Cybersecurity in SMEs: The Smart-Home/Office Use Case

1st Nikolaos Vakakis
CERTH-ITI
Thessaloniki, Greece
nikovaka@iti.gr

2nd Odysseas Nikolis
CERTH-ITI
Thessaloniki, Greece
odynik@iti.gr

3rd Dimosthenis Ioannidis
CERTH-ITI
Thessaloniki, Greece
djoannid@iti.gr

4th Kostantinos Votis
CERTH-ITI
Thessaloniki, Greece
kvotis@iti.gr

5th Dimitrios Tzovaras
CERTH-ITI
Thessaloniki, Greece
tzovaras@iti.gr

***Abstract*—Today, Small and medium-sized enterprises (SME) can be considered as the new big target for cyber attacks, while the cybercrime prevention is often neglected within their environment. This paper aims to investigate the characteristics of cybersecurity threats in the Digital Innovation Hub (DIH) ecosystem of a Smart-Home/Office environment being constituted by Small and Medium Enterprises (SMEs) that contains various smart-devices and IoT equipment, smart-grid components, employees workstations and medium size networking equipment.. As the Cyber-security in such an ecosystem is really demanding and challenging because of the various communication layers and the different supported IoT devices, we introduce a more robust, resilient and effective cybersecurity solution that can be effortlessly tailored to each individual enterprises evolving needs and can also speedily adapt/respond to the changing cyber threat landscape. Thus, this Cyber-security framework will be evaluated through three major types of Smart-Home/Office data-sets and will be supported from SME/ICT clusters under the framework of the Secure and Private Smart Grid (SPEAR) H2020 project. The first promising results of our work indicate the potential of implementing strong defence mechanisms for SMEs environments within DIHs.**

***Index Terms*—cybersecurity, anomaly detection, SME, dataset, Smart-Home/Office environment, SPEAR,**

I. Introduction

Cybersecurity in SMEs is an ever growing concern due to the increasing adoption of digital technologies like Cloud, Computing and Internet of Things (IoT)but also the constitution of a wide range of devices (e.g. PCs, servers, mobile devices, etc.) and business practices (e.g. BYOD, remote access, use of cloud-based apps and services, etc.). In this context, the increased connectivity that comes with IoT technologies entails also bigger attack surface for adversaries, by leveraging security gaps of interconnected IoT devices within a smart home/office environment. One the one hand providing secure protocols can protect organization data from being eavesdropped, as well as from unauthorized access to their systems or even denial of service attacks and save them from security breaches that could cost a lot in terms of capital or customer dissatisfaction, but one the other hand security imposes slower communications, thus it is important to find the right balance between those goals. Therefore it can be understood that it is crucial to have security mechanisms that can detect attacks not only in the network layer but also on the application layer and compensate for any protocol vulnerabilities.

Recent studies show not only that the percentages of small or medium size businesses that suffer from cyber attacks are quite high, but also that SMEs appear to have significant perception gap when it comes to cyber awareness and preparedness. Similarly, according to one FireEye claim 77% of all cyber-crimes target SMEs. Simple endpoint protection through antivirus has become by far inadequate due to the complexity and variety of cyber threats as well as the integration of a wealth of digital technologies in business processes even of the smallest enterprises.

The fact that SMEs represent the vast majority of total businesses and that hackers consider them as easier target than large enterprises, reinforce the need for new robust security mechanisms that achieve early and accurate detection of attack patterns. The SPEAR H2020 project aims to develop tools that enhance the security by design and introducing cyber security prevention mechanisms in the area of smard grids, including also smart homes (like the smart home of CERTH in Thessaloniki which operates both as a living and working environment). This dual nature renders the smart home an appropriate testbed for applying advanced anomaly detection methodologies and tools with data deriving from a real environment that can be a powerful weapon for SMEs security administrators support.

The discussion on this paper will proceed as follows:

Section 2 describes the derived from the smart home environment, different kinds of datasets that have been collected efficiently, while in section 3 we analyse how anomaly detection techniques could be applied in these data sets. Thus we present some advanced methods based deep learning techniques, while in section 4 the first results obtained from the previously defined models are presented. Finally in section 5 we conclude with the future discussion and work.

II. Datasets

There are three kinds of datasets produced by the smart home environment, netflows dataset concerning the network traffic of the smart home, application layer protocol dataset concerning the application layer protocols that are used by the smart home equipment's communications and electricity measurements dataset stemming from the measurements collected by smart meters.

N. Vakakis, O. Nikolis, D. Ioannidis, K. Votis, and D. Tzovaras, "Cybersecurity in SMEs: The Smart-Home/Office Use Case," in 2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2019, doi: 10.1109/camad.2019.8858471.

A. Netflows dataset

The netflows dataset is produced by monitoring the smart home network traffic using Wireshark tool. After capturing the traffic in the form of pcap files, Cicflowmeter network traffic generator and analyzer tool is used to generate csv files with 84 netflow features like source IP, source port, destination IP, destination port, flow duration, total forward packets, total length of forward packets, packets per second, bytes per second and many more. In the produced dataset, there is also a column that can be used for manually labelling a netflow as normal or abnormal. Based on those features, feature extraction or dimensionality reduction methods can be applied to extract the most representative information about the netflows and apply state of the art machine learning techniques to train classifiers.

B. Application Layer Protocols dataset

Detecting attacks on network layer is the first step, but is not sufficient on its own, because many attacks known as layer 7 or application layer attacks target the application layer. The smart home use case includes smart devices that utilize three very common application layer protocols i.e. BACnet for the smart home's Heating HVAC system, MQTT as a message broker for sensor measurements and MODBUS for the smart meters. The second dataset is derived from the network traffic capture of the smart home by utilizing t-shark a network protocol analyzer that supports extracting protocol features from pcap files. T-shark filters include a wide variety of features for many protocols making it possible to dissect the protocols' packets and extract the information needed to train machine learning models that learn the common traffic of a specific environment in terms of packets and exchanged messages. Some indicative features for the protocols used in smart home use case are listed below:

MQTT:

- mqtt.clientid that specifies the id of the client that wants to establish a connection with the MQTT broker
- mqtt.hdrflags that holds information relevant to the MQTT control packet type
- mqtt.conflags that contains parameters specifying the behavior of the MQTT connection. It denotes the presence or absence of fields in the payload
- mqtt.len that specifies the length of an MQTT publish message

BACnet:

- bacapp.confirmed _service that indicates the type of application service choice in a service request message
- bacapp.type which defines the type of the Application Protocol Data Unit (APDU) message type and the fields that appear

MODBUS:

- mbtcp.trans_id to synchronize communication between devices
- mbtcp.len that identifies the remaining length of the packet
- modbus.func_code that identifies the function to execute
- mbtcp.modbus.data which are the actual trasmited data

For the formation of the application layer dataset in the smart home use case, all the protocol features related to the communications of the smart home equipment are collected and some of them are used for the anomaly detection methods that will be presented in section 3.

C. Electricity Measurements dataset

The third and final type of dataset consists of electricity measurements that are collected by the smart meters related to the smart devices supporting the smart grid functionalities. These are measurements like total active, reactive and apparent power, phase voltage, phase amperage and more. The collection of those measurements for a certain time frame results to time series of measurements that are used to train sequence to sequence LSTM neural networks that are able to learn sequences of data and effectively reconstruct them.

III. Anomaly Detection Methods

In this section two initial anomaly detection methods will be presented. The first concerns application layer anomaly detection and focuses on MQTT protocol, but can be adjusted to other protocols as well and the second method is related to electricity measurements time series anomaly detection. Both methods utilize LSTM neural networks, for the reason that this type of neural networks exhibit remarkable results in learning sequences of data containing long term patterns by efficiently modeling complex multivariate sequences.

A. Stacked LSTM neural network for application layer anomaly detection

The first is an unsupervised learning method which implements a stacked LSTM neural network topology, that has the ability to learn normal data and detect any unusual data it receives. More specifically the network is being continuously trained with normal streaming data consisting of two MQTT protocol features, message type and message length. The goal of this method is to learn the patterns of MQTT message sequences that appear in the smart home network considering their type and length. If at some point in time the network receives as training input an unusal sequence of message types or message lengths which could indicate an infiltration in the network or a Denial of Service (DoS) attack, then the training loss is expected to exhibit a local peak. After experimenting a neural network with six LSTM layers with ten units (fig 1) each and a time distributed output layer seemed to produce the most satisfying results (fig 2). Furthermore, Adam the state of the art optimization algorithm for the adoption of the learning rate was used. The network receives in its input layer ten training samples with two features each and each layer

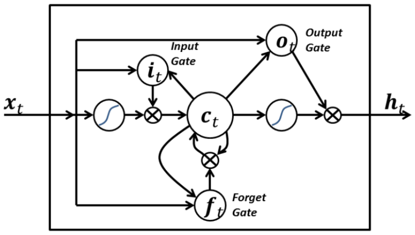

Fig. 1. LSTM cell

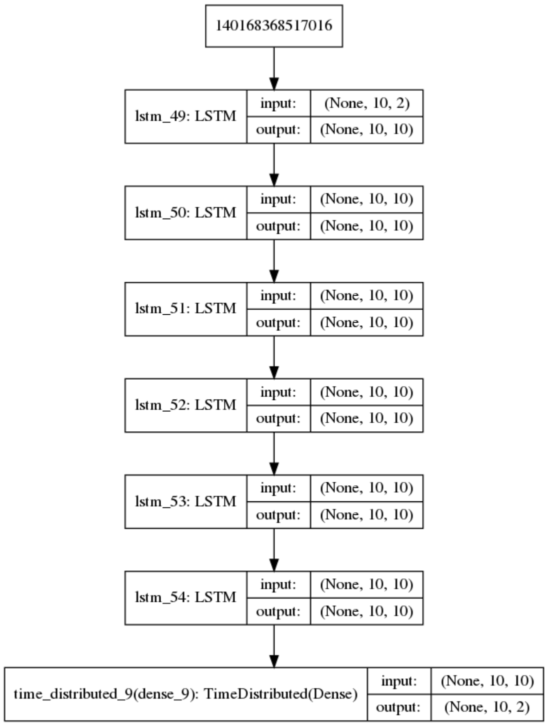

Fig. 2. stacked LSTM model

passes one hidden state output for each input time step to the next layer. Finally the time distributed layer applies a dense layer for each input and thus outputs one time step from the sequence for each time step in the input.

B. Seq2seq LSTM encoder-decoder electricity measurements anomaly detection

The second model is a self-supervised sequence to sequence LSTM network in encoder-decoder format. The aim of a sequence to sequence model is to map a fixed length input with a fixed length output where the length of the input and output may differ. The encoder part of the network produces a representation of the time series data it receives as input in the form of a vector which then the decoder part receives as input and tries to reconstruct the initial data. The specific architecture of the model (fig 3) that after experimenting achieved the best results with the smart home electricity measurements dataset consists of an encoder with 2 LSTM layers, a decoder also with 2 LSTM layers and a dense output layer. The proposed model is trained on manually annotated normal day sequence samples and is tested on mixed normal and abnormal data. Data are fed to the network in 3D format (100, 96, 1), which corresponds to 100 days, where each day consists of 96 samples and each sample represents a single measurement such as active power or reactive power. The detection is based on a user-specified threshold on the Mean Squared Error (MSE) between the input and the reconstructed output of the decoder.

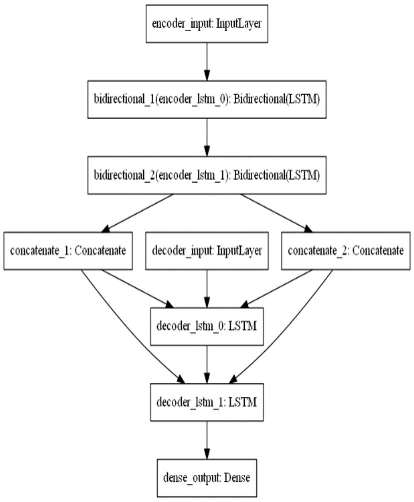
encoder_input: InputLayer
bidirectional_1(encoder_lstm_0): Bidirectional(LSTM)
bidirectional_2(encoder_lstm_1): Bidirectional(LSTM)
concatenate_1: Concatenate
decoder_input: InputLayer
concatenate_2: Concatenate
decoder_lstm_0: LSTM
decoder_lstm_1: LSTM
dense_output: Dense

Fig. 3. seq2seq LSTM encoder-decoder

IV. Initial Results

The first results obtained from the previously defined models are quite promising and pave the way for further improvement of the models.

A. Application Layer anomaly detection results

The stacked LSTM neural network was tested for its efficacy by using normal MQTT protocol data captured from the smart home network traffic and synthetic abnormal data which were produced by randomly shuffling the sequence of a portion of the normal data. More specifically the initial dataset which consisted of approximately 10000 messages, was split to chunks of 500 messages that were fed consequently to the neural network in the 3D form (50, 10 ,2), which means 50 batches of 10 timesteps each, where each timestep corresponds to a message with 2 features, message type and message length. At some point a chunk of 500 shuffled messages was intepolated and then again the normal data continued as before. As can be observed in figure 4 after some time the neural network fits to the data that it considers as normal and the training loss decreases to very low levels, but when the shuffled data are fed to the neural network a peak at the loss is observed which means that the network received data that has never faced before. After that point the training loss falls again to normal levels due the inserted normal data. Furthermore the sensitivity of the model was tested by forming a second synthetic batch of abnormal data, this time only with 50 from

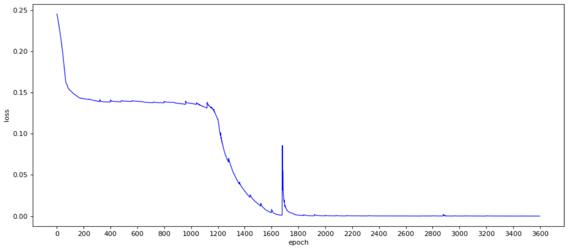

Fig. 4. Loss diagram indicating abnormal traffic

500 messages shuffled. Same as before a peak was detected in the training loss, but this time smaller thus proving that the model can efficiently detect even smaller changes in the network.

