PSW (Hypothesis Formation)

What is the best ML algorithm to detect malware in real time network traffic so that cyber attacks to IoT devices can be spotted with 80% of accuracy; basing the detection on data in communication protocols, device category and malware types.

H

1 Context

IoT (Internet of Things) is an emerging topic of interest among research and industrial fields which has proven to have a significant impact on human life by the integration with smart devices used in healthcare, home security, lifestyle, city dynamics, automobilism and many others.

Malware is a software designed to cause damage to a computer, server or network and jeopardize the safety of the user, company, infrastructure or even a whole population. Malwares have the potential of producing damages of around $15 billion USD (like in 2003) once they successfully attack.

IoT devices can be misused as infrastructure to botnets, being the mean for one of the largest DDoS attacks registered in the US where Mirai malware infected around 500,000 IoT devices in 164 countries and almost took down the entire internet in 2016. There is no doubt that malware in IoT represents a major problem due to the amount of devices connected today and tomorrow (estimated 125 billion by 2030). However, malware infections leave a trace in their network traffic, which makes them identifiable and predictable at some extent by analysing devices' log files (basic principle of Security Information and event Management Systems).

Even though safety protocols and security mechanisms are quite advanced, cyber attacks are getting more sophisticated and popular. In particular, to enhance the level of security in IoT devices, it is necessary to use specific methods, datasets and tools to develop SIEM Systems. Therefore, being able to develop deployable technology in the form of algorithms, frameworks or even programs, could be extremely useful to get a better understanding of the behaviour of malware infections, IoT devices per se and their network traffic.

2 Criteria for Success

- **Implementing a machine learning algorithm that can successfully detect at least 80% of malicious traffic running in IoT devices.**
- **Additional to the binary classifier, implementing a type-of-malware detector**
- **Having a deployable production code to detect real time threats in real IoT devices.**

3 Scope of Solution Space

The analysis and modelling will be made at the IoT-23 Dataset provided by Avast AIC laboratory. It is planned to take the 3 benign scenarios along with other 4 malware captures to prevent class imbalance and cope with the computational resources necessary for this project.

The **machine learning detection algorithm** could be applicable to any SIEM integrated to devices such as: Amazon Echo, Philips Hue and Somfy Door Lock. However, any device that handle the same communication network protocols can be suitable for this implementation.

H D E I P

PSW (Hypothesis Formation)

What is the best ML algorithm to detect malware in real time network traffic so that cyber attacks to IoT devices can be spotted with 80% of accuracy; basing the detection on data in communication protocols, device category and malware types.

H

4 Constraints within solution space

- **Not having enough computational resources to get a proper analysis in all data.**
- **Information being outdated and not valid for new malwares.**
- **Not sufficient data on ports like Telnet that could represent a representative sample.**
- **Class imbalance considering the number of packets on each benign and malignant scenario.**

5 Stakeholders to provide key insight

Avast AIC Laboratory – Sebastian Garcia, Agustin Parmisano and Maria Jose Erquiaga

Springboard Mentor – Yadunath Gupta

Other Antivirus Companies – Bitdefender, Norton, McAfee, Panda and Kaspersky

6 Key data sources

IoT-23 Dataset provided by Avast AIC Lab – 20 malware and 3 benign captures (scenarios) executed in real IoT devices ranging from 2018-2019 in the form of *pcap* files (log files) captured by the Zeek network analyser. These files will have the network traffic and identification of the devices as well as communication protocols found

IoT-23 Dataset:

"Sebastian Garcia, Agustin Parmisano, & Maria Jose Erquiaga. (2020). IoT-23: A labeled dataset with malicious and benign IoT network traffic (Version 1.0.0) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4743746"

H D E I P