Behavior-Based Method for Real-Time Identification of Encrypted Proxy Traffic

Ping Luo, Fei Wang, Shuhui Chen
College of Computer
National University of Defense Technology
Changsha, China
{ping.l, wangfei09a, shchen}@nudt.edu.cn

Zhenxing Li
College of Computer
National University of Defense Technology
Changsha, China
372652329@qq.com

2021 13th International Conference on Communication Software and Networks (ICCSN) | 978-1-6654-3182-8/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCSN52437.2021.9463594

***Abstract*—Encrypted proxy is often used to hide malicious behavior or criminal activity on the Internet. Therefore, identifying encrypted proxy traffic is essential for network management and communication security. Existing researches usually use statistical features to profile network flows, which only have limited effects on encrypted proxy traffic, and are not suitable for real-time identification. In this paper, a novel behavior-based approach for encrypted proxy traffic detection is proposed. Two unique behavior features, IP proxy and data encryption behaviors, which are highly related to the activity of accessing network through encrypted proxies, are defined as learning features. Machine learning techniques are adopted for encrypted proxy traffic identification. The experiments on a real V2Ray traffic dataset demonstrate that the behavior-based method can identify encrypted proxy traffic with high accuracy, up to 99.86%. Besides, the method can timely seek out target flows, as all those behavior features can be obtained in the first packet.**

***Index Terms*—encrypted proxy, traffic identification, behavior feature**

I. Introduction

The encrypted proxy is a special type of network application. In contrast to general proxies, it utilizes proprietary protocol to encrypt and encapsulate communication data between client and proxy server. As a typical example, V2Ray is one of the popular proxies which encapsulate raw traffic with the VMess protocol. This kind of proxy usage has grown dramatically in recent years. Internet users can use proxies to preserve their privacy and expect to bring some level of confidentiality. On the flip side, attackers can utilize it for hiding their identities and anonymize malicious behaviors. In addition, it also can help circumvent Internet censorship to access blocked content. Thus, having the ability to identify encrypted proxies and prevent potential security threats becomes paramount.

Traffic identification is a worth trying detection method for encrypted proxies because of its high accuracy and concealment. It detects proxy in real-time by monitoring network traffic. According to the techniques difference, traffic identification methods can be divided into port-based, deep packet inspection (DPI)-based, and machine learning (ML)-based [1]. Port-based approaches use the port number assigned by the Internet Assigned Numbers Authority (IANA) [2], to identify specific protocols or applications. Methods based on DPI achieve high accuracy by discovering application signatures in traffic payload. However, traditional port-based and DPI-based methods become less effective due to dynamic port assignments and encrypted payloads by encrypted proxies. To solve this problem, machine learning algorithms have been applied to network traffic identification and obtain effective performances. ML-based methods generally consist of three phases: feature extracting and selecting, model designing, and application traffic identifying. The feature extracting and selecting is a crucial step, owing to the quality of features directly affects the model performance. In previous research, ML-based methods often use statistical features such as packet size and inter-arrival time to represent traffic. Although these characteristics can distinguish encrypted proxy traffic from non-proxy traffic to a certain extent, some shortcomings remain. On the one hand, the extraction of statistical features needs to check every packet, which introduces a large computational overhead. It not suitable for real-time identification. On the other hand, these features are easily affected by the network environment. In other words, a model based on a static dataset could be less accurate when works in the open Internet word, on account of the numerous applications and various packet types.

To tackle the issue in the above-mentioned method, we propose a method of encrypted proxy traffic identification with machine learning. Our method utilizes unique behavior features to distinguish proxy traffic, which is more relevant and effective than statistical features. Firstly, we reveal the IP proxy behavior and data encryption behavior of encrypted proxies with an in-depth analysis of the operation principles about it. Subsequently, we use flow relationship, burst, and information entropy for quantization, and extract 6 features from these two aspects. Finally, we construct encrypted proxy traffic identification models with machine learning algorithms. The experiment results show that our method can easily identify encrypted proxy traffic with high accuracy in our dataset.

The main contributions of this paper are as follows:

- Two unique behavior features, IP proxy and data encryption behaviors, are defined to represent a flow. Those features get right to the running mechanism of encrypted proxies, which can be extracted from relevant flows and first TCP payload and used as learning features to identify

978-1-6654-3182-8/21/$31.00 ©2021 IEEE

encrypted proxy traffic.

- A novel real-time identification method for encrypted proxy traffic is given, which extracts behavior features of inspected flows and identify encrypted proxy ones through machine learning models. The method is suitable for real-time identification, without deep packet inspection or long-term statistics of inspected flows.
- Experiments are conducted on a real dataset of V2Ray encrypted proxy traffic to perform comprehensive evaluation of the proposed method. The results show that our method can significantly improve identification performance comparing to widely used methods, with high accuracy up to 99.86%.

The rest of this paper is organized as follows. Section 2 summarizes the related work. Section 3 describes the behavior features of the encrypted proxy and gives the details of our method. Section 4 evaluates our proposed approach using the traffic dataset and various machine learning architectures. Finally, Section 5 presents our conclusions and directions for future works.

II. Related Work

A. ML-based Traffic Identification

Over the last few years, a large and growing body of research on traffic identification pays particular attention to machine learning approaches. These studies commonly extracted one or more categories of features from network traffic, then employed different ML algorithms to construct identification models [3]. The work in [4] reported 249 statistical features of network traffic, which are often used to support ML-based traffic identification. Iglesias et al. [5] defined a time activity vector consist of time-related characteristics to detect various network events. In another major study, Ding et al. [6] proposed a traffic classification method based on expanding vector by analyzing 7 types of relationships between flows. Furthermore, Stöber et al. [7] introduced the traffic burst feature and applied it to discriminate different smartphones. Dorfinger et al. [8] proposed a privacy preserving traffic detection method relayed on the entropy of the first packet payload. Agrawal et al. [9] proposed an optimized feature set to identify P2P traffic.

For the ML model, the work in [10] tested several well-known ML approaches to classify VoIP encrypted traffic with statistical features. Singh [11] evaluated different unsupervised techniques for encountering unlabeled network traffic. Different from the classical machine learning method, deep learning (DL) is an end-to-end method that can directly learn features from raw traffic automatically [12]. Several comparisons between DL-based methods are presented in [13], which aim at mobile application traffic classification. Together, these studies indicate the potential feasibility of ML-based traffic identification.

B. Proxy Detection

Recently, there has been an increasing amount of research on proxy and VPN traffic identification. Foroushani et al. [14] used the C4.5 decision tree to classify different behaviors of Squid proxy and achieved a high detection rate. Janbeglou et al. [15] reported a proxy detection method by analyzing the DNS activity and traffic pattern. In terms of VPN traffic identification, the work in [16] proposed a classification approach with time-related features and ML techniques (C4.5 and KNN). What is far more important is that it generated and published an extensive labeled dataset of VPN and non-VPN traffic. Based on this dataset, work [17]–[21] applied different ML algorithms into the VPN traffic identification and obtained over 90% accuracy respectively.

There are relatively few historical studies in the area of encrypted proxy traffic identification. By using the random forest algorithm, Deng et al. [22] proposed a Shadowsocks traffic detection method with statistical features and got over 85% detection accuracy. In the same vein, Zeng et al. [23] firstly utilized flow content and host behavior to identify Shadowsocks traffic, which achieved up to 93% accuracy on the dataset collected from the campus network. Zhang et al. [24] used convolutional neural networks (CNN) to distinguish 6 types of proxies including Shadowsocks and VMess.

In the above research, ML-based methods usually need to extract features from traffic, directly or indirectly. However, these features have not made the anticipated impact in the real-time identification scenario.

III. Identifying Encrypted Proxies

A. Traffic Identification Framework

Fig. 1 shows the framework of our encrypted proxy traffic identification method, which consists of three steps, data pre-processing, feature extraction, and traffic identification.

Data pre-processing includes two main tasks. First, the raw traffic packets collected from the Internet are divided into flows according to the 5-tuple (source IP address, source port number, destination IP address, destination port number, and the protocol id) with a given timeout. For encrypted proxy traffic in this paper, only TCP flows are considered as most of the encrypted proxy protocols over TCP. All TCP flows compose a sequence according to the start time. Then, by using a sliding window, our method constructs neighbor flows set for each flow within a time window. In the part of feature extraction, the behavior features are computed from two aspects, namely, IP proxy behavior and data encryption behavior. The IP proxy behavior measurement works on the time window flows set, and the data encryption behavior quantification only relies on the first packet. This step is efficient and lower-cost without inspecting all packet content. Finally, on the basis of flow features, our method uses machine learning algorithms to construct models to identify encrypted proxy traffic. In the following subsection, we describe the details of behavior features.

B. Behavior Features

Encrypted proxy is a special type of network application that benefits the exchange of data between end users and network servers. A encrypted proxy system generally consists of a

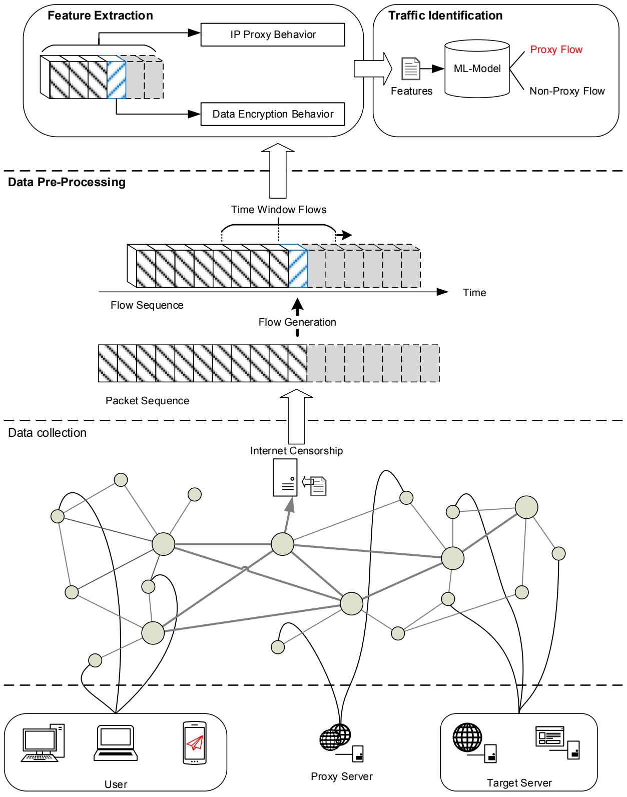
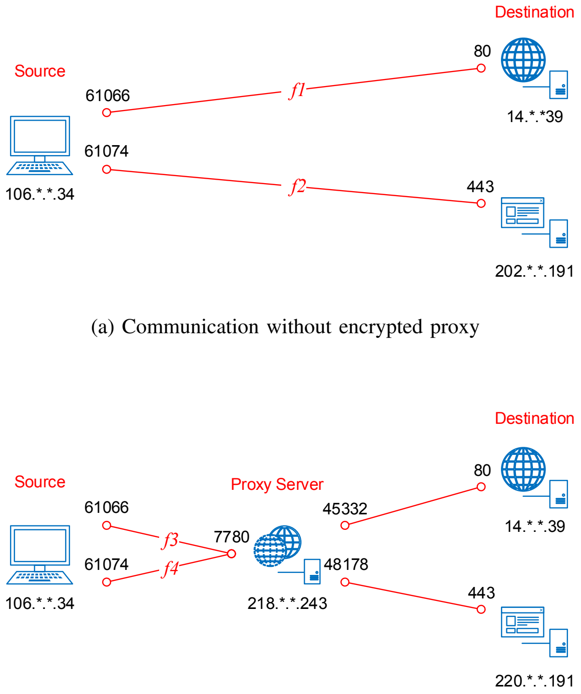


Fig. 1. Encrypted proxy traffic identification framework.



