Doctoral Thesis Report

Name of student: Ing. Tomáš Čejka
Thesis title: Stream-wise Parallel Anomaly Detection in Computer Networks
Reviewer: prof. Ing. Václav Přenosil, CSc.
Institution: Faculty of Informatics, Masaryk University

Up-to-dateness of the dissertation

The topic of dissertation is very up-to-date. Securing computer and communication networks plays a vital role in the modern world of digital communication. At present, the Internet of Things is a reality. It is the research area of many researchers and security companies around the world. However, there are still many open research questions. The presented dissertation deals with several problems related to network security and parallel processing of anomalies hidden in the large volume of network traffic data.

Formal structures and organization of the dissertation

The presented work has 120 pages including the references and appendices. The thesis document has the following structure:
1. Introduction
2. State-of-the Art
3. Topics and Contribution in Details
4. Conclusions
5. Includes Papers

General structure and organization of the dissertation is adequate, correct and logical, although I would prefer to see the subchapter 1.3 "Goals and Contribution of the Dissertation Thesis" as a separate chapter.

The author articles in Appendix A describe in detail the individual aspects of detection and parallel processing of detected anomalies in computer networks, which are discussed in the dissertation text.

The contents of the thesis document are technically sound and deep, covering all aspects of the described work. In addition, the dissertation provides an excellent overview of the current state of the art of computer network traffic monitoring, detection of the traffic anomalies and big data processing. The bibliography section contains majority of relevant work and is fully adequate to the presented thesis.
Completion of the dissertation objectives

The aims of the dissertation are presented somewhat unclearly in the first paragraph of chapter 1.3. From my point of view, the aim of this dissertation is processing the flow data in large backbone networks. In this text, I have found three partial goals:
- to design mechanisms to process huge volume of flow data as soon as possible especially for network security purposes,
- to extend the monitoring and analysis infrastructure to support additional information,
- to design an algorithm for splitting, which ensures that the machine receives sufficient data to identify security threats, allowing for parallel processing based on splitting a flow of data into independent subsets.

I can say that both the main goal and all three sub-goals are met in the chapters 3.2, 3.3, 3.4 a 3.5 and in appendix A.

Assessment of the methods used in the dissertation

The methods used for the dissertation work are in accordance with the methodology of the scientific work. The dissertation contains description of the problem, problem formulation, hypothesis determination, deduction from the hypothesis, experiment and evaluation of the experiment. I have no comments on the methodology used.

Evaluation of the results and contributions of the dissertation

There are five main results:
- description of a new stream-wise approach to network traffic analysis and anomaly detection using flow data without long-term storage,
- design and development of stream-wise and application-aware detection methods discovering threats on application layer through a framework that allows rapid prototyping of the detection methods,
- design and development of the scalable architecture for parallel processing the flow data with respect to witnesses,
- realization and evaluation experiments with real backbone traffic using the developed detection algorithms,
- implementation the formal symbolic description of the algorithms as the modules for the NEMEA system (open source framework and a set of modules for a stream-wise analysis of flow data).

All results and contributions are convincingly documented in this dissertation.
Remarks, objections, notes, and questions for the defense

The author managed to design the system that can run in a parallel distributed environment to handle big data. So he solved the crucial problem the stream-wise concept, especially for large networks. He has managed to find a way to minimize delays of detection and keeps the storage requirements low. For increasing reliability and precision of detection he uses some additional information from the L7 layer.

Questions for oral defense:

1) In the Chapter 1.1, Motivation, you mentioned Distributed on-the-fly anomaly detection. Please indicate the difficulty of this type of detection and how it should be addressed?

2) Extended flow records increase the accuracy and reliability of detection of the anomalies, but this tool works with unencrypted traffic only. Do you have any idea of how to detect anomalies in encrypted communication?

The overall evaluation of the dissertation

The use of English language is adequate, although a number of minor formal issues can be found. However, the overall quality of text is good and the thesis as a whole document provides clear description of the subject matter and the contributions made by the student. Quality of all graphics and illustrations provided in the thesis is very good, further contributing to the overall merit of the thesis.

The whole work is abundantly accompanied by the author's publication work. Citations and publishing activities are in line with the usual standards. I have not found indication of plagiarism.

Statement on the recommendation of the dissertation for the defense

Author in this thesis has demonstrated the ability to work independently in the specified field, and the thesis meets the standard requirements on a dissertation in the field. In accordance with par. 47 letter (4) of the Law Nr. 111/1998 (The Higher Education Act) I recommend that the dissertation is accepted and the candidate proceeds to oral defense with the aim of receiving the Ph.D. degree.

