A Survey of QUIC-Based Network Traffic Identification

Xiaolin Gui(✉), Yuanlong Cao, Longjun Huang, Yong Luo, and Jianmao Xiao

School of Software, Jiangxi Normal University, Nanchang, China
{xlgui,ylcao,luoyong1020,jm_xiao}@jxnu.edu.cn

Abstract. With the QUIC (Quick UDP Internet Connection) protocol recognized by the Internet Engineering Task Force as the core protocol of HTTP/3, network traffic based on the QUIC protocol (also known as "QUIC traffic") will become one of the primary traffics on the Internet. Network administrators use QUIC traffic identification as the foundation for network management. Numerous studies on QUIC traffic identification and application are already underway with the goal of assisting network operators, and pertinent results are beginning to emerge. We evaluate the topic of QUIC traffic identification to help future researchers rapidly grasp the research frontier of QUIC traffic identification and to summarize the present research and understand the obstacles in the field.

Keywords: QUIC · HTTP/3 · Network traffic identification

1 Introduction

1.1 Backgroud

At present, network traffic encryption is mainly implemented by adding an SSL/TLS (Secure Sockets Layer/Transport Layer Security) layer on top of the TCP (Transport Control Protocol) protocol at the transport layer. To increase communication effectiveness, Google suggested the QUIC (Quick UDP Internet Connection) protocol [1]. To ensure reliable communication, the protocol incorporates multiplexing, traffic encryption, congestion control, and forward error correction. Additionally, it avoids the requirement for numerous handshakes and key negotiations in the TCP protocol by employing the UDP protocol at the lowest layer. Since then, other academic academics have examined the QUIC protocol, assessed its effectiveness, and verified its benefits [2–5]. The application of the QUIC protocol has also been monitored and supported by the industry. Web browsers represented by Chrome, Firefox, and Safari have successively announced their support for the QUIC protocol. Internet companies represented by Google and Akamai have successively deployed QUIC applications on the server side. The IETF (Internet Engineering Task Force) recognized the QUIC protocol as a global standard for HTTP/3 in May 2021. With the promotion of academia, industry, and the International Organization for Standardization, it is foreseeable that network encrypted traffic based on the

Y. Cao and X. Shao (Eds.): MONAMI 2022, LNICST 474, pp. 365–372, 2023.
https://doi.org/10.1007/978-3-031-32443-7_26

QUIC protocol (also known as "QUIC traffic") will become one of the main traffics on the future Internet. Therefore, there is an urgent need for network operators to identify QUIC traffic in order to effectively manage the network. To this end, many researchers have started research on QUIC traffic identification and have achieved some results. This article aims to review these results and discuss possible future research directions.

2 Overview of QUIC-Based Network Traffic

2.1 QUIC-Based Network Traffic Datasets

As the QUIC protocol is not yet commonly used, datasets available for QUIC traffic identification are relatively scarce. At present, the widely used public QUIC traffic datasets are released by the team led by Professor Xin Liu at the University of California, USA. The datasets were captured in the lab at the University of California, Davis, and include 5 Google services: Google Drive, Youtube, Google Docs, Google Search, and Google Music [6]. The datasets were acquired on several systems with different configurations, including Windows 7, 8, 10, and Ubuntu 16.4, using Selenium WebDriver and AutoIT tools to write scripts to mimic human behavior when surfing the Internet. During the preprocessing of this dataset, all non-QUIC traffic is removed, and all flows in this dataset are marked, suitable for validating some of the methods in this project. In addition, the researchers constructed some private datasets according to their own research needs.

2.2 Categories of QUIC-Based Network Traffic Identification

QUIC traffic identification refers to the output form of the identification result. The identification level is determined by the requirements of network operators. QUIC traffic can be gradually refined from attributes such as protocol, application, and service to realize protocol identification, application identification, and abnormal traffic identification. We summarize the varied kinds of identification as follows:

1) Protocol identification. It is to distinguish QUIC traffic from hybrid network traffic. In the future, the traffic based on the QUIC protocol on the Internet will coexist with the traffic based on the traditional protocol for a long time. How to identify the QUIC traffic and reveal its characteristics is a problem that needs to be studied.
2) Application identification. It is to identify the application to which the traffic belongs, such as WeChat, BitTorrent, or YouTube. These applications can be further refined; for example, WeChat can be divided into text short messages, voice short messages, voice calls, video calls, and file transfer.
3) Service identification. It is to identify the type of service to which QUIC traffic belongs, such as web browsing, streaming media, instant messaging, and cloud storage.
4) Abnormal traffic identification. It is to identify malicious traffic such as DDoS, botnet, and APT.

3 Review of Recent QUIC-Based Traffic Research

Due to the multiplexing, traffic encryption, congestion control, forward error correction, and 0RTT (0 Round Trip Time) connection establishment of the QUIC protocol, there are fewer feature dimensions extracted from QUIC traffic than from traditional protocols. The characteristics extracted from QUIC traffic mainly include time series dimension information, statistical information, and byte stream information. At present, there are few works that take QUIC traffic as the research object, and the existing research can be divided into four categories: (1) Protocol identification [7]; (2) Application identification [8, 9]; (3) Service identification [6, 10]; (4) Abnormal traffic identification [11, 12]. Details will be given below.

3.1 Protocol Identification

There are varying data encapsulation formats as well as interaction processes for different communication protocols. Therefore, we need to understand the interaction process of different protocols, and find out the characteristics and laws that can be used to distinguish different applications in the interaction process. And then, it is possible to summarize the best feature attributes of each application protocol in network traffic [7]. And finally, these features lay the foundation for improving the granularity and accuracy of overall flow identification. Protocol identification can be at different levels in communication, such as identification of QUIC and non-QUIC protocols, identification of encrypted protocols, etc. Among them, the identification of QUIC and non-QUIC protocols is relatively simple and can be distinguished by the fingerprint feature of the packet header. The identification of encryption protocols is the key and most difficult content of protocol identification. The interaction process of encryption protocols can be roughly divided into two stages: (1) the first stage is to establish a secure connection, including a handshake, authentication, and key exchange. During this process, both parties negotiate supported encryption algorithms, mutual authentication, and key generation; (2) the second stage uses the key generated in the first stage to encrypt and transmit data. At present, the three mainstream encryption protocols are IPSec, SSH, and SSL.

3.2 Application and Service Identification

Traditional machine learning methods are widely used in the field of traffic identification. For QUIC, a new type of traffic, researchers also try to use traditional machine learning methods to identify QUIC traffic. The authors in [9] proposed a method to extract the characteristics of combined application data units from network traffic and used a variety of traditional machine learning methods to evaluate the quality of video streams based on the QUIC protocol, and obtained good results. This method could solve the difficult problem of video stream quality assessment caused by QUIC stream multiplexing. However, this method could not be used for other QUIC application network flows and could not be effectively applied to the application-level classification of QUIC traffic.

With the application of deep learning in various situations, more and more researchers try to use deep learning methods for QUIC traffic identification. The authors in [10] proposed a method that uses a convolutional neural network (CNN) to integrate feature

extraction and classification and applies it to five services based on Google's QUIC protocol. This kind of method integrates feature extraction and classification by using CNN, and these methods always have high accuracy in private datasets. However, the method mentioned in [10] classifies a single QUIC stream ten times, which in turn consumes a lot of CPU time and memory space, and at the same time extracts fewer feature dimensions. In [6], the authors propose a semi-supervised CNN to classify QUIC traffic. Specifically, the authors use a large amount of unlabeled encrypted traffic data for initial training of the CNN and then use a small number of labeled QUIC encrypted traffic data sets for secondary training of the model. It is suitable for the case when there are only a few labeled samples. This method can avoid the problem that a large amount of labeled data is required for model training. However, the generalization performance is poor, and the classification accuracy on some datasets cannot meet the requirements of practical applications.

At present, although there are few works on QUIC traffic classification, in the field of non-QUIC traffic research, there have been many research results in recent years, such as using traditional machine learning algorithms to classify encrypted network traffic [13, 14], malicious traffic detection [15], using deep learning to classify encrypted traffic such as VPN and HTTPS, etc. [16–19]. These results have important reference significance for the research on QUIC traffic classification.

3.3 Abnormal Traffic Identification

Website fingerprinting attackers can infer the website visited by network users from network traffic. And website fingerprinting attacks (WFP) can be implemented through multi-classification tasks. In [11], the authors studied the safety of the three protocols, that is, QUIC, gQUIC, and HTTPS/2, to resist WFP from the perspective of traffic analysis. The authors collect network traffic through a controlled environment consisting of three Web servers running Ubuntu 18.04. In addition, the author selected the official landing pages of the top 100 schools from the Times World University Rankings. The main Web pages of these websites were downloaded and stored on three hosting servers. Each server uses Docker to isolate resources from these websites. To compare the effectiveness of different protocols against fingerprinting attacks, the authors used five machine learning models for testing, namely random forest, decision tree, k-nearest neighbor, naive bayes, and support vector machines, and used 10-fold cross-validation to obtain experiment results. According to the experiments, the following conclusions can be drawn: (1) At the beginning of the connection, the gQUIC and QUIC protocols are more vulnerable to the threat of WFP than HTTPS/2, but if considering the full-traffic situation, the performances of the three protocols are similar. (2) When considering the full traffic of both parties in communication, most of the characteristics of the three protocols can be converted to each other. While the traffic characteristics of the three protocols are quite different if only considering the traffic at the beginning of the connection. (3) When tested with only 40 packets and some simple features, the attack accuracy of gQUIC reaches 95.4%, QUIC is 95.5%, and HTTPS/2 is only 60.7%. Since the QUIC protocol includes the function of network padding, the authors in [12] studied the effectiveness of network layer padding in preventing website fingerprinting attacks. To this end, the author prepared two datasets, the hybrid datasets and the QUIC datasets.

The authors found that in a mixed dataset with only 4% of QUIC traffic, the classifier was biased towards TLS-specific features. Therefore, the author constructed a dataset with QUIC traffic accounting for 70%. In this paper, the random forest model is used for classification, and the results show that the network layer padding is almost ineffective against network fingerprinting attacks. In addition, the author also proposes the idea of padding data at the application layer to counteract website fingerprinting attacks. To sum up, although the QUIC protocol has been improved in many aspects, it does not have a big advantage in resisting website fingerprinting attacks. And in some scenarios, it is even weaker than HTTPS/2.

4 Evaluation

At present, the evaluation of traffic identification and classification is mainly based on the use of accuracy-related indicators. This indicator is relatively simple. To meet the network traffic analysis requirements, there are many new evaluation indicators proposed. The following introduces several common evaluation indicators currently used in QUIC traffic identification and classification.

(1) Accuracy

The percentage of the total number of samples that are correctly predicted. The details of the parameters are described in Eq. (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The accuracy indicator has a disadvantage, that is, the data samples are not balanced, and this indicator cannot evaluate the performance of the model. Suppose a test set has 99 positive samples and 1 negative sample. The probability that the model predicts all samples as positive will be 99%.

(2) Precision.

Precision is an evaluation index for prediction results. Among the results predicted by the model as positive samples, the percentage of true positive samples can be calculated in Eq. (2).

$$precision = \frac{TP}{TP + FP} \tag{2}$$

The meaning of the precision is how many of the results were predicted as positive samples.

(3) Recall

Recall is an evaluation index for the original sample. Among the actual positive samples, the percentage of predicted positive samples can be described in Eq. (3).

$$recall = \frac{TP}{TP + FN} \tag{3}$$

5 Discussion

With the growth of QUIC traffic on the Internet, identifying the type of traffic carried by the QUIC protocol and effectively managing and controlling it has become a problem that network operators are about to face. Compared with the current protocol of the transport layer, the QUIC protocol integrates more functions, and its QUIC packet organization and traffic characteristics are quite different from the current transport layer protocol. Therefore, the existing encrypted traffic classification methods cannot be directly used for QUIC traffic classification. In order to cope with the management and service of the future network calmly, it is urgent to research the classification of QUIC traffic. We believe that there may be the following three challenges in the field of QUIC traffic classification.

(1) How to build a classification model for QUIC traffic?
The QUIC protocol has the functions of multiplexing, traffic encryption, congestion control, and forward error correction and data transmission using the UDP protocol. These functional characteristics create the unique traffic characteristics of the QUIC protocol, which can be used for traffic classification. QUIC traffic classification can be subdivided into multiple levels, such as: protocol classification, application classification, user behavior classification, and so on. For these different levels of classification requirements, it is necessary to understand the corresponding QUIC traffic characteristics and build a suitable model to classify QUIC traffic.
(2) How to solve the stability of the QUIC traffic classification model?
The training data set in the field of network traffic classification generally has an unbalanced distribution of traffic data. Generally speaking, the classification result always has the problem of majority bias if you train the unbalanced data set directly. That is, it is easy to misreport the category of the traffic with a small number of samples. Therefore, it is necessary to consider how to keep the classification accuracy stable when building a QUIC traffic classification model.
(3) How to enhance the scalability of the QUIC traffic classification model?

In the face of complex and changeable network environments, the scalability of the QUIC traffic classification model is important for network operators. That is, the classification model can quickly identify new types of traffic based on the existing classification functions. Therefore, this kind of model needs a new structure to meet the requirements. Therefore, it is necessary to study the scalability method for the QUIC traffic classification model.

Acknowledgment. This work was supported by the Natural Science Foundation of Jiangxi Province under Grant No. 20192ACBL21031.

References

1. Langley, A., et al.: The quic transport protocol: Design and internet-scale deployment. In: Proceedings of the conference of the ACM special interest group on data communication, 183–196 (2017)
2. McMillan Kenneth, L., Zuck Lenore, D.: Formal specification and testing of QUIC. In: Proceedings of the ACM Special Interest Group on Data Communication, pp. 227–240 (2019)
3. De Coninck, Q., et al.: Pluginizing quic. In: Proceedings of the ACM Special Interest Group on Data Communication, pp. 59–74 (2019)
4. Shreedhar, T., Panda, R., Podanev, S., Bajpai, V.: Evaluating quic performance over web, cloud storage and video workloads. IEEE Transactions on Network and Service Management, 1–16 (2021)
5. Chiariotti, F., Deshpande, A.A., Giordani, M., Antonakoglou, K., Mahmoodi, T., Zanella, A.: QUIC-EST: a QUIC-enabled scheduling and transmission scheme to maximize VoI with correlated data flows. IEEE Commun. Mag. **59**(4), 30–36 (2021)
6. Shahbaz, R., Xin, L.: How to achieve high classification accuracy with just a few labels: a semi-supervised approach using sampled packets (2018). arXiv preprint arXiv:1812.09761
7. Zhao, J., Jing, X., Yan, Z., et al.: Network traffic classification for data fusion: a survey. Information Fusion **72**, 22–47 (2021)
8. Van, T., Anh, T.H., Souihi, S., Mellouk, A.: A novel quic traffic classifier based on convolutional neural networks. In: 2018 IEEE global communications conference (GLOBECOM). IEEE, pp. 1–6 (2018)
9. Hua, W., Guang, C., Xiaoyan, H.: Inferring adu combinations from encrypted quic stream. In: Proceedings of the 14th International Conference on Future Internet Technologies, pp. 1–6 (2019)
10. Peng, Y., He, M., Wang, Y.: A federated semi-supervised learning approach for network traffic classification (2021). arXiv preprint arXiv:2107.03933
11. Zhan, P., Wang, L., Tang, Y.: Website fingerprinting on early QUIC traffic. Comput. Netw. **200**, 108538 (2021)
12. Barman, L., Siby, S., Wood, C., et al.: This is not the padding you are looking for! On the ineffectiveness of QUIC PADDING against website fingerprinting (2022). arXiv preprint arXiv:2203.07806
13. Giuseppe, A., Giampaolo, B., Domenico, C., Antonio, M., Valerio, P., Antonio, P.: Characterization and prediction of mobile-app traffic using Markov modeling. IEEE Trans. Netw. Serv. Manage. **18**(1), 907–925 (2021)
14. Shi, D.: Multi class SVM algorithm with active learning for network traffic classification. Expert Syst. Appl. **176**, 114885 (2021)
15. Chang, L., Longtao, H., Gang, X., Zigang, C., Zhen, L.: Fs-net: A flow sequence network for encrypted traffic classification. In: IEEE INFOCOM 2019-IEEE Conference On Computer Communications, pp. 1171–1179. IEEE (2019)
16. Cong, D., Zhang Chen, L., Zhigang, L.B., Bo, J.: CETAnalytics: comprehensive effective traffic information analytics for encrypted traffic classification. Comput. Netw. **176**, 107258 (2020)
17. Giuseppe, A., Domenico, C., Antonio, M., Antonio, P.: DISTILLER: encrypted traffic classification via multimodal multitask deep learning. J. Netw. Comput. Appl. **183**, 102985 (2021)

18. Lin, X., Xiong, G., Gou, G., Li, Z., Shi, J., Yu, J.: ET-BERT: a contextualized datagram representation with pre-training transformers for encrypted traffic classification (2022). arXiv preprint arXiv:2202.06335
19. Cao, Y., Ji, R., Ji, L., Lei, G., Wang, H., Shao, X.: l2-MPTCP: A Learning-driven Latency-aware Multipath Transport Scheme for Industrial Internet Applications. IEEE Trans. Industr. Inf. (2022). https://doi.org/10.1109/TII.2022.3151093
20. Cao, Y., Ji, R., Huang, X., Lei, G., Shao, X., You, I.: Empirical mode decomposition-empowered network traffic anomaly detection for secure multipath TCP communications. Mobile Networks and Applications (2022). https://doi.org/10.1007/s11036-022-02005-6