CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Behavioral Analysis and Detection of IoT malware using theIRC protocol**
Author's name:	**Bc. Ondřej Preněk**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Yury Kasimov
Reviewer's department:	External - Avast

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — **Above average challenging**

How demanding was the assigned project?

The topic of the thesis is relevant given the increasing number of IoT devices and as consequence the increased use of IoT devices in attacks. There are multiple challenges: setup of infrastructure, executing malware on IoT devices, capturing of traffic, manual inspection of what is captured, cleaning of normal IRC pcaps etc.
The research in the thesis requires understanding of network security, machine learning and to some extent of natural language processing.

Fulfilment of assignment — **A - Excellent**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

The thesis fulfills all the objectives defined in the assignment.

Methodology — **B - Very good**

Comment on the correctness of the approach and/or the solution methods.

The description of dataset creation is extensive and clear. The student uses the best practices for training and evaluation of the models. The code in notebooks is not clean, there is commented code, execution errors in some cells and requirements miss a couple of libraries which make reproducibility more complex.

Technical level — **B - Very good**

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

The student shows that he has a good overview of the machine learning field as well as network security field. Every step of the research is defined clearly which shows the student understands what he is doing very well. The used features are sound and they are well explained. The applied ML algorithms are suitable for this problem and they are widely used in the field. The proposed method is compared to well known detection systems and it outperforms them.
I would appreciate more extensive discussion about the results and why models perform differently in various scenarios. There is a mix up with confusion matrices in tables reporting results: for example, table 5.7 the number of False Positives for Random Forest is 1 and for xgboost it is 3 however the number of true negatives does not change.

Formal and language level, scope of thesis — **A - Excellent**

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

The thesis is very well structured, it is easy to follow and understand the main ideas. The thesis is written in English which is not the native language of Ondrej. There are a few typos which do not affect the overall intelligibility of the thesis.

Selection of sources, citation correctness — **A - Excellent**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?
The student reviews sources relevant to the topic of the thesis. The student's work is clearly distinguished from earlier work in the field.

Additional commentary and evaluation (optional) *Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.*

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading. Pose questions that should be answered during the presentation and defense of the student's work.

The thesis presents an approach to detect botnets which use IRC protocol. There is a lot of novelty in the thesis and I think it can be useful in practice. The student has created a module for a popular open source project *zeek*. The applied methodology is sound and the best practices are used. The work shows a good understanding of machine learning and network security by the student.

There are issues with confusion matrices in the reported results and there is an issue in the part 6.4.2 where results in the corresponding table do not correspond to the results in the text. The thesis is missing discussion about reasons behind model performance.

Questions:

- Is there benign IRC communication on IoT devices? Would it be enough to detect IRC communication used by an IoT device and classify it as malicious?
- Why do you think unsupervised methods perform so much worse compared to supervised?
- In figure 5.3 there are two malicious samples surrounded by benign samples. have you examined them? What is going on there?
- What could be possible reasons that random forest outperformed xgboost in many of the conducted experiments?

The grade that I award for the thesis is **B.**

Date: **29.08.2020** Signature: