Data and Services Discovery projects - Transformative Data Collections

Title

A Pilot Database of Industrial Internet of Things Networks for Cyber Security Applications

Approach

This project includes three key methodological phases: 1) extending the testbed of IoT network at the Cyber Range Labs of UNSW Canberra; 2) collecting and filtering heterogeneous datasets; and 3) initial evaluation of datasets using statistical and deep learning models. The three phases are separately described as follows:

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

In the Cyber Range and IoT Labs of UNSW Canberra, a testbed has been designed for Industrial Internet of Things (IIoT)/ industry 4.0 network which includes IoT services and network elements. The testbed has been extended to include many IoT services and security events for creating new realistic datasets, as presented in Figure 1. The testbed was 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 IIoT thermostat and Modbus sensors, was connected to MQTT gateways to publish and subscribe to various topics, such as measuring temperature and Modbus register values.

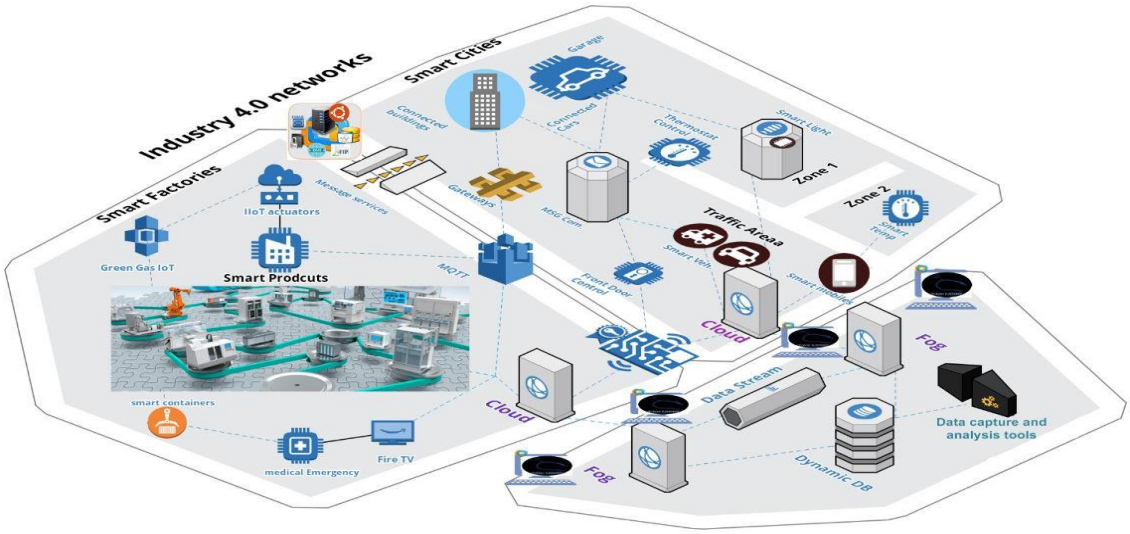

Figure 1: An architectural design for generating datasets from the Industry 4.0/IIoT network

2) Collecting and analysing heterogeneous datasets

From the testbed network, there are four heterogeneous data sources collected from telemetry data of IIoT 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 have been utilised to extract multiple features for evaluating the efficiency of the datasets for validating cyber applications-based machine learning and improving big data analytics tools. The new datasets and tools used have been explained in the CloudStor.

3) Initial evaluation of datasets using statistics and deep learning models

The datasets have been pre-processed and filtered to evaluate new cybersecurity applications. 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. Statistical and deep learning and algorithms have been used for evaluating the new datasets compared with current benchmark network and IoT datasets. **The results have revealed that the datasets have a broad range of normal and attack observations that can be utilised for evaluating the fidelity of several cyber applications better than the state-of-the-art datasets.**

Participants

Participants from Data61 CSIRO, Cyber Security CRC (CSCRC), Australian Federal Police (AFP), UNSW Sydney, RMIT, and University of Texas at San Antonio in the USA, were involved and consulted during the collecting of the new datasets. The participants have advised to consider generating data sources for various cyber security problems. More specifically, the participants of Data61 and RMIT assisted in collecting data sources that will be used to solve privacy-preserving challenges, as they guided to collect heterogeneous features with different data types from the four new datasets. The participants of CSCRC, AFP and UNSW Sydney supported the project team in determining how the layers of IoT, Fog and Cloud systems should be linked to each other. They also gave feedback to include multiple IoT services to determine the validation of threat intelligence models in smart systems. The USA collaborator provided some more insights to collect features that can be used to train and validate digital forensic mechanisms-based machine learning algorithms. The participants also reviewed the outcomes of the datasets and they committed to using the datasets in their research community.

Outputs

Heterogeneous Datasets

This project has produced heterogeneous datasets that have a broad range of normal and attack patterns that can be utilised for evaluating the fidelity of several cyber applications better than the state-of-the-art datasets.

The datasets have been described in the UNSW ResData catalogue and harvested into Research Data Australia at https://doi.org/10.26190/5d7ac9bfe8487 to sustain its public availability.

The datasets have been publicly published to a cloudstor space, within which there is a file named '*ReadMe.pdf*' that describes the folders and files which contain the raw and filtered datasets and their description.

Conference Paper

There is also a conference paper that has been accepted as an oral presentation at the eResearch 2019 to describe the new features of the four datasets.

> Nour Moustafa, "*New Generations of Internet of Things Datasets for Cybersecurity Applications: TON_IoT Datasets*", eResearch AustraliaAsia 2019, Brisbane, Australia.

Output Datasets combined with Existing Datasets

The new datasets, so-called TON_IoT, have been integrated with our existing datasets, UNSW_NB15 and Bot-IoT, that have been widely used in academia and industry, as published in the following outputs:

- Nour Moustafa and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015. https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
- Nickolaos Koroniotis, Nour Moustafa, Elena Sitnikova, and Benjamin Turnbull. "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset." *Future Generation Computer Systems* 100 (2019): 779-796. https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/bot_iot.php

FAIR

The output Datasets have had FAIR principles applied during this project, so that they may further benefit research and industry developments in cyber security, in Australia and globally.

The FAIR assessment spreadsheet has been attached. All the FAIR assessment principles have been successfully considered in the spreadsheet based on the guidance and feedback of the ARDC contact, Melanie Barlow.

Collaboration and coverage

There are multiple contributors and national collaborators who worked together to create the new datasets to the Cyber Security research and industry community. The contributors from UNSW Canberra generated and processed the datasets from a realistic testbed of IIoT network installed at the Cyber Range and IoT Labs of UNSW Canberra. Several national and international collaborators engaged with the contributors to provide tangible contributions to create the datasets for evaluating new Cyber Security applications-based machine learning. The collaborators of Data61 and RMIT added new features that will be utilised for solving privacy-preserving and differential privacy problems. The collaborators of CSCRC, AFP and UNSW described new services that should be included in the IIoT network of the testbed. The USA collaborator guided to add new features, which could assist in validating digital forensics models based machine learning.

The datasets that we produced between 2015 and 2019 have been widely used and cited by several academic and industry institutions. More importantly, the UNSW-NB15 dataset has become a benchmark for validating intrusion detection systems, as the anomaly detection models of Oracle and Microsoft based deep learning systems were validated using this dataset. This is an excellent indicator that the new datasets will also be used by developers and cyber security researchers in Australia and throughout the world.

Sustainability

The new datasets and their metadata have been publicly published through the sustained UNSW ResData catalogue, hence Research Data Australia, and CloudStor, as listed in the following:

- ResData is used for sustainably storing the datasets, as it is the online catalogue of UNSW datasets and collections of research materials. It records UNSW research data, where records are sustainably published to the online portal Research Data Australia (RDA).
- CloudStor is also used for storing the datasets because it is designed to meet the specific needs of researchers, and one terabyte of storage is available free to each individual researcher at AARNet-connected institutions. It sustainably provides quick and secure data transfer with no file size restrictions.

The dataset metadata from UNSW ResData has been harvested into Research Data Australia at https://doi.org/10.26190/5d7ac9bfe8487 and the record contains a link to the data on CloudStor at https://cloudstor.aarnet.edu.au/plus/s/ds5zW91vdgjEj9i

The datasets have been stored in files of a Giga-byte size at maximum to assert the download persistence at any Internet speed. The UNSW ICT department has a regular update to the datasets and keeps all UNSW datasets available. The department has a backup plan to ensure that the UNSW public datasets will be available all the time to support researchers and developers for easily downloading the datasets at anywhere and anytime.

Learnings

There are many key lessons learned in bringing together a Transformative Data Collection (TDC) in the Cyber Security and Internet of Things domains, as explained in the following:

- Development of many IoT services and integration of them into cloud and network systems are one of the issues of scalability and operability. This issue has been handled by consulting the collaborators who guided the project team in managing IoT sensors and network elements into separate layers of edge/fog, cloud and network.
- Collection of heterogeneous data sources, along with ensuring the correctness of security events, demanded the integration of data science and network security skills. The lesson learned is handling the challenge of collecting structured and unstructured data sources and processing their high dimensional space in real-time.
- Launching hacking scenarios to multiple systems of IoT, network, Windows and Linux is an arduous task because a cyber-attack that could exploit a system vulnerability is different at every system involved in the testbed network in most cases. The challenge of this task has been tackled by

applying a standard cyber threat framework for various systems. A cyber-kill chain model was utilised to making homogenous vulnerabilities for breaching them by homogenous exploits.

The ARDC should provide a suitable time-frame for the short projects in future to make the process easier for growing other Transformative Collections. The actual time frame of this project should be at least a year to provide detailed analysis and offer detailed problems and solutions to the cybersecurity community.