(a) Communication without encrypted proxy

(b) Communication with encrypted proxy

Fig. 2. An example of IP proxy behavior.

local client and a remote server. The proxy server forwards the communication data between the user and the destination server. These data are usually encrypted and obfuscated by specific protocols, such as Shadowsocks used in SS and VMess used in V2Ray. We analyze the mechanism of encrypted proxies, discover unique behavior features that could identify its traffic in real-time with high accuracy.

1) IP proxy behavior: When user access destination servers through encrypted proxies, the proxy server step into the communication process as an intermediate node. The proxy server performs requests on behalf of the client, potentially masking the true origin of the request to the destination server. We term such transmission mode IP proxy behavior. This behavior changes the connection pattern between the end user and the destination server, all packets are transferred by the proxy server. In terms of the source side, it causes flow aggregation and influences the flow correlation within a time window.

As an example demonstrated in Fig. 2, suppose one user access two different services successively in a time window. Normally, two flows are generated, flows f_1 and f_2 in Fig. 2a, which just have the same source IP address. When using encrypted proxies, the source side host connects to the proxy server rather than the destination server. There are also two flows in this process, flows f_3 and f_4 in Fig. 2b. In contrast to the scenario without proxy, these two flows not only have the same source IP address but also the same destination IP address and port.

2) Data encryption behavior: Unlike traditional proxies, encrypted proxies utilize proprietary protocols to encapsulate payload during the session. Once the client establishes a connection by the TCP three-way handshake, it can directly transfer encrypted data to the proxy server. There is no obvious handshake process for key exchange and parameter negotiation. Its traffic appears as encrypted TCP flows without any plaintext headers. However, this communication mode that neither contains a handshake process nor carries plaintext headers is unusual. For instance, it is trivial to distinguish SSH or VPN traffic, as these are some unique fields in their unencrypted protocol handshake messages. We start with the intuition that such data encryption behavior is uncommon on the Internet, extract features to support traffic identification.

C. Feature Extraction

Considering the unique behavior mentioned above, our method constructs a 6-dimensional vector to represent a flow, which composes of features listed in Table I.

On one hand, we compute relevant flows number, destination IP entropy, and flow burst to measure IP proxy behavior based on the time window flows set. The IP proxy behavior changes the connection pattern between the end user and the

TABLE I: Behavior Features

Feature		Description
IP Proxy Behavior	rf_num	Number of flows with the same source IP address, destination IP address and port number.
	dip_etp	Entropy of destination IP address accessed by the source side host.
	fb_num	Number of flow brust within a time windows.
	fb_len	Total length of all flow burst within a time windows.
Data Encrypted Behavior	p_etp	Information entropy of TCP payload of the first packet.
	vc_ratio	Ratio of visible characters in TCP payload of the first packet.

destination server. Given a flow f, the number of the relevant flow of f is calculated as follows:

$$rf\_num = crad(R) \tag{1}$$

$$\begin{aligned} R = (g \mid g.src_{ip} = f.src_{ip} \; and \; g.dst_{ip} = f.dst_{ip} \; and \\ g.dst_{port} = f.dst_{port}, \; g \in F) \end{aligned} \tag{2}$$

where F is the time window flows set of f, function $crad(S)$ provides the number of elements in set S. Simultaneously, we count the destination IP address distribution of all flows that have the same source IP address with f in the time window. Then we can calculate the entropy value with

$$H(X) = -\sum_{x \in X} p(x) \ln p(x) \tag{3}$$

Otherwise, IP proxy behavior results in flow aggregation, flow burstiness is a better way to estimate it. The flow burst refers to a series of time adjacent flows whose interval is less than a threshold. We record the number and total length of the burst within a time window as flow features. To provide an intuitive illustration, an example of flow burst is displayed in Fig. 3. In the time window flows set of flow f, when taking the minimum burst length to 3 and choosing 100 milliseconds as the time threshold (as suggested in [23]), flows f_1, f_2, and f_3 form a burst. Likewise, flows f, f_6, f_7, and f_8 form a burst, $fb\_num = 2$, and $fb\_len = 7$.

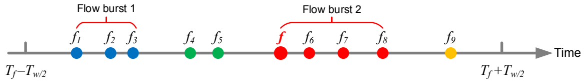


Fig. 3. An example of flow burst.

For data encryption behavior, we use information entropy and visible characters ratio of the first payload to quantify it. Firstly, we acquire the first 16 bytes TCP payload of the first packet from a flow. Then calculating information entropy and percentage of visible characters in bytes. The computation of entropy depends on Formula 3, and the statistical process of visible characters ratio is formulated as:

$$vc\_ratio = \frac{\sum_{i=1}^{n} sgn(128 - b_i)}{n} \tag{4}$$

$$sgn(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases} \tag{5}$$

Combining these two group behavior features, we obtain the flow feature vector. A natural question is whether these features can distinguish encrypted proxy traffic from others. To this end, we select a sample from our dataset in Section IV-A, extract the behavior features from both encrypted proxy and non-proxy traffic. The cumulative distribution function (CDF) results are set out in Fig. 4. It can be seen there are varying degrees of differences in each dimension feature between the two types of traffic. For example, only about 30% of encrypted proxy flows with the number of relevant flows less than 20, however, this value of non-proxy flows exceeds 90%. Similarly, the payload entropy of most encrypted proxy flows is higher than 2.5, while 73% of non-proxy flows have entropy values lower than 2.5. Overall, these comparison results indicate that our method is feasible in identifying encrypted proxy traffic.

IV. Evalution

In this section, we evaluate our method using a traffic dataset, to verify the effectiveness of our behavior features on encrypted proxies traffic identification. After the description of the dataset and metrics, we carry out a comparison on different ML-algorithms and identification methods respectively, then briefly analyze the experimental results. In particular, We show that our method accord with the real-time demand.

A. DataSet

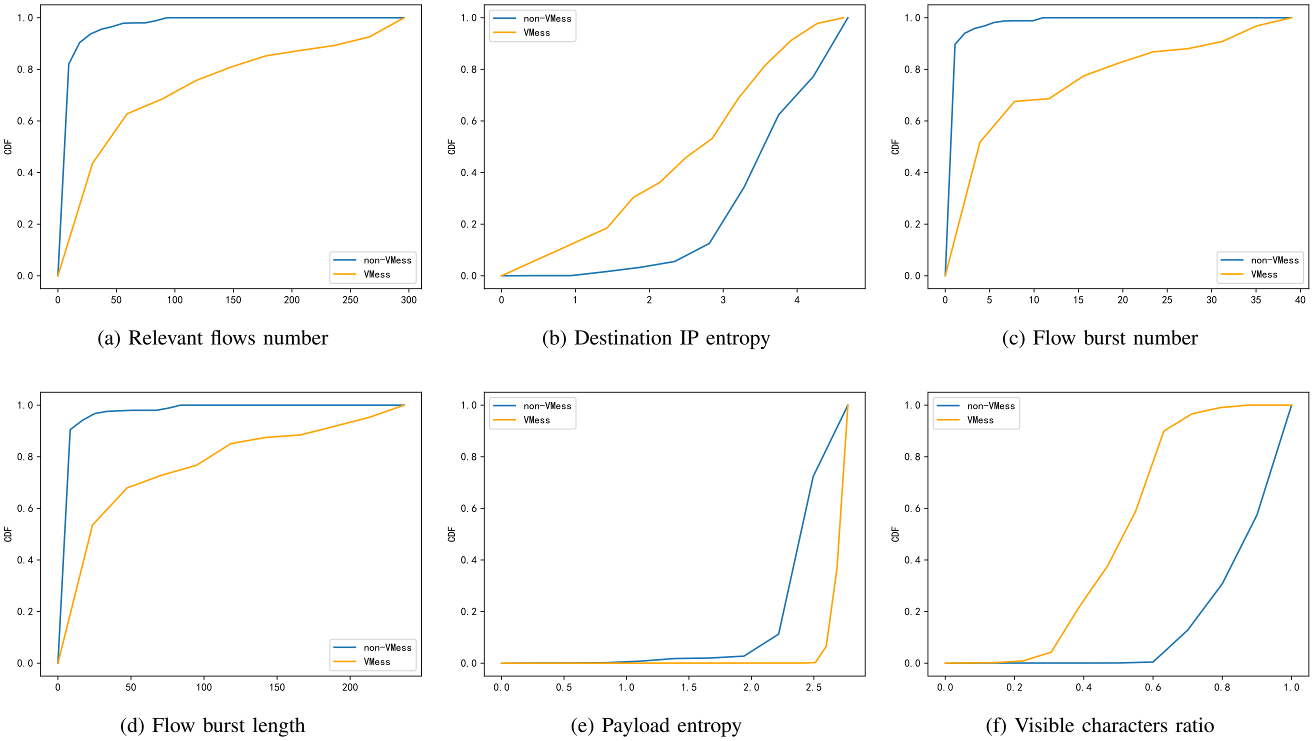
Due to the reasons such as data privacy, there is no public dataset available for encrypted proxy traffic identification. To evaluate our method, we set up an encrypted proxy system by V2Ray and collect a real traffic dataset under local proxy mode, which finally obtains a dataset that includes encrypted proxy traffic and non-proxy traffic. Table II. list the details of our dataset. The ratio of positive and negative samples is about $1:2$.

