Review of a Doctoral Thesis at FIT BUT

Doctoral thesis (hereinafter referred to as "thesis"), title of the thesis: BIG DATA ANALYSIS TECHNIQUES FOR NETWORK TRAFFIC MONITORING: THE STORY OF DNS OVER HTTPS DETECTION

Name of the doctoral student (hereinafter referred to as "student"), name and surname:

KAMIL JEŘÁBEK

Name and institution of the reviewer (full name of the reviewer, full name and country of the institution):

Peter Shaojui Wang

National Taiwan University of Science and Technology, R.O.C. (Taiwan)

Please state your opinion on the following aspects of (I) the student's thesis and (II) the student's overall achievements, and (III) state your conclusion (a minimum of approx. 300 characters foreach item below is recommended):

I. Thesis

Appropriateness and relevance

Is the area addressed by the thesis appropriate to the particular scientific discipline of the thesis and does the thesis address relevant problems within the chosen area?

Yes. This thesis tackles the challenge of detecting DNS over HTTPS (DoH) in the context of Network Traffic Monitoring. Initially, the DNS protocol allowed hosts to translate domain names into IP addresses. However, due to potential privacy concerns with the original DNS protocol, encrypted alternatives such as DNS over HTTPS (DoH), DNS over TLS (DoT), and DNS over QUIC (DoQ) were developed. These methods encrypt DNS traffic between users and resolvers to safeguard privacy. While DoT and DoQ utilize port 853, DoH uses port 443, the same as the standard HTTPS protocol. However, for security operators, these encryption protocols pose a risk by complicating the monitoring of network traffic, as they obfuscate the content, including potentially malicious packets. Consequently, some security operators and organizations focus on identifying and blocking the use of these encrypted DNS protocols. While detecting DoT and DoQ usage through port 853 is feasible, DoH presents a challenge as it shares port 443 with regular HTTPS traffic, blending in seamlessly. This situation raises a critical question for security operators: how to distinguish DoH traffic from HTTPS traffic? This is the central issue addressed in this thesis. The thesis begins by discussing the background and related work, then shares security observations from both server (resolver) and client (browser) perspectives, including analyses of well-known DoH resolvers and browser DoH traffic. The thesis also details the datasets used, introduces a novel approach for DoH detection, and compares it with existing methods.

A summary of the contributions of the thesis

Review of a Doctoral Thesis at FIT BUT

From your point of view, please summarize what the goal of the thesis is, what the main contributions of the thesis are, and whether the thesis has achieved the chosen goal.

Please indicate also specific contributions of the student.

Detecting DNS over HTTPS (DoH) poses a greater challenge than identifying other encrypted DNS alternatives within communication traffic. The primary objective of this thesis is to develop a reliable method capable of accurately identifying DoH traffic and distinguishing it from regular HTTPS traffic. Moreover, the effectiveness of the proposed method is evaluated under various scenarios to assess its characteristics, practical applications, and its comparison with existing methods. The major contribution of this thesis is the introduction of a novel DoH detection method that integrates IP-based, machine learning-based, and active probing detection techniques, significantly outperforming methods that rely solely on machine learning and often result in suboptimal outcomes. This success is attributed to the limited effectiveness of using machine learning alone for network detection tasks, as high-accuracy models may generate a significant number of false positives over time, potentially overwhelming network operators. In contrast, combining machine learning with precise, commonly used network approaches like filtration and blocklists has shown superior performance. Furthermore, the incorporation of domain knowledge enhances the creation of comprehensive datasets and the design of effective solutions. The proposed method tackles the challenge of detecting DoH by offering a practical, lightweight, and compatible solution that is stable, generalizes well across different networks, maintains high accuracy, and minimizes false positives. It also addresses the detection of short flows. Utilizing just four lightweight features that can be extracted from most current monitoring infrastructures, including high-speed and backbone networks, the method stands out from other approaches. These alternative approaches typically rely on complex, hard-to-compute statistical flow features that cannot be computed on running sequences, limiting their applicability and demonstrating low generalization capabilities. Therefore, the solution is considered satisfactory, fulfilling the goals set at the outset of the thesis.

<u>Novelty and significance</u>:

Please assess the level of novelty of the results and their significance for the given scientific area, for its further development, and if applicable for possible applications in practice.

Given that none of the existing proposals provide a satisfactory solution for reliable DoH detection in real environments, it is clear that a new approach is necessary. Current methods rely on complex, hard-to-compute features and machine learning-based classifiers, which, despite their low false positive rates, still produce unacceptable levels. Considering a network with a throughput of 1000 flows per second, even a 0.001 false positive rate would result in one alarm every second, inundating security personnel with a high number of false alarms. Moreover, these detectors require additional data about connections, such as individual packet lengths or the median and mode of packet sizes—information that standard network monitoring tools on high-speed networks typically do not support, significantly restricting their deployability. To address these limitations, the proposed method diverges from a reliance solely on machine learning. Instead, it adopts a heterogeneous detection approach, incorporating three different types of detection: IP-based, ML-based, and active probing. The detection pipeline employs a feed-forward loop, where a resource-intensive but reliable verification step generates a blocklist/allowlist, which is then utilized by a swift IP-based detector. To reduce the number of active probes, an ML-based classifier is used to select DoH-suspicious flows that warrant further verification. This tripartite approach mitigates the shortcomings of each detection method: the obsolescence of IP lists is countered by active verification and continuous updates; the inaccuracy of machine learning is offset by active verification; and the resource demands of active verification are minimized by preliminary IP list and machine learning filtration. This innovative DoH detection strategy can utilize standard flow data sources, making it applicable to nearly any flow monitoring infrastructure—from local area networks to high-speed backbone lines—while still achieving an accuracy rate of over 99.9%.

