Prof. Ing. Róbert Lórencz, CSc.
Department of Information Security
Faculty of Information Technology
Czech Technical University in Prague
Thákurova 9, 160 00 Prague 6
Czech Republic
Tel.: +420 22435 9812
E-mail: lorencz@fit.cvut.cz

Doctoral Thesis Review

Representation of Communication in Computer Networks Security

Submitted by Ing. Jan Kohout

The thesis has been submitted to the Faculty of Electrical Engineering, Czech Technical University in Prague in Ph.D., study program: Electrical Engineering and Information Technology and branch of study: Information Science and Computer Engineering.

Up-to-dateness of the dissertation

The growth of the complexity of Internet structures and the growth of data volume transmitted by them requires the development of new methods for monitoring and detecting the status of computer networks. More recently, these methods increasingly include machine learning algorithms that facilitate and speed up monitoring of networks and detection of unwanted or malicious activities. In this sense, the dissertation is very up-to-date. The effort of the author to design an abstract model of network communication representation is valuable.
The author in his work also pursues the urgent question of optimization of complex computational artificial intelligence algorithms used in the analysis of network flows.

Formal structure and organization of the thesis

The dissertation thesis structure reflects the effort of the author to combine several partial themes into a compact product. For the most part the author succeeded in doing that, with a few exceptions. The common denominator of the work should be artificial intelligence algorithms involved in the processing and analysis of network flows. The structure of the work corresponds to the requirements for the composition of the dissertation.
In the first chapter, the author presents his motivation for the research in the chosen field and outlines the goals of the work and the key contributions. In the second chapter, the author performed reviews of network protocols and network traffic representations, which have appeared in prior art works. The third chapter contains motivation of needs to create a formal model describing network communication independently of specific cases. For this purpose, the author introduced the term *message* as a basic unit of communication and the term *message set* dedicated to modeled communication. In the fourth chapter, all the datasets used

by the author for his experiments are described. Chapter five deals with a method for representation of network communication based on empirical histograms. The author's proposed representation uses so-called soft histograms, which help to suppress noise from experimental data.

The chapters six, seven and eight represent main contributions of the author. In chapter five, he introduced a method for effective similarity search on large-scale data, which uses the histogram representation. In addition, in this chapter he presents a method of using histogram approach especially for malware detection in encrypted HTTPS traffic. Then in chapter seven, he introduces a trade-off approach that uses representations building upon kernel embedding of probability distributions, thereby increasing performance the cost of increased storage requirements. Chapter eight deals with dictionary representations of persistent behavior of network entities. The last chapter nine summarizes author's achievements and shows further research.

The author starts the individual chapters by briefly describing the current state of the given research field and concludes each chapter by its summarization. The text of the work, however, is often interleaved by the parts that describe the contribution of the author with the parts that describe prior art. This makes it difficult for the reader to understand the author's own contribution.

Completion of the dissertation objectives

In chapter 1.2 Thesis goals and key contributions, the author lists four main objectives of the dissertation work:

1. Introduction of a basic model of network communication from which the representations can be derived in different scenarios and use-cases. The model treats the communication as identically and independently distributed messages drawn from a probability distribution, which characterizes the given communication.
2. Proposal of two main representation frameworks based on that model and present their applications – a histogram-based representation, which excels in its sparsity and scalability and a representation derived from kernel embedding of probability distributions, which significantly improves performance of algorithms working on top of it.
3. Experimental verification using data from real computer networks, showing that the representations are able to provide solid background for algorithms that leverage pair-wise similarities between the analyzed objects and improve the results achieved by prior methods.
4. Demonstration that in the case of each proposed representation they fulfil the following main requirements: The are general enough, such that they allow to model and compare samples of communication from different sources of data, at various levels of abstraction, and are independent of any specific algorithm working on top of them. This is important, because representations having these properties allow effective development of modular analytical systems, enable easy comparisons of individual algorithms, and can be quickly deployed when new types of input data or modeled entities appear. Experimental testing of the representations on different types of network traffic logs and together with algorithms for anomaly detection, clustering as well as classification.

I confirm that the stated goals of the dissertation thesis are consistent with the achieved results.

