ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Identifying Malicious Hosts by Aggregation of Partial Detections**
Author's name:	**Ondřej Lukáš**
Type of thesis :	bachelor
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Cybernetics
Thesis reviewer:	Luciano Bello
Reviewer's department:	IBM Thomas J. Watson Research Center

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — **challenging**

Evaluation of thesis difficulty of assignment.

The assignment is interestingly challenging. It provides the opportunity to develop a solid theoretical base while having practical impact. The practical output is very well measurable and impactful in a real existing tool. By asking to extend a tool instead of developing it from zero, the student had the chance the attack the most common kind of problem from real life infosec industry: "We had a problem. We developed a tool to solve it. The tool is an improvement but falls short somehow. Make it better."

Satisfaction of assignment — **fulfilled with major objections**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

We can divide the assignment in two aspects: The practical (the engineering effort to make the idea work) and the theoretical (the formal aspect of the thesis, in terms of explanation and scientific thoroughness). The student clearly and fully fulfilled the practical aspect of the assignment, i.e. to improve the quality of the detections in a given IPS. However, there is room for improvement in terms of the theoretical aspect of the assignment which are: review the existing literature, experimentally evaluate solutions, and critically analyzing the results. Given that the reporting on the second part has significant issues, I cannot confirm if the student fully understand the results or knows the reason of the practical success.

Method of conception — **correct**

Assess that student has chosen correct approach or solution methods.

From the purely result achievement perspective, the approach seems to work and improved the current implementation. Nonetheless, it is very hard to know if this is far or close to be the optimal solution, or even a local maximum. From the explanation in the thesis, there are many *out-of-the blue* design decisions. For example, why is the SDW size in 12 TWs? Is there any reason to suspect that one hour of history is better than two?

Similarly, there is no evaluation on the features listed in Sections 5.1 and 5.2. At first glance, some of the features look redundant (like (i), (ii), and (iii) in 5.1). Without any further discussion, I cannot assume the student is aware of *feature relevance extraction* techniques, such as the Chi-Squared test.

Again, the practical results show a significant increment in the IPS results. However, this positive outcome is overshadowed by the lack of well-supported justifications for the design decisions.

Technical level **C - good.**

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

If the hypothesis is "Can the history of the connection help to classify traffic as malicious?", the results match the intuition and confirms it. This is clearly new valuable knowledge and a great first step for further work.

Although, and as mentioned in *Method of conception*, it is hard to evaluate if the student considered alternatives (beyond comparing TW and SDW) as the best way to reach the results.

It seems that the student is on a good technical path but an outstanding solution is more than working code. It needs to be supported by well-understood results and detailed explanations.

Formal and language level, scope of thesis **E - sufficient.**

Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.

The main value of the thesis is its implementation. The formal aspects of it is the weakest aspect. The document seems unstructured and chaotic, probably finished in a rush. Since the thesis has most to improve in this regard, let me continue with a point-by-point explanation:

Structure:

It is hard to see the thread that runs through the text. Section 1 explains the problem in a very hermetic way, using lingo without definition. For example, the relation between *flows* and *connection* (both concepts mentioned in the introduction) is not clear until Section 4. Even further, the first paragraph mentions *web flows*, a concept that is fully ignored during the rest of the thesis.

It is not obvious from a first read that Section 3 is an explanation of the current SLIPS. It is easy to miss that the reason of that section is 3.4, i.e. explaining the current limitations. Section 4 is the core of the contribution and could have done a better job explaining key aspects, such as the fact that the system groups by source IP address. It is not until later that the reader is able to understand that source IP, host, and source address are synonyms. In general, it is better for the writer to stick with a single term when a concept is referred to. The choice of terms is also important. The name *Source Address Layer* invites to think that there are other layers that interact among them but no more layers are mentioned in the rest of the thesis. Subsection 5.4 looks like belonging to Related Work but is not. And Section 6 goes back to Section 4 in a very unexpected way. The subsection 8.1 is a collection of figures without any explanation (Subsection 9.1 makes very limited contributions in that respect). The explanation of Subsection 8.2 is done by Section 9.2 and others, which looks very disconnected.