There are two reasons for successfully completing this project on time. First, UNSW Canberra allowed using another fund scheme until signing the project contract and transferring the fund to UNSW Canberra. Using the other fund scheme enabled us to hire two casual academics to work with Dr Nour Moustafa for completing the project on time. Second, the ARDC team, in particular, Melanie Barlow, Data Consultant and the ARDC contact of the project and Elizabeth Lopes, Senior Project Officer, have been supportive. Melanie Barlow has helped us in every stage of the project and has provided feedback and comments to ensure the successful deliverables of the project. **We strongly thank Melanie Barlow.**

Impact

The raw and processed datasets and their tools of collection and analysis have been publicly published to enable developers and researchers in other domains, such as data science and general machine learning applications, for using the datasets. The datasets will be Australian benchmark data assets for cyber applications-based Artificial Intelligence that could significantly enhance the Australian cyber security ecosystem. This research outcomes of the project have a great impact as explained in the following achievements:

Research Publications

The outcomes of the project have been partially published in the following three outputs:

1. Moustafa, Nour, "*New Generations of Internet of Things Datasets for Cybersecurity Applications: TON_IoT Datasets*", eResearch AUSTRALASIA2019, Brisbane, Australia. https://conference.eresearch.edu.au/wp-content/uploads/2019/08/2019_eResearch_59_New-Generations-of-Internet-of-Things-Datasets-for-Cybersecurity.pdf
2. Moustafa, Nour. "A Systemic IoT-Fog-Cloud Architecture for Big-Data Analytics and Cyber Security Systems: A Review of Fog Computing." *arXiv preprint arXiv:1906.01055* (2019). https://arxiv.org/abs/1906.01055
3. AlKadi, Osama, Nour Moustafa, Benjamin Turnbull, and Kim-Kwang Raymond Choo. "Mixture Localization-Based Outliers Models for securing Data Migration in Cloud Centers." *IEEE Access* 7: 114607-114618, **DOI:** 10.1109/ACCESS.2019.2935142 (2019).

The new TON_IoT datasets and existing datasets, UNSW_NB15 and Bot-IoT, have been widely used in academia and industry, and they have become benchmark datasets for evaluating intrusion detection systems. In academic venues throughout the world, the datasets have been totally cited more than 460 times, as published at Dr Moustafa's Google Scholar Profile (05/09/2019). In Industry, the datasets have been used to validate real-world Cyber Security applications of Oracle and Microsoft.

Research Grants

This project has helped the project team to win and submit other research grants related to the design of IoT testbed and cybersecurity applications based Artificial Intelligence, as listed in the following:

- Nour Moustafa (Lead Investigator), 'Threat Intelligence tool based Artificial Intelligence for Smart Airports', funded by Cybersecurity Collaborative Research Center (CSCRC), Australian Federal Police, Data61, and Australian Cyber Security Center (Salary of Senior Research Associate for two years) (**won**)
- Nour Moustafa (Co-investigator), UNSW Infrastructure Grant for purchasing new IoT infrastructure to the IoT Lab of UNSW Canberra (In progress, AU$100K)
- Nour Moustafa (Lead Investigator), ' Development of Platform as a Service for Cyber Security Applications based Artificial Intelligence', (will submit EOI for Platforms 2019 Call of the ARDC)

New Research Areas and Approaches

The outcomes of the project have identified a new research area, so-called 'Intelligent Security', which demonstrates the interconnection between Cybersecurity, Industrial IoT and Artificial Intelligence fields. The Intelligent Security area investigates new capabilities of Artificial Intelligence approaches to address current cybersecurity problems in the environments of network, IoT, IIoT, cloud and fog. The full description of this area has been written by Dr Nour Moustafa and it will be released in the report of UNSW Capability Statement 2020. This project also allows to Dr Nour Moustafa to become the Lead of Offensive Security Theme at UNSW Canberra Cyber.

Broader Anticipated Impact

The research outcomes of the project will benefit the cybersecurity research and industry community such as academic institutions and companies of Microsoft and Oracle. The generated datasets will be used for training and validating new cybersecurity applications-based Artificial Intelligence for measuring their fidelity in research and real-world cybersecurity systems. These datasets will be an excellent value to the Australian data assets in the cybersecurity ecosystem. For the first time in this domain, there are new features of IIoT services have been created in the datasets to discover new cyber threats. This will help in testing new security systems for protecting Australian critical infrastructure and saving a lot of money that Australia loses yearly due to cyber threats.

In the upcoming two years, the future pathway of the project will be developing a Platform as a Service for cybersecurity applications based Artificial Intelligence. The platform will include open-source techniques tools that allow researchers and developers to generate realistic datasets in real-time. Moreover, the platform will include open-source applications of statistics and deep learning techniques for developing new generations of security systems in Australia that could enrich the Australian cybersecurity ecosystem.

Report prepared by: Dr Nour Moustafa and University of New South Wales at Canberra
Date: 03/09/2019