Assessment of the methods used in the thesis

The methods used in the dissertation thesis are relevant to the present research in the field of network security analysis. Almost all methods use machine learning algorithms. Particularly these are the kNN method, OutRank algorithm for outlier detection, Louvain method, spectral clustering, ECM classifier, and the method of kernel embedding of distributions. To verify the achieved results the author uses standard evaluation metrics as for example Recall, Precision, FP-50 error measure, Purity, ARI, etc.

Evaluation of the results and contributions of the thesis

The dissertation thesis does not always provide a clear connection between a proposed solution and the prior art works. It is difficult to recognize which parts of the text represent proposed methods of the author and which parts were already solved in other works.
The author performed a number of experiments on the data representations described in the thesis. These data represent a closed set for performing all experiments in the work. It would be interesting to use another data set for verification of correctness of achieved results. With a given set of data, the results obtained are convincing using the proposed methods.

Remarks, objections, notes, and questions for the defense

The DT exhibits several formal issues in descriptions, but I will not go into those here. List of abbreviations is missing. For this reason, the text is in some parts hard to read.
Questions:

- Page 20: "The model treats the communication as identically and independently distributed messages..." – Are these prerequisites actually fulfilled? (I would guess that the independence condition does not apply to packets as they are ordered according to some rules.)
- Page 36: "Therefore, the model ignores order of the messages within a message set. While this might seem limiting, we are aware of this possible oversimplification ..." – What would be the formal definition of a message (set), which would also take into account the order of messages? (Question is related to the question from the previous point.)
- Page 38: "Malicious connections produced by malware were obtained from 14 different malware binaries executed in a malware laboratory." – From what "malware laboratory" are the malware samples obtained?
- Page 44: "In the simplest form, the sample's contribution to two nearest bins depends linearly on distance to them ..." – The two closest bins do not need be determined unambiguously. What should be done correctly in such a situation?
- Page 44: "This way the soft histograms are able to estimate the underlying probability density in more accurate way than the commonly used hard histograms." – This statement is made without proof, or at least a reference to the literature. Why is this statement true?
- Page 61: "The exact number of pivots can be determined based on the volume of data and parameters of the MapReduce cluster." – How exactly was the number of pivots determined?

- Page 64: "All the 8 642 368 message sets composed of 145 822 799 messages were used for this experiment." – How many GB of data was that? (because the student used a computational cluster with up to 100 nodes)
- Page 70: How is the inner product <f, g> calculated in the Hilbert function space?

Remarks:

- Page 43: Here are two different names for the same: histogram bins and bins' centers (or on page 44 bins with centers) and their definition is in conflict with the sentence (last on this page) "For example, in a one-dimensional case, i-th bin with bounds [b_i, b_i + 1) is updated ... ", i.e. bins are first understood as real numbers and then as intervals.
- Page 43: "... probability that the next realization of f will be close to the center of bin b_i." - "close" is inaccurate, it should be "closest".
- Page 44: TF-IDF at the end of the page is not much explained, moreover in the sentence: "It multiplies each bin b_i in all histograms by..." it should not be b_i but h_i. Furthermore, this procedure does not work if no histogram has an empty bin (i.e., h_i is always positive).
- Page 47: "This confirms that the soft histograms indeed contribute to improved quality of the detection independently of the outlier detection method." – this is too strong a claim because hard and soft histogram representations were compared for only two methods (KNN and OutRank). Also, only one distance (cosine) was used.
- Page 51: The order of figures 5.4 and 5.3 is reversed.
- Page 55: "Then, in Section 6.2 we propose a method for effective k-NN similarity search..." – it is not clear from the text, what is exactly the author's contribution? I am asking because the lower-bounding principle with Voronoi partitioning was used in [102].
-

Overall evaluation

I consider the dissertation thesis successful, despite the critical remarks listed in the previous paragraphs. The reason for this claim lies in the amount of work that the author had to do to achieve the presented results. Despite its wide thematic spread and a great amount of text, the dissertation thesis does not contain significant formal or factual deficiencies.

Recommendation Statement

The dissertation thesis contains original results that have been published in prestigious magazines and conferences. The author of the dissertation proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act) I do recommend the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

V Praze dne 3. 1. 2020.

Róbert

Róbert Lórencz