Brno September 11, 2018

prof. Ing. Václav Přenosil, CSc.
Faculty of Informatics Masaryk University
DISCUSSION REVIEW

Author: Ing. Tomáš ČEJKA, Czech Technical University in Prague

Title: Stream-wise parallel Anomaly Detection in Computer Networks

Reviewer: Doc. Ing. Ondřej RYSAVÝ, Ph.D., Brno University of Technology

A. Novelty and contribution
The presented dissertation focuses on specific challenges in network security monitoring aiming to provide on-the-fly solution for anomaly detection. The considered approach is based on Netflow record analysis. Although this is not a novel idea, the author's goal is to develop a new method enabling near real-time analysis of a high volume of incoming Netflow data. Contrary to existing solutions that require huge data storage for saving and processing Netflow records, the presented approach works with streams of data. It enables to perform certain operations without the need to store the entire dataset.
Achieving the stated goal is essential for monitoring of large network installations. The work is oriented towards providing the practical solution to the problem of large-scale network monitoring.
The contribution consists of a new network monitoring system based on the author's original idea. The novelty can be seen in adapting stream processing for anomaly detection based on analysis of extended Netflow records. Also, the author elaborates on the possibility of parallel processing by proposing a method for mapping incoming data to processing nodes that respects required relations among flow data.

B. Formal structure and organization
The thesis conforms to the formal requirements. It has the form of a collection of papers accompanied by about 40 pages of information that explains main ideas and provides references to the original papers written by the author. The text itself is structured into four chapters. The first chapter contains an introduction of the problem being solved and describes the context of the proposed solution. The second chapter consists of a brief state-of-the-art. The main contribution is presented in chapter 3. In this chapter, the author discusses the main principles of the proposed stream-wise approach to NetFlow analysis. The method is demonstrated on several detection algorithms. The thesis is concluded in chapter 4. The text is clearly written, all terms are well explained and it is very easy for a reader to understand the presented information. It is partly because the text itself does not contain many technical details nor results. These can be found in Appendix A, which contains 8 original papers written by the author. Each paper is provided with short information on the contribution of individual authors and the relevance to the presented thesis. All included papers were published in proceedings of well-recognized conferences.

C. Objectives and Methods
Objectives of the dissertation were stated by the author in sections 1.2 and 1.3 as follows:
- to enable near real-time processing of a high volume of Netflow data,
- to avoid the necessity of maintaining a long-term storage for Netflow data,
- to extend Netflow records with additional data to support application-aware monitoring, and
- to enable parallel processing of flow data in order to provide scalability of the proposed solution.

The author demonstrated the completion of all these objectives in the presented dissertation. The solution is based on the so-called stream-wise approach that performs stream processing of flow data. Stream processing enables to reduce memory requirements and also offers parallel execution of detection algorithms. Instead of using some existing stream processing engine, the author designed and implemented a dedicated system called NEMEA. The system integrates the presented ideas and was used in experiments. NEMEA is designed as the open framework enabling further extensions.

All stated objectives were completed. To achieve the intended goal, the author employed suitable methods. He started with detailed problem analysis and came with a new idea for a possible solution. Then he performed a feasibility study (by designing and implementing NEMEA tool). Finally, he demonstrated the properties of the solution by experimenting with the tool and available datasets.

D. Evaluation of Results

The significant results of the presented dissertation are methods and algorithms that were integrated into the NEMEA system. These results enable to perform near real-time monitoring of large computer networks. All main results were presented at scientific conferences. Except presenting the NEMEA system itself, the author also studied methods for detecting different types of security problems ranging from simple host/port scanning to SIP fraud detection. It was shown that the proposed system is capable of handling these security threats. Though the author mentions related work the direct comparison to other systems is not provided.

Experiments conducted with NEMEA system demonstrate the feasibility of the proposed approach. The most significant contribution can be seen in the successful application of stream processing approach to the efficient analysis of Netflow records that allows the scalability of the system.

E. Remarks

I have found the following minor issues in the document:

Equation (3.3) case for $i = 0$ is not necessary as the second case also includes the situation in which $F_0(\text{attr} \not\in \{\}$. $$

Equation 3.6 does not probably express the author's intention. Reading the text I assume that $I$ is an equivalence class on $F$. The provided definition rather express that $I$ is any strict subset (possibly empty) of $F$ such that its members have the same attribute "srcip".