TABLE II: Dataset Details

Traffic Type	Size(GB)	Flows
VMess	0.54	5202
Non-Proxy	2.63	11502

B. Traffic Identification Result

In our experiments, we use scikit-learn [25] to build traffic identification models. Scikit-learn is an open-source machine learning library with simple and efficient tools for predictive data analysis. We keep its default setting of 5-fold cross-validation during the following tests.



(a) Relevant flows number

(b) Destination IP entropy

(c) Flow burst number

(d) Flow burst length

(e) Payload entropy

(f) Visible characters ratio

Fig. 4. Distribution of features from encrypted proxy and non-proxy traffic.

1) Performance of different models: For the initialization, we set the IP proxy behavior feature extraction time window as 30 seconds. We utilize five well-known ML-algorithms for modeling after extracting behavior features. Table III shows the metrics of different models. The overall results indicate our method can identify encrypted proxy flows well. The precision and recall of KNN, Decision Tree, and Random Forest models are both more than 95%. The tree-structure models obtain better performance, which is consistent with previous studies in network traffic identification. Random Forest achieves the best result with the precision of 99.89%, a little higher than Decision Tree.

TABLE III: Metrics of Diferent Models

ML algorith	Accuracy	Pression	Recall	F1 Score
SVM-rbf	82.10%	86.14%	75.90%	80.70%
Naive Bayes	88.71%	97.58%	79.06%	87.35%
KNN	97.10%	96.95%	97.16%	97.06%
Decision Tree	98.87%	99.67%	98.04%	98.85%
Random Forest	99.03%	99.89%	98.15%	99.01%

In summary, the ML-based identification models achieve high performance, confirm the validity of our behavior features.

2) Variation of time window: The measurement of IP proxy behavior works on a given time window, which is initialized to 30 seconds. In theory, the larger time window is beneficial for identification performance, because it contains a complete set of relevant flows. To verify this, we chose to focus the attention on different time windows. For each time value, we have two different representations as shown in Fig. 5, one corresponds to the Decision Tree result and the other one to the Random Forest.

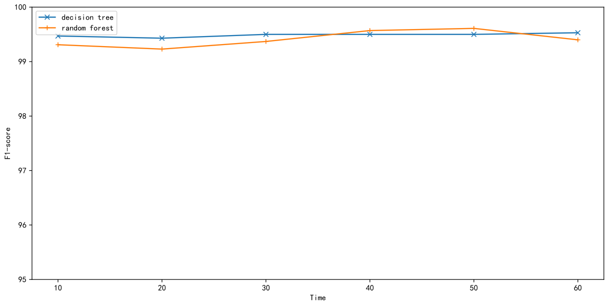


Fig. 5. Performance of different time window.

However, the results reveal that time windows values have little effort on identification performance. Larger time win-dows actually increase the computational complexity. On balance, a time window of 40 seconds is suitable for both identification performance and computational overheads.

3) Comparison with other methods: For further evaluation, we compare our method with three representative methods. We first examine two method mentioned in previous researches. One obtain the side-channel information and construct 2-layer CNN to identify proxy traffic, suggested in [24]. Another method uses flow content and host behavior features as the work in [23]. Besides, statistical features are extensively used for traffic identification. We extract the first 7 packets of a flow, then record packet size and inter-arrival time as flow attributes, in common with most ML-based traffic identfy method. We conduct the experiment on our dataset. Here, except the first method, all methods are performed with the Random Forest model, and the time window of our method is 40 seconds according to the above measure. TABLE IV. shows the results.

TABLE IV: Comparision Results

Method	**Accuracy**	**Pression**	**Recall**	**F1 Score**
2L-CNN	50.06%	89.62%	50.04%	64.22%
Flow-based	96.07%	96.19%	88.05%	91.94%
Packet statistic	96.59%	93.49%	93.05%	93.27%
Behavior-based	99.86%	99.87%	99.68%	99.78%

It can be seen from the table that our method significantly better than the other three methods, with an F1-score of 99.01%, which achieves 5.74%, 7.09% and 34.79% improvement respectively.

4) Real-time performance: Trafic identification ordinarily required to execute online, real-time performance is another important aspect. We analyze the number of processed packets and time consumption of different methods mentioned in the last fraction when identifying a flow. The comparison results are listed in TABLE V. Our method process only one packet of each flow and have a certain time cost of 20 seconds, which is able to operate in real-time.

TABLE V: Real-time Performance

Method	**Processed packets number**	**Time consumption**
2L-CNN	20	Uncertain
Flow-based	1	5min
Packet statistic	7	Uncertain
Behavior-based	1	20sec

V. Conclusion

In this paper, we focus on encrypted proxies detect and propose a novel traffic identify method based on behavior features. We discover the unique IP proxy and data encryption behavior of encrypted proxies. According to these behavior features, we define a 6-dimensional vector to represent a flow. Then we utilize ML-algorithms for modeling to identify encrypted proxy traffic. To evaluate our method, we capture a real traffic dataset by V2Ray. The results confirm the effectiveness of our proposed behavior features in encrypted proxy traffic identification. Compared with existing methods, our method achieves better performance with a low computational overhead, which is suitable for real-time identification.

For future works, we might explore the model update problem in the real network environment to support lifelong identification. Besides, our method detects encrypted proxies in a passive manner by monitoring network traffic, active probing would be a fruitful area for further research.

References

[1] T. T. Nguyen and G. Armitage, “A survey of techniques for internet traffic classification using machine learning,” *IEEE communications surveys & tutorials*, vol. 10, no. 4, pp. 56–76, 2008.

[2] M. Cotton, L. Eggert, J. Touch, M. Westerlund, and S. Cheshire, “Internet assigned numbers authority (iana) procedures for the management of the service name and transport protocol port number registry.” *RFC*, vol. 6335, pp. 1–33, 2011.

[3] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, “Towards the deployment of machine learning solutions in network traffic classification: A systematic survey,” *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1988–2014, 2018.

[4] A. Moore, D. Zuev, and M. Crogan, “Discriminators for use in flow-based classification,” Tech. Rep., 2013.

[5] F. Iglesias and T. Zseby, “Time-activity footprints in ip traffic,” *Computer Networks*, vol. 107, pp. 64–75, 2016.

[6] L. Ding, J. Liu, T. Qin, and H. Li, “Internet traffic classification based on expanding vector of flow,” *Computer networks*, vol. 129, pp. 178–192, 2017.

[7] T. Stöber, M. Frank, J. Schmitt, and I. Martinovic, “Who do you sync you are? smartphone fingerprinting via application behaviour,” in *Proceedings of the sixth ACM conference on Security and privacy in wireless and mobile networks*, 2013, pp. 7–12.

[8] P. Dorfinger, G. Panholzer, B. Trammell, and T. Pepe, “Entropy-based traffic filtering to support real-time skype detection,” in *Proceedings of the 6th International Wireless Communications and Mobile Computing Conference*, 2010, pp. 747–751.

[9] S. Agrawal and B. S. Sohi, “Feature optimization and performance evaluation of machine learning algorithms for identification of p2p traffic,” *Journal of Advances in Information Technology*, vol. 3, no. 2, pp. 107–114, 2012.

[10] R. Alshammari and A. N. Zincir-Heywood, “Identification of voip encrypted traffic using a machine learning approach,” *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 1, pp. 77–92, 2015.

[11] H. Singh, “Performance analysis of unsupervised machine learning techniques for network traffic classification,” in *2015 Fifth International Conference on Advanced Computing & Communication Technologies*. IEEE, 2015, pp. 401–404.

[12] S. Rezaei and X. Liu, “Deep learning for encrypted traffic classification: An overview,” *IEEE communications magazine*, vol. 57, no. 5, pp. 76–81, 2019.

[13] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, “Mobile encrypted traffic classification using deep learning,” in *2018 Network traffic measurement and analysis conference (TMA)*. IEEE, 2018, pp. 1–8.

[14] V. Aghaei-Foroushani and A. N. Zincir-Heywood, “A proxy identifier based on patterns in traffic flows,” in *2015 IEEE 16th International Symposium on High Assurance Systems Engineering*. IEEE, 2015, pp. 118–125.

[15] M. Janbeglou and N. Brownlee, “Identifying tunnelled proxies through passively monitoring network traffic,” in *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 2016, pp. 63–69.

[16] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, “Characterization of encrypted and vpn traffic using time-related,” in *Proceedings of the 2nd international conference on information systems security and privacy (ICISSP)*, 2016, pp. 407–414.

[17] S. Bagui, X. Fang, E. Kalaimannan, S. C. Bagui, and J. Sheehan, “Comparison of machine-learning algorithms for classification of vpn network traffic flow using time-related features,” *Journal of Cyber Security Technology*, vol. 1, no. 2, pp. 108–126, 2017.

[18] J. A. Caicedo-Muñoz, A. L. Espino, J. C. Corrales, and A. Rendón, “Qos-classifier for vpn and non-vpn traffic based on time-related features,” *Computer Networks*, vol. 144, pp. 271–279, 2018.

[19] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, “End-to-end encrypted traffic classification with one-dimensional convolution neural networks,” in *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE, 2017, pp. 43–48.

[20] L. Guo, Q. Wu, S. Liu, M. Duan, H. Li, and J. Sun, “Deep learning-based real-time vpn encrypted traffic identification methods,” *Journal of Real-Time Image Processing*, vol. 17, no. 1, pp. 103–114, 2020.

[21] A. Saber, B. Fergani, and M. Abbas, “Encrypted traffic classification: Combining over-and under-sampling through a pca-svm,” in *2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*. IEEE, 2018, pp. 1–5.

[22] Z. Deng, Z. Liu, Z. Chen, and Y. Guo, “The random forest based detection of shadowsock’s traffic,” in *2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, vol. 2. IEEE, 2017, pp. 75–78.

[23] X. Zeng, X. Chen, G. Shao, T. He, Z. Han, Y. Wen, and Q. Wang, “Flow context and host behavior based shadowsocks’s traffic identification,” *IEEE Access*, vol. 7, pp. 41 017–41 032, 2019.

[24] Y. Zhang, J. Chen, K. Chen, R. Xu, J. Teh, and S. Zhang, “Network traffic identification of several open source secure proxy protocols,” *International Journal of Network Management*, p. e2090, 2019.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.