B. *Electricity Measurements anomaly detection results*

The training set for the Seq2Seq LSTM model is a manually annotated dataset of more than 100 normal days. The test set consists of the rest of the days which are a mix of normal and abnormal data. Figure 5 presents with red the real data of a normal day and with blue the prediction outcome of the trained model. The prediction follows the trend of a normal day. On the other hand, figure 6 presents an abnormal day. The anomaly is detected at the right side of figure 6 where the last samples do not follow the normal day pattern. Specifically, the measurements represent the total power consumed in a Smart-Home/Work environment and the abnormal event represents the power consumption during the last hours of the day. This event detection can be exploited for energy efficiency and cost reduction from the electricity bills.

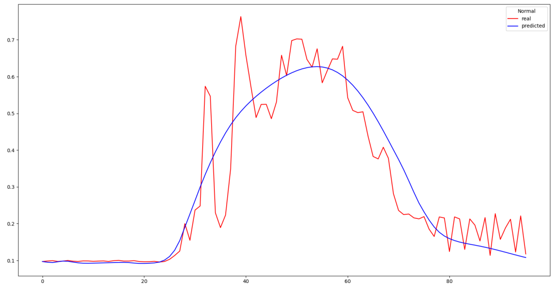

Fig. 5. Diagram with the predicted and the normal measurements

V. Conclusions and Future Work

Securing SMEs in the era of IoT is a very challenging but also extremely important task that cyber-security researchers need to focus on. Legacy signature based anomaly detection techniques are not sufficient anymore as rapid technology improvements lead to new attack schemes. In the context of SPEAR project the smart home use case offers the opportunity for implementing novel anomaly detection methodologies that could be used for other similar environments. The initial experimental results show the value of the datasets of the smart home, which constitute the solid basis upon which effective and robust anomaly detection models can be build. Those datasets offer a wealth of useful information that can be leveraged not only from the models presented in this paper but also from other kinds of predictive models. Furthermore, the two LSTM-based models can be even more improved by experimenting further with their hyper-parameters and that is the next step towards implementing a state-of-the art security mechanism that operates on several different levels and can efficiently detect different attack patterns.

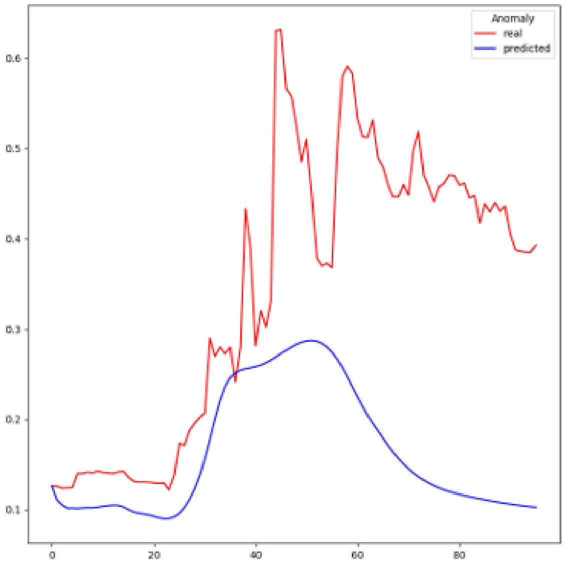

Fig. 6. Diagram with the predicted and the abnormal measurements

VI. Acknowledgement

The aforementioned work effort in this paper is conducted under the framework of the SPEAR project, a Horizon 2020 program, funded by the European Union under the grant agreement No. 787011.

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] Dalamagkas, Christos Sarigiannidis, Panagiotis Ioannidis, Dimosthenis Iturbe, Eider Nikolis, Odysseas Ramos, Franscisco Rios, Erkuden Sarigiannidis, Antonios Tzovaras, Dimitrios. (2019). A Survey On Honeypots, Honeynets And Their Applications On Smart Grid.

[3] H2020 - SPEAR - Official Website, https://www.spear2020.eu/.

[4] Seldon-Core, Github repo, https://github.com/SeldonIO/seldon-core/blob/master/components/outlier-detection/seq2seq-lstm/doc.md.