In algorithms 3.2 - 3.5 there is a mismatch in a notation of flow subsets. For instance in 3.2 there should be $S_1 = \text{COUNT}([\text{DISTINCT_SUBDOMAIN(}[\text{domain}])]. Similarly in other algorithms.

Equation 3.8 suffers from the same problem as Equation 3.6.

Algorithm 3.8: CHECKBRUTEOFORCE function is not defined anywhere in the text.
F. Questions
Most of the presented detection algorithms rely on the threshold value to raise an alarm.
• Have you evaluated these algorithms in terms of their accuracy?
• Is it possible to include "history" in the detection methods to improve accuracy?

The author proposes hash-based scattering as the method of assigning flows to computational nodes:
• The method has an almost uniform distribution of flows. Each flow can be sent to up n-nodes, where n is the number of hash functions. What is the cost of flow distribution comparing to the cost of execution of detection algorithms?
• By construction, this method obeys "semantic" relations between flows. However, results presented in paper A.7 show that it misses some alerts. Why?
• Have you measured some other performance indicators than the number of flows that the scatter can process?
• What is possible speed up when adding new processing nodes?

For flow representation, the system uses UniRec structure.
• What are the benefits of UniRec in comparison to other serialisation formats?
• What is the performance gain of using UniRec comparing to other binary formats, e.g., ProtocolBuffers, Apache Thrift?

Overall Evaluation
The author of the dissertation demonstrated that he is capable of conducting independent research in Computer Science. He delivered original scientific contribution and performed necessary evaluation of achieved results. In accordance with par. 47, letter(4) of the Law Nr. 111/1998 I recommend the thesis for the presentation and defence with the aim of receiving Ph.D. degree.

Brno, 30th July 2018

Doc. Ing. Ondřej Ryšavý, Ph.D.
Zürich, September 10, 2018

Letter of Recommendation for Mr. Tomáš Cejka's Doctoral Thesis on “Stream-wise Parallel Anomaly Detection in Computer Networks”

Dear Sirs,

the thesis of Tomáš Cejka on “Stream-wise Parallel Anomaly Detection in Computer Networks” has addressed a very timely and relevant topic in the areas of today’s networking sector, especially within the areas of a flow-based network monitoring and network security. Since a stream-wise processing determines a suitable principle for performing a security analysis of large-scale computer networks’ data in transmission, all flow records have to be processed on-the-fly upon reaching the flow collector. Thus, the NEMEA framework had been developed and investigated in that respect as a proof-of-concept.

Tomáš shows that such an approach is possible and provides for a technically feasible approach, including the addressing of a key parameter of ever growing traffic based on parallel flow-based analysis. Major considerations have been evaluated and experimented with, where these had been performed on data sets from real backbone traffic of the Czech National Educational Network. Thus, Tomáš defines and investigates in this research work in detail his approach proposed to enforce a practically deployable solution for parallel anomaly detection and ensures that available resources are effectively controlled for a parallel use. Such an approach is new and in goes well beyond major related work.

Content and Contributions

Section 1 introduces the major basics of the thesis and summarizes the motivation, defines the targeted at challenges, and lists the thesis goals. This is carefully phrased into the problem statement, which determines as the challenging part the distribution of a single stream of data among multiple nodes, whereas each node should process just a subset of the original stream. Thus, a suitable mechanism of data distribution must be found, which preserves all results of a possible detection. Furthermore, since traditional flow records with very basic information are not sufficient for a detection of advanced threats — they may be similar to legitimate traffic — the reliability of the detection algorithm needs to be maintained for application layer traffic (L7). To address these aspects, an open source tool, due to lacking any suitable elements as of today, was determined and proposed to be developed.
Section 2 describes the current state-of-the-art and lists briefly related work in the area of network security, anomaly detection, and parallel processing using flow data. While monitoring approaches are discussed, relevant detection methods are summarized, too. The "Big Data Processing" is focused on parallel processing, while adding semantics of data as a discussion item. Thus, Section 2 does address the key input factors, but Tomáš does not deliver a clear set of dimensions against which related work and the newly proposed approach could be compared to.

Section 3 determines the core of the thesis, especially the solutions proposed to answer the research challenges as posed above. Unfortunately, Tomáš's responses do focus initially on the NEMEA framework, which was introduced in earlier chapters of this thesis. Thus, the scientifically relevant focus of a thesis remains a bit hidden within Section 3. The content of NEMEA — once called framework and once termed system or even tool —, however, is the key of this work and develops well beyond state-of-the-art knowledge, which was researched, obtained, designed, and evaluated. Thus, the flow-based analysis is explained by its essential features (application-awareness and stream-wise processing). Key terminology important for this work is defined, especially "witness"(an abstract representation of semantic relations in the flow data, significant in parallel processing). And the application of these terms onto the detection algorithms gains the major observations from this thesis, well beyond details, which are not part of the referenced set of eight papers published earlier. The set of papers written by Tomáš are integrated into the flow of arguments, such that the major and minor elements of the new approach are clearly interrelated.

Furthermore, in Section 4 Tomáš summarizes and partially concludes his work, and proposes next steps. While more general summarizing elements are discussed on a per-challenge basis, the real conclusions of his work are not made that explicit. The advantages reached by the work performed are not made explicit within the thesis, but within the papers, thus, conclusions in terms of user, provider, or operator views are a bit hidden. The key observation, however, is made very clear: "This research showed that some advanced security threats are invisible for the traditional NetFlow tools." And the solution provided within the thesis perfectly well demonstrates that these application-aware detection methods studied and designed are capable to overcome this lack of visibility! And very good for a systems-based thesis is the deployment success of the NEMEA within the Czech National Research and Education Network (NREN) — CESNET2. Unfortunately, the view into the future is very short and only states three out of the many interesting steps possible, without any minimal argumentation, why these three steps may add further advantages to the solutions and research findings already achieved.

Finally, the appendix includes the set of eight papers, which outline, define, and refine a number of aspects, relevant for a suitable answer toward the thesis' goals. Thus, the cumulative type of a thesis submission is well documented by these papers attached.

**Evaluation and Specific Comments**

In summary, Tomáš' work is very timely and was driven by a dedicated problem (theoretically and practically motivated at the same time) in which the technology of flow-based analysis in a real-time manner, while applying a stream-wise approach in terms of parallel analysis steps, was successful. These features are used in the detection modules designed and developed to evaluate the concept on data, which is originates from real network data.

It is a key strength of Tomáš' work that the core of his work is a combination of formalized detection algorithms and metrics (semantic relations termed "witnesses", and parallel flow data processing based on a Flow Scatter component) and its evaluation as well as its application in an operational prototype, called tool, too.
On one hand, Tomáš' thesis is broad in scope (flow-based analysis applied to anomaly detection), but it is well constrained (by the stream-wise view being prototyped in a parallel compute model), too. Besides the clear scientific perspective (not perfectly described in the thesis, well done within the papers, but excellently evaluated in the thesis) of this thesis' advances are very well documented beyond the state-of-the-art, especially the work in the application-aware stream-wise flow data processing. This needs to be considered to handle the practically complex approach, but being operationally useful. On the other hand, the prototype designed and evaluated is operational and well set-up and serves for way more than just a pure proof-of-concept, since it was experimented on real-life data and within CESNET2.

Tomáš' combination of methods chosen, experiments/evaluations and analysis performed, and interpretations being undertaken is well documented. His approach in this thesis is fully in-line with instruments and standards of the scientific communities of networking, which is well documented by the fact, too, that Tomáš published his research results in 5 first and 3 second author papers in peer-reviewed conferences and workshops (AIMS 2017, CNSM 2016, AIMS 2016, AIMS 2015, and IEEE CAMAD 2014). Overall, this is an excellent result for the duration of the thesis, since 9 additional conference papers have been published besides the thesis' topic.

Recommendation

Overall, this thesis contributed in detail to those areas listed above, for which these findings are new, unique, and evaluated in many details throughout Tomáš Cejka's written thesis. The excellent combination of theoretical and practically applicable results, combined with a thorough and in-depth evaluation — again both based in real-life data and with the running/operational system experiments — leads to a highly relevant, applicable, and very well founded knowledge advance well beyond today's state-of-the-art. The written thesis may contain, mainly due to the use of those papers considered relevant, a few inconsistencies with respect to terminology applied.

Based on this evaluation of the Ph.D. thesis submitted, Tomáš's very good performance in the thesis, and his descriptive, extensive, and constructive work, I can state that at all in all, the work determines a very good Ph.D. thesis.

Tomáš Cejka, the author of the thesis, proved the ability to conduct research and achieved scientifically valid results. Thus, I do recommend to the Czech Technical University in Prague this dissertation for the presentation and defense with the aim of receiving a Ph.D. degree. While evaluating the overall work seen, a very good grade should be considered to be awarded.

Kind regards,

Prof. Dr. Burkhard Stiller