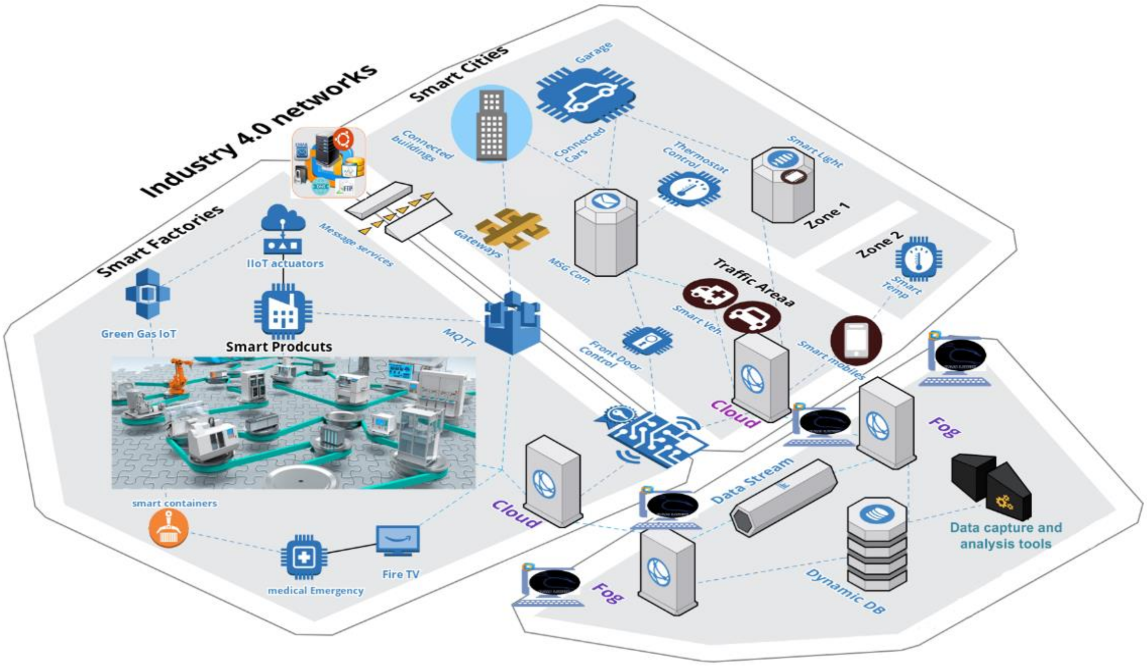

Figure 1: An architectural design for generating datasets from the Industry 4.0/IIoT networks

2) Collecting and analysing heterogeneous datasets

From the designed testbed network, there are four heterogonous data sources collected from telemetry data of IoT systems, data of Windows and Linux Ubuntu systems and their network traffic. The datasets contain a wide range of new attack surfaces and vectors, as well as legitimate events. For analysing the datasets, existing and new tools are utilised to extract multiple features for evaluating the efficiency of the datasets for validating cyber applications and improving big data analytics tools.

3) Initial evaluation of datasets using deep learning models

The datasets have diverse patterns and large-scale events to assess different cyber applications-based learning models such as intrusion detection, privacy-preserving, and digital forensics systems. Deep learning and statistical algorithms can be used for evaluating the new datasets compared with current benchmark network and IoT datasets.

CONCLUSION

This research proposes new testbeds for IoT networks, datasets and their analysis for validating cybersecurity applications. Further investigation and evaluations of the datasets will be publicly published at the UNSW website.

REFERENCES

1. Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In2015 military communications and information systems conference (MilCIS) 2015 Nov 10 (pp. 1-6). IEEE.
2. Koroniotis N, Moustafa N, Sitnikova E, Turnbull B. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. Future Generation Computer Systems. 2019 May 22.
3. Moustafa N, Adi E, Turnbull B, Hu J. A New Threat Intelligence Scheme for Safeguarding Industry 4.0 Systems. IEEE Access. 2018;6:32910-24.
4. Moustafa N, Creech G, Sitnikova E, Keshk M. Collaborative anomaly detection framework for handling big data of cloud computing. In2017 Military Communications and Information Systems Conference (MilCIS) 2017 (pp. 1-6). IEEE.