ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Identifying similarities in malicious network behaviour**
Author's name:	**Michal Stanke**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Ing. Ondrej Fikar
Reviewer's department:	Department of Cybernetics

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

Evaluation of thesis difficulty of assignment.

The goal of the thesis is to study algorithms to determine similarities in the behavior of different actors in the network environment. The student was supposed to search the literature and to evaluate suitability of the algorithms found for the specific task of reducing the amount of data needed to describe security events while preserving the important informations about the events. Algorithms selected by the student should have been implemented and connected into a framework. The reduction of the amount of data should have been evaluated.

I evaluate this assignment as challenging. Navigating the landscape of security literature is non trivial and the student is supposed not only to find the suitable algorithms but also to judge which of the algorithms are the best for this specific task and how they fit together for a joint application in one framework.

Also the metric for the final decision whether the framework fulfills its goal is not completely obvious from the assignment and the correct fulfillment of the last point of the assignment would require to specify such a metric clearly. Designing a suitable metric and its expression in a formal way might be in the end more challenging task than just implementing a set of algorithms.

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.

In general the thesis addresses all of the tasks given in the assignment. However, I have objections to all of them.

The student mentions few approaches to describe the behavior of the actors in the network and few ways how to aggregate them into groups of actors exhibiting similar behavior. It would be very difficult to give an exhausting list of the algorithms used in the domain but the given set of algorithms seem to be a bit incomplete and inaccurate. For example IP blacklisting might be considered an approach utilizing the similarity in the behavior (the same IP address) but there might be other algorithms meeting the criteria better. Another example would be the given list of clustering and partitioning algorithms which contains two well know approaches and two approaches adopted from the supervisor of the thesis. There is definitely more clustering algorithms in the literature. Mentioning few approaches which are not taught in the basic machine learning course would definitely make the thesis more interesting.

Furthermore the student decided to implement some of the algorithms he mentions in the literature survey but does not provide sufficient justification for his choice. A deeper analysis of the given algorithms and their comparison with the specific goal of thesis in mind would be very beneficial.

The selected algorithms are implemented and connected into a framework in a reasonable way and it seems the produced software yields reasonable results. I would appreciate more details describing the implementation but I do not see it as crucial given the research orientation of the thesis.

The last point of the assignment was to measure the data reduction achieved by the framework. The student did some measurements and the resulting numbers seem to show some reduction. The problem with this part of the thesis is that the results came somehow out of the blue. I would expect the student to precisely define how to measure the volume of the data at the input and at the output and what does it mean to preserve the information value in the data in the early stage of the design. Then the framework should be designed to optimize this measure and at the end the measure would be used to evaluate whether the framework works as intended. This approach would also provide guidelines for many design decisions which were made through the course of the work.

Without the precise definition of the evaluation measure many of the decision must be based on the intuition. This could actually be considered a valid engineering approach but it is difficult for the reviewer (or the potential user of the framework) to be sure that the decisions made were optimal. Furthermore, and I find this especially important, it is difficult to evaluate whether the framework actually brings any added value. If the goal of the thesis was to design, implement, and deliver a working piece of software then it was definitely fulfilled. If the actual goal was to answer the question whether the amount of data given at the input could be reduced without significant loss of information then the goal was somewhat fulfilled but it is not based on strong arguments and there remain many unanswered questions which in my opinion could be solved during the course of the work if it was done in a more rigorous way.

Method of conception **partially applicable**

Assess that student has chosen correct approach or solution methods.

In general the chosen approach is sound. The student had chosen some similarity measures and aggregates the input data according to them. Giving the information about the similar groups instead of individual data items may be considered a data reduction at least in the eyes of the analyst who is using the system.

Major objection here is that for the reasons given in the previous section, i.e. the lack of in the indepth analysis and the rigorous definition of the goal of the framework, it is not clear whether the proposed solution is the best (or at least reasonably good) or whether some other approach would be more suitable.

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.

From the software engineering point of view the work done seems to be good. The student proposes an architecture of the framework with extendability in mind, defines the interfaces for the classes used, and then connects them into a working software. I appreciate that the design allows adding more similarity measures easily.

I would appreciate if more attention was given to the algorithmic part of the problem, i.e. how exactly are the methods described theoretically in section 2 transformed into the algorithms to be implemented? What is the complexity of the problems? For example, some clustering algorithms may be very costly. Is it an issue given the assumed amount of data to be processed? What are the trade-offs implied by these facts? For data intensive tasks, which I believe is the case of this thesis, these questions should have been asked and answered.

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

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

I highly appreciate the thesis was written in English. Unfortunately, there is lot of mistakes, typos, and incorrectly used English words (such as in the sentence "I need to notice it was a great experience working on the thesis" on page 57). While I find this distracting it does not harm the readability of the text significantly. Anyway, I believe the thesis should have been proof read more carefully.

What is an issue is the structuring of the thesis which seems to be a bit chaotic. For example the methods used for the similarity computation are partially described in section "Analysis" and partially in section "Proposed solution design"

without any obvious reason for this division. I would, for example, propose to discus the mathematics behind the methods in one section and the algorithmic issues in the other section but in the thesis the mathematics is discussed in both sections and algorithmic point of view is omitted.

The included equations are sometimes insufficient to understand the problems at hand. Many times they are included out of context and sometimes there are variables and functions used without specifying what they actually mean (as in sections 2.4.4 at page 31 and 2.5 at page 35). Also it would be very helpful if the equations were numbered.

Selection of sources, citation correctness **C - good.**

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 distinction between information adopted from the literature and the students own contribution is clear.

The information adopted from the literature give very broad overview of the computer security field but, in my opinion, the given information are too general (e.g. there is a section describing the history of the Internet) and not enough attention is given to the problems directly related to the topic of the thesis. Also the literature recommended by the thesis supervisor is mostly ignored. One of my questions for the student would be whether there was a good reason for the omission.

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.

In general the thesis presents a design and implementation of a non-trivial piece of software. The design seems to be reasonable given the information presented in the thesis and the student implemented the software successfully.

I believe that significantly more attention should have been given to the initial analysis of the literature and the choice of the algorithms to be implemented. Also the final evaluation of the results should have been done in a more rigorous way and with respect to the initial analysis.

The language level of the thesis is not very high. The proofreading should have been done more carefully. Also more attention should have be given to the clarity of the presented equations and to the structure of presented information.

To conclude: In my opinion all parts of the assignment of the thesis were fulfilled to some extent even though some of them with major objections.

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.

1. In section 2.4.7 you describe the usage of histograms to compute similarity of transferred bytes and connection timing. How are the bins of the histograms chosen? Do you ensure consistency of the bins over different datasets? If so how do you do it?
2. It might happen that the similarity is undefined for some pairs of data items. What are the consequences for the clustering? How do you deal with the pairs with undefined similarity?
3. It was not clear to me what the equations in section 2.4.4 mean. Could you explain again how does the method described by the equations work?

4. You use terms "incident" and "event" in the thesis. Could you provide some formal definitions of the two terms?

I evaluate handed thesis with classification grade D - satisfactory.

Date: **24/01/2018** Signature: Ing. Ondrej Fikar