Lack of explanation of the figures:

The thesis is full of figures with processes which are not self-explanatory and have no further explanation. The figures should assist an explanation. Especially when a workflow is explained, a figure can help explain how a stage interacts with another. However, the explanation of the stage needs to be done somewhere. Otherwise, the figure creates confusion instead of assistance. For example, the whitelisting check that is mentioned in several figures is not explained at all. If it is not relevant, why is it in the figure? If it is relevant, the student should explain it.

Subsections 9.1 explains that "SVM and XGBoost have similar results", although Figures 16 and 20 have different scales. Probably they are similar in some other aspect which is not explained.

Figure 1 looks very similar to Figure 11, although their differences create more questions than answers.

Figures 3 and 4 look redundant.

Figure 6 invites to think that there are two sets of Markov Models. Again, without any explanation, it is very hard to tell.

Lack of a single driving example:

Section 5 makes a good use of examples to explain concepts. However, the examples change unnecessarily often which is confusing. The connections in Figure 7 are not the connections in the subsequent Figures, even when they have the same name. The example in Figure 9 looks very similar to the one in Figure 10, with the exception of C3, with an extra flow V, which does not look relevant at all.

Unnecessary or confusing internal aspects:
Probably, the class structure in Figure 12 represent the implantation exactly. At the same time, it contains unnecessary details like the existence of a "IpDetectionAlert" class. Similarly, the full WHOIS information machinery seems irrelevant to the goal of the thesis.
Another internal aspect that looks irrelevant is the internal representation of the Markov chain states used by SLIPS (Figure 2). I understand that this format is used in the implementation, but in the explanation a more intuitive notation is probably better. For example, why not calling <SSize,SDur,SPeriod> the state "a"? The symbol "." could be replaced by a more illustrative <5s>.

Be careful with some claims:
Sentences like "The goal is to eliminate [SLIPS] limits" (Section 1) are very strong and should be avoided.
The existence of "models for most common types of malware" (Subsection 3.2) is an expression that is very hard to support without a reference.

Proof read before submitting:
The broken English makes the thesis very hard to read (e.g., some of the sentences do not have any verb). There are also multiple errors that can be caught by a spell checker (e.g. "... protocol ale combined ...", Section 4). A sentence in Subsection 6.1 finishes abruptly. In general, there is a lot of repetition, like in the first two paragraphs of Subsection 6.2.

Be consistent:
The student should introduce acronyms the first time and stick to them. If a particular capitalization or wording is used once, it should be used throughout the thesis if you are referring to the same concept. For example, "False Positive rate", "False positive ratio", "False positive rate", "FP rate" they all look like synonyms and are used in various part of the thesis, without a clear criterion.

Oher style issues:
- The time unit in Table 3 is missing.
- I think a clearer way to notate the range in the codomain of Equation 3 is [0,1]
- Sliding frame is a synonym of SDW?
- The footnote 1 should probably be a link to the RFC3912 instead of the Python package.
- There is a spacing issue in first paragraph of Section 10.

Selection of sources, citation correctness **D - satisfactory.**

Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.

The *Related Work* section can be improved in terms of explaining the existing literature and the relative positioning of the thesis in the use of machine learning for traffic classification. The cited sources could be linked with the presented work in more detail.

Additional commentary and evaluation

Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

The thesis made an interesting practical step forward but needs further work before it can be established. While the results are undoubtedly positive, they are clouded by a lack of detail in the explanations. In particular, a stronger experimental ground needs to be made to gain confidence on this good first step to improve SLIPS.

The extensive raw data attached to the thesis lack explanations and, because of that, deep understanding of the results cannot be assumed. From the scientific perspective, the maturity of the works seems to be suboptimal.

REVIEWER'S OPINION OF FINAL THESIS

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

I evaluate handed thesis with classification grade **D - satisfactory.**

Date: **8.6.2017**

Signature