THESIS OPPONENT'S REPORT

I. IDENTIFICATION DATA

Thesis title:	Clustering Network Security Data for Efficient Human Analysis
Author's name:	Jan Svoboda
Type of thesis :	Bachelor
Faculty/Institute:	FEL, CVUT
Department:	Dept of Cybernetics
Thesis reviewer:	Sebastian Garcia
Reviewer's department:	Dept Computer Science, FEL, CVUT

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **Challenging**

How demanding was the assigned project?

The assignment was challenging due to the problem of working with URL data and the differences in the distribution. It is hard to evaluate.

Fulfillment of assignment **C- Good**

How well does the thesis fulfill 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 fulfilled the assignment in a good way since it is a hard topic. Some more considerations should have been given to the separation of data, evaluation, description of how the evaluation was done, and more description of the data. The dataset is a critical part of the problem, maybe even more than the classifiers or clustering algorithms.

Activity and independence when creating the final thesis

Assess whether the student had a positive approach, whether the time limits were met, whether the conception was regularly consulted, and whether the student was well prepared for the consultations. Assess the student's ability to work independently.

As reviewer this is not up to me to decide.

Technical level **B- Very good**

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

The thesis is technically sound as an approximation to the problem and exploration of ideas. The techniques were applied in a good way. It is nicely explained what the thesis does, but better explanations would have been better. Some points are still unclear: composition of dataset, details of the methodology, evaluation criteria on dataset C, etc.

Formal level and language level, scope of thesis **C- Good**

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 well organzed with minor comments on the sections. However some sections would greatly benefit for a deeper explanation and more graphs and diagrams. Especially in the analysis of the dataset.

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?

Some parts can be better cited, specially in the introduction. The rest of the citations are ok, but given that there has been so much work in the subject, some extra understanding of the limitations of the problem would have been good.

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

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

General comments for the Author:

- Saying that concept drift is a problem because old data does not resemble the new data is ok, but then this old data is used to help select which clusters are good candidates to be classified by training an ensemble. How come that that same old data is good if it still has concept drift? The errors are still there and they are not measured. The old data is treated as if it is not training. But it is training data. So there should be an analysis of how the distributions differ. What is that 'old data' used? where is it coming from?
- Why 'very certain' classified data is used to train the 'end classifier'? Would it make more sense to use the 'least certain' data for the end classifier since this data is the most interesting one for the analyst? The analyst already knows the 'certain data' well, so no need to use it in the end classifier.
- It would have been nice to have a real expert trying to label the clusters, since that is the whole idea of the research. But I get that here the 'best expert' was approximated. Should be good enough.
- Using the performance of a classifier as a measurement of how good the cluster is is very risky since many factors affect the classifier. Why not to use some of the most known cluster evaluation metrics, such as Silhouette Score and others.
- There is no good description of where the data comes from. Where it was captured, from whom, and what it is. There are some descriptions for some labels, like phishing, but not for others like benign.
- The whole idea of clustering is to find natural patterns in the data because they have some patterns together or relationship. But if the data is too different (another distribution) then what is the principle behind the clustering? that all 'malicious' URL look similar?
- Many URLs do not have 'www' in front, so deleting 'www' from the whole dataset inserts a huge bias. I would strongly advise against it without confirmation.
- The cross-validation technique is mostly used to compare different models/methodologies and to find which one works best. However, many decisions were taken in the thesis 'by hand' after some results were achieved in the training data. Example in page 33 "based on manually inspecting the clusters, we decided to relax the threshold". But these decisions are not 'verified' with a cross-validation.
- The classifier of the clusters is explained after the end classifier, which is confusing.
- If dataset C is the testing dataset and it is unlabeled, how were you able to compute an accuracy of 87%? There are no labels to compute any accuracy.

Date: **2024.06.3** Signature: