Doctoral thesis review

Thesis author	Mgr. Jan Kohout
Thesis Title	Representation of Communication in Computer Networks Security
Submission year	2019
Ph.D. Program	Electrical Engineering and Information Technology
Branch of study	Information Science and Computer Engineering

Reviewer	Mgr. Marta Vomlelová, Ph.D.	**Role**	oponent
Affiliation	Department of Theoretical Computer Science and Mathematical Logic Faculty of Mathematics and Physics Charles University		

Review:

a) The thesis addresses the network traffic analysis based on the encrypted communication monitoring. This is a highly actual topic.

b,c) It is focused on a very specific goal, the representation of the message set of the network communication. These sets come from the web proxy logs and similar sources.

The author presents a deep analysis of the idea to represent the set of messages as a histogram. The thesis is well written. It contains a review of relevant works, proposed ideas are tested on four flaw-based datasets and the results are presented in tables and graphs. The text is accompanied by illustrative figures and pseudo-code for algorithms.

d) The work brings several new results. First, the feasibility study to represent the message set as a histogram with several recommendations: the use of soft histograms, 11 bin discretization, pivot indexing, and a MapReduce algorithm for fast parallel (approximate) evaluation of k-nearest neighbour queries.

From the theoretical point of view is the most interesting the Section 7. The author reviews reproducing kernel Hilbert spaces. He selects a kernel for probability distributions, the maximum mean discrepancy MMD, and reviews its approximations. Compared to a random selection of basic features ($\mathbb{L}$ matrix) he presents an algorithm that select the basis with respect to a specific problem. He shows the improvement of such approach compared to the state-of art algorithms in three experiments. Furthermore, an application of this algorithm in SVM is presented.

The thesis concludes with an experiment on categorical (bag of words) data.

e) The results can be applied to practical network traffic monitoring. The work relates to four issued and one pending patent.

   The work is published in three journal papers that have citations already.

Topics for discussion and/or further research:

1) Does your histogram representation replace other features (statistics, fft, initial sub-sequences, other logs or knowledge bases) or how would you combine multiple sources of information?

2) It is nice to see a model based approach in the flood of deep learning. Can your relate to these works in terms of the performance and the possible idea exchange?

3) Technical question: Why are the results of Any-P better than Any-F (page 82)? Any-F has more information, I would expect better results.

The dissertation fulfills the conditions of independent creative work and it contains original and published results of scientific work and patents.

I recommend to accept the work as a doctoral thesis.

Prague, December 18th, 2019

Mgr. Marta Vomlelová, Ph.D.