DISSERTATION REVIEW

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

Title:Stream-wise parallel Anomaly Detection in Computer Networks

Reviewer: Doc. Ing. Ondřej RYŠAVÝ, 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[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 = COUNT(DISTINCT_SUBDOMAIN(I[domain]). Similarly in other algorithms.

Equation 3.8 suffers from the same problem as Equation 3.6.

Algorithm 3.8: CHECKBRUTEFORCE 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.