Evaluation of the formal aspects of the thesis:

Please evaluate formal qualities of thesis and its language level.

The formal qualities of this thesis are commendable. It begins with a well-defined structure that provides essential information about the relevant technologies, followed by a discussion on the current state-of-the-art related works. It then analyzes well-known DoH servers, which serve as representative samples likely to be chosen by users, noting how different server deployments may significantly influence traffic characteristics. The thesis continues with an analysis of DoH behavior, starting with the performance of single query DoH (which could be employed by applications such as malware), then examines the impact of different HTTP methods on DoH in browsers, and concludes with a traffic shape analysis that reveals the characteristics of DoH traffic. Subsequently, a reliable DoH detection method is proposed. This method is rigorously tested, including assessments with real-world traffic and data drift testing, and it is compared to other published methods which were thoroughly reproduced specifically for this purpose. The work concludes in the final section. Overall, this thesis presents its information in a logical sequence. In the experimental design and results, it also offers rich details and clear explanations. Additionally, the use of language throughout the thesis is well-executed and easy to understand.

Quality of publications

Has the core of the thesis been published at an appropriate level? Please judge the quantity and quality of the publications. When judging the quality, please take into account internationally recognized standards (WoS/Scopus quartiles, CORE ranks, specific knowledge of flagship publication channels of agiven community, etc.) in a way appropriate for the given area of the thesis.

Yes, the core content of the thesis has been published in two journals and two conferences. The first journal paper, titled “DNS over HTTPS Detection Using Standard Flow Telemetry,” was published in the IEEE Access journal in 2023, which is ranked Q1 in the 2022 SJR and Q2 in the 2023 WOS. The second journal paper, “Collection of Datasets with DNS over HTTPS Traffic,” appeared in the Data in Brief journal in 2022, ranked Q4 in the 2022 SJR and Q1 in the 2022 AIS. On the conference front, the first paper, “Analysis of Well-Known DNS over HTTPS Resolvers,” was presented at the 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), where it received the Best Paper award in category, and is indexed in WOS (not in Core). The second conference paper, “Big Data Network Flow Processing Using Apache Spark,” was published in the Proceedings of the 6th Conference on the Engineering of Computer Based Systems (ECBS 2019) in 2019, and was honored as the Best Conference Paper. It is ranked B in the CORE 2018 and B1 in Qualis. Overall, the core of the thesis has been published in internationally recognized journals and conferences, demonstrating a convincing quality.

II. Student’s overall achievements

Overall R&D activities evaluation:

Does the student's thesis, the results included into it, and possible other scientific achievements listed in the list of scientific activities indicate that he/she is a person with scientific erudition and creative abilities?

Yes. The student's thesis presents a compelling research direction by integrating various detection techniques, distinguishing itself from past research that focused primarily on machine learning. The thesis shows that the research ideas were developed through detailed observations in experiments and by building on previously published results, providing a solid foundation to support the student's perspectives. This demonstrates the student's significant potential in scientific research. Additionally, the student's rich industrial experience in product development has enabled him to create practical, real-world solutions. The student's experience with international stays also enhances his global perspective on research.

Assessment of other characteristics (optional):

More characteristics of the student may be added here (e.g., awards, grant participation, international collaboration, etc.).

The student has received the Best Paper Award from the CCWC Conference in 2023 and from the ECBS Conference in 2019. Additionally, they have earned the Expert Panel Award and the Professional Community Award at the Excel@FIT 2016, a conference for students.

III. Conclusion

The conclusion should contain an explicit statement saying whether, in your opinion, the thesis and the student´s achievements until now meet the generally accepted requirements for the award of an academic degree (in accordance with Section 47 of Act No. 111/1998 Coll., on higher education institution).*

* Short overview of both the Act and corresponding internal BUT regulations is enclosed.

Review of a Doctoral Thesis at FIT BUT

Yes, the student demonstrates performance that meets the generally accepted criteria for the award of an academic degree. In his thesis, he has developed a reliable method capable of accurately identifying DNS over HTTPS (DoH) traffic and differentiating it from regular HTTPS traffic. His innovative approach incorporates three distinct types of detection: IP-based, machine learning-based, and active probing, which together achieve the best performance results to date, surpassing traditional methods that rely solely on machine learning and yield suboptimal outcomes. This groundbreaking DoH detection strategy utilizes standard flow data sources, making it applicable to almost any flow monitoring infrastructure, from local area networks to high-speed backbone lines, while still maintaining an accuracy rate of over 99.9%. This approach results in a more practical solution than ever before. Therefore, in my opinion, the student meets the accepted standards for the award of an academic degree.

Place (e.g. Brno) DD.MM.YYYY

Taipei 13.04.2024

Signature of the reviewer: