Hindawi
Security and Communication Networks
Volume 2018, Article ID 6124160, 15 pages
https://doi.org/10.1155/2018/6124160

Research Article

Fingerprinting Network Entities Based on Traffic Analysis in High-Speed Network Environment

Xiaodan Gu, Ming Yang, Yiting Zhang, Peilong Pan, and Zhen Ling

School of Computer Science and Engineering, Southeast University, Nanjing, China

Correspondence should be addressed to Ming Yang; yangming2002@seu.edu.cn

Received 24 August 2018; Accepted 28 October 2018; Published 16 December 2018

Guest Editor: Yuan Yuan

For intrusion detection, it is increasingly important to detect the suspicious entities and potential threats. In this paper, we introduce the identification technologies of network entities to detect the potential intruders. However, traditional entities identification technologies based on the MAC address, IP address, or other explicit identifiers can be deactivated if the identifier is hidden or tampered. Meanwhile, the existing fingerprinting technology is also restricted by its limited performance and excessive time lapse. In order to realize entities identification in high-speed network environment, PFQ kernel module and Storm are used for high-speed packet capture and online traffic analysis, respectively. On this basis, a novel device fingerprinting technology based on runtime environment analysis is proposed, which employs logistic regression to implement online identification with a sliding window mechanism, reaching a recognition accuracy of 77.03% over a 60-minute period. In order to realize cross-device user identification, Web access records, domain names in DNS responses, and HTTP User-Agent information are extracted to constitute user behavioral fingerprints for online identification with Multinomial Naive Bayes model. When the minimum effective feature dimension is set to 9, it takes only 5 minutes to reach an accuracy of 79.51%. Performance test results show that the proposed methods can support over 10Gbps traffic capture and online analysis, and the system architecture is justified in practice because of its practicability and extensibility.

1. Introduction

With the rapid development and widespread application of computer networks, mobile communications, smart devices, and the Internet of Things technology, cyberspace is becoming more and more integrated into people's social life. People can access services through various devices anytime and anywhere, thereby realizing the interconnection between people and people, people and things, and even things and things. However, while cyberspace brings a lot of convenience to people, network attacks such as DDoS attacks, worm attacks, information theft, and cyber fraud have become increasingly severe. Therefore, it is imperative to effectively protect against cyber threats.

The intrusion detection system is used to monitor a network or system, which can identify malicious activities or policy violations from both inside and outside intruders. As an important and dynamic research area, the network intrusion detection technology can identify malicious activities by monitoring and analyzing inbound and outbound traffic [1, 2]. But there is less work that can effectively identify potential threats if an intruder has no abnormal activity. To address this issue, we introduce the identification technologies of network entities to detect the intruder with no abnormal activity, which mainly consist of device identification and user identification. The basic idea is that if we detect unauthorized devices or unauthorized users using authorized devices, we can indicate that a network intrusion may be taking place.

However, the conventional identification technologies of network entities are usually based on explicit identifiers. For example, user devices are always identified by the MAC addresses or browser cookies, and the user is identified by intercepting and verifying his account information in the network traffic through the Deep Packet Inspection (DPI). These explicit identifiers can be easily hidden or tampered, causing an identification failure. In response to this problem, the researchers have demonstrated [3] that the absence of explicit identifiers brings no harm as long as well-chosen implicit

identifiers are reasonably combined. The potential selections are instantiated as the SSID information in the 802.11 active probe frame, the plug-in installed in the browser, the font library in the system, etc. Although these implicit identifiers cannot uniquely identify a network entity individually, they are hard to be concealed because they usually reflect the user's personalized configuration, historical behavior records, or subtle differences between entities. Therefore, in practice, a combination of the implicit identifiers is usually utilized to generate a fingerprint for device identification and user tracking. The fingerprinting technology based on implicit identifiers is essentially a method of traffic analysis and side-channel attack. Although its effectiveness has been initially verified by existing work, there are still many problems to be solved in practical application, including the selection of efficient features, realization of real-time processing of network traffic in a high-speed network environment, and quick identification of devices and users in a short period of time.

In view of the problems mentioned above, we use PFQ kernel module to realize high-speed capture of network packets and use Storm, a well-known distributed real-time streaming data processing technology, to realize online analysis of network traffic. Based on these, a device identification method based on runtime environment analysis and a network user identification method based on behavioral fingerprinting are proposed separately. The main contributions of the paper are as follows:

(i) A distributed traffic analysis framework for high-speed network environments is designed. The framework uses the PFQ kernel module to implement packet capture, Kafka for packet distribution, and Storm for packet content analysis and information extraction of applications, operating systems, HTTP User-Agents, domain names, and Web access records.

(ii) An online device identification technology based on the analysis of user device operating environment is proposed. This technology selects 961 features such as applications, operating systems, and HTTP User-Agent fields to constitute the fingerprints of the devices, while a variety of offline classification models are trained and verified. Finally we select the logistic regression algorithm to identify the user device in a sliding window manner.

(iii) In order to realize network user identification, Web access records, domain names, and HTTP User-Agent fields are selected to constitute user behavioral fingerprints according to the user's network behavior habits. These fingerprints, containing a total of 57593 feature columns, are trained and verified by the two offline classification models using machine learning, which are Naive Bayes and random forest models. By comparison, the Multinomial Naive Bayes model is found to outperform the random forest model, so it is chosen as the classification algorithm for identifying online users in a sliding window manner.

(iv) In order to achieve high-efficiency identification, the impacts of different time window sizes on the recognition rate are tested. Specifically, the accuracy rate of device identification reaches 77.03% within 60 minutes, and 79.51% of user identification accuracy can be achieved within 5 minutes. The packet capture rate, distributed processing speed, and online recognition response speed are also evaluated to verify the practicability of the proposed identification technology.

An early version of the network entities identification is presented in [4]. In the early version, we design a distributed high-speed traffic analysis framework to recognize devices based on runtime traffic. In this journal version, to deal with the cross-device scenario, we further analyze the user's network behavior habits and generate fingerprints to identify the network users. We also evaluate the identification performance difference between the Boolean and numerical type fingerprints in this extended version.

The rest of this paper is organized as follows. In Section 2 we overview the related work. In Section 3 we describe the overall design of our network entities fingerprinting technology. Section 4 introduces a distributed traffic analysis framework for high-speed network environments. In Sections 5 and 6, we present the details of the identification technologies of network device and user, respectively. Section 7 tests the performance of the identification technologies. Finally, the paper is concluded in Section 8.

2. Related Work

The identification of network users and that of devices are two differentiated research directions but are closely related. Earlier device identification technologies mainly obtain the information of the hardware, operating systems, network protocols, and other parameters by collecting and analyzing the physical signals or traffic generated by the device. For example, in the physical layer [5], the TCP packet time stamps are analyzed to obtain the clock skew [6], and the Ethernet frames are analyzed to obtain the differences among the analog signals of different devices [7]. While in the operating system layer, the active scanning algorithm used by the wireless device driver might be inferred by analyzing the interval time of 802.11 probe request frames [8]. In the application layer, the User-Agent field, IP address, browser cookie, user login ID, and other identity information are extracted through the traffic analysis in clear text [9]. The interval time, number, direction, and other attributes of the encrypted wireless packets are analyzed for the distinction of different terminal applications [10]. Other researches have applied different threat models to achieve the identification of devices, e.g., the device recognition based on browsers [11–17] and that based on mobile applications [18, 19].

The above-mentioned identification technologies are merely, in essence, the identification of a single browser [20] or a single terminal device. They are far from being capable enough to identify the user's cross-device activities. For example, in the scenario of intrusion detection, if some

intruder occupies an authorized device, we cannot detect the intrusion by using the device identification technologies. So it is necessary to carry out the research on the user identification technology based on behavioral fingerprints.

Essentially, user identification technologies based on behavioral fingerprinting are biometric, which use the inherent physiological characteristics or behavioral characteristics of the human body for identification. They can be categoried into two types. The former one has been widely used by employing the characteristics of human body parts, such as fingerprint identification, face identification, and iris identification.

The behavior-based identification technology [21] extracts the features for identification with the information of the user's operation skills, knowledge, styles, preferences, and strategies revealed in behaviors. For example, researchers have found that different users differentiate from each other in moving, clicking, dragging, and releasing the mouse [22]. Some may be different in key stroking when keyboarding [23]. All of these differences can help to extract fingerprints for effective identification. In the network area, users have different behavioral patterns for network access due to different preferences, habits, etc. Different behavior patterns lead to different traffic flows. Therefore, researchers believe that the network traffic generated by the user can be regarded as biometric for user identification [24].

In order to identify users based on network traffic, early solutions are implemented by extracting explicit identifiers such as IP addresses or MAC addresses [3]. However, such method based on explicit identifiers is not reliable for that it will fail once the user changes the IP address or devices. So far, the dynamic address allocation scheme adopted by ISP makes users change the IP more frequently. To address this issue, the researchers apply the behavior fingerprinting technology to user identification based on network traffic. Padmanabhan et al. [25] find that different users may have different behaviors when browsing the same website. By analyzing the real data, they extract the users' clickprints to generate behavioral fingerprints. Pang et al. [3] propose to explore the destination address, network name, 802.11 option configuration, and broadcast frame length so as to identify the user from the perspective of protocol and user preferences. This is actually a comprehensive application of user-related and device-related implicit identifiers.

Yang [26] uses data mining techniques on the Web browsing dataset in order to mine association rules for each user's behavior, proposes three strength evaluation criteria based on support and lift to generate fingerprints, and finally calculates the distance between fingerprints for identification. Kumpošt et al. [27] believe that the websites visited by the user and the corresponding frequencies, which reflect individual preferences, can be identified as a behavioral fingerprint. They store the source IP, destination IP, and frequency in a two-dimensional matrix and perform the inverse document frequency and cosine similarity algorithm to identify users. Similarly, Herrmann et al. [28] extract user's destination domain names and the corresponding visiting frequency to derive the behavioral fingerprints and use Multinomial Naive Bayes classifier to classify them. Experiments are conducted on a dataset containing HTTP traffic generated by 28 volunteers, and a 73% accuracy rate is obtained.

Since the data set used in the experiments [28] is not big enough to prove the feasibility of the method, the author conducts a larger-scale experiment in the later work [29]. He tests a dataset containing more than 2,100 users' DNS requests, uses the cosine similarity algorithm to filter the noise data, and finally obtains an accuracy of 88%. In addition, Herrmann et al. [30] also compare and evaluate three classification methods through a large number of experiments, including the Multinomial Naive Bayes classifier, the nearest neighbor algorithm, and association rules mining technology. Kim et al. [31] get the user's behavioral fingerprints based on DNS traffic by analyzing the domain name, the sequence of domain names, and the requested periods. Gu et al. [32] infer the users' preferences through the semantic analyses of search records and achieve an accuracy of 93.79% on the dataset of 509 network users.

Overall, the current device and user identification researches have some defects. For example, only a few features are explored and the identification effect is prone to jitter. In addition, the existing studies are not time-efficient enough as it needs to aggregate a whole day's traffic as a fingerprint.

To avoid the aforementioned problems, we adopt the distributed processing technology to extract information such as applications, operating systems, HTTP User-Agent, domain names, and Web access records in real time from high-speed network traffic. Then we propose two online identification methods based on the runtime environment and the behavioral fingerprints, respectively, which are made possible in the pattern of sliding windows. In addition, we also focus on testing the impact of different traffic window sizes on the identification rate and thereby prove the high efficiency and practicality of the proposed technologies.

3. Overall Design of Fingerprinting

The overall design of our network entities fingerprinting technology is shown in Figure 1. The initial step leverages the PFQ-based high-speed packet capture module to capture high-speed network traffic and then forwards the packets to distributed high-speed network traffic processing modules through the Kafka message queue. Then the processing module parses the message content, extracts the relevant feature data, and stores it in the HBase. Finally, the Spark-based online identification module periodically reads the feature data from the distributed database and realizes the online identification of network devices and users by employing the machine learning algorithms in a sliding window mechanism. The working mechanism and functions of each module are specified as follows:

(i) **PFQ-based high-speed packet capture module.** Configure a mirror port on the switch or router, or use an optical splitter to mirror the traffic to a data distribution server. The high-speed packet capture module on this server achieves highly efficient packet capture by adopting the memory mapping mechanism, zero

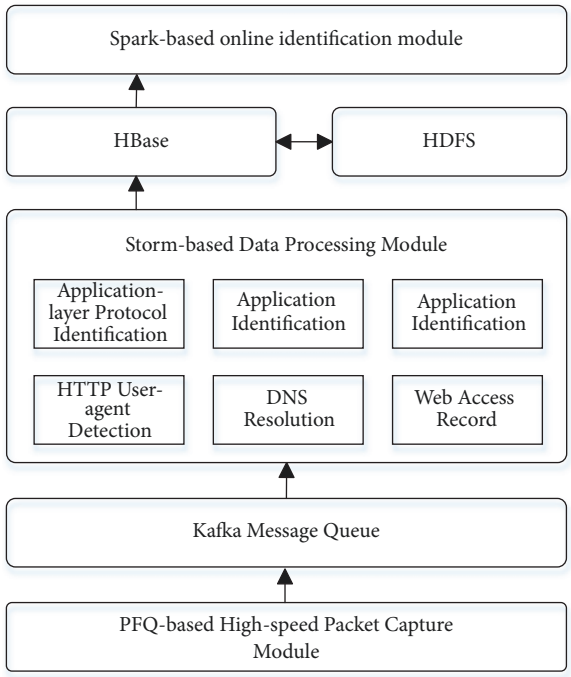


Figure 1: Overall design of network entities identification.

copy technology, and double-buffering mechanism based on the PFQ Linux kernel module.

(ii) **Kafka message queue.** The distributed message queue is a data transmission channel between the packet capture module and the distributed processing module. More specifically, it is a buffer zone for coordinating the producer and consumer. We use Kafka to implement distributed message publish, which has a high linear extensibility in adapting to the high-speed data transmission scenario.

(iii) **Storm-based data processing module.** The distributed data processing module, as a core functional module, undertakes all processing and analysis tasks for network traffic, including parsing of network messages, identification of application protocols, identification of applications, and finally extracts and stores data of applications, operating systems, Web access records, domain names, and HTTP User-Agent fields. We realize distributed streaming data processing based on the Storm platform, which can achieve high-speed data reading combined with Kafka queues and can achieve high-speed data writing combined with HBase. Meanwhile, data transmission performance between the internal components of Storm is also very efficient.

(iv) **HBase.** The online identification module is proposed to read and analyze the data periodically. Thus, as a distributed column-oriented database, the HBase functions as a data medium between the distributed data processing module and the online identification module.

(v) **Spark-based online identification module.** Our identification module is based on the Spark platform and designed to fit into two scenarios. For the device identification scenario, the module extracts device features about runtime environment to generate the fingerprints. For the cross-device identification scenario, the user behavior data are collected to implement fingerprinting for network users. The relevant feature data are read from the HBase distributed database, and the machine learning algorithms in Spark MLlib along with the sliding window mechanism are employed to identify the network devices and users online.

4. Collection and Distributed Processing of High-Speed Network Traffic

4.1. Collection of High-Speed Network Traffic. Compared to Pcap, which is a traditional packet sniffing toolkit, PFQ is a better designed network packet capture framework customized for the optimization of multicore CPUs and multihardware queue network interfaces. It is primarily used for efficient packet capture and transmission on Linux. In its internal implementation, PFQ eliminates the cost of copying packets from kernel space to user space by adopting a memory mapping mechanism, and performs concurrency operations of user-space applications and PFQ kernel packet grabbing programs on the buffers by means of double-buffering technology. The core components of PFQ fall into three parts: packet extracting program, packet forwarding module, and socket queue. First, the packet extraction program directly obtains the packets from the network interface card (NIC) driver and transfers them to the batch queue. Then, the packet forwarding module selects the socket and sends packets to the user-space applications.

After the packets are captured, librdkafka is used to write them into the Kafka message queue. In this paper, we decouple the high-speed packet capture module and the Kafka message queue through the open source project Blockmon [33].

4.2. Distributed Processing of Network Traffic. When the captured packets are written into the Kafka message queue, we implement distributed analysis and processing of packets based on the Storm platform. The following are the key concepts related to the Storm:

(i) Topologies: the logic of the application which defines various components and the ways of communication between them.

(ii) Streams: data flows consisting of message tuples transmitted between Storm components. All Streams are parallel transferred in a distributed way.

(iii) Spouts: data sources. Usually, a Spout reads messages from an external data source and transfers them into the Topology in the form of tuples.

(iv) Bolts: Storm processing units. Each Bolt completes one or more processing tasks and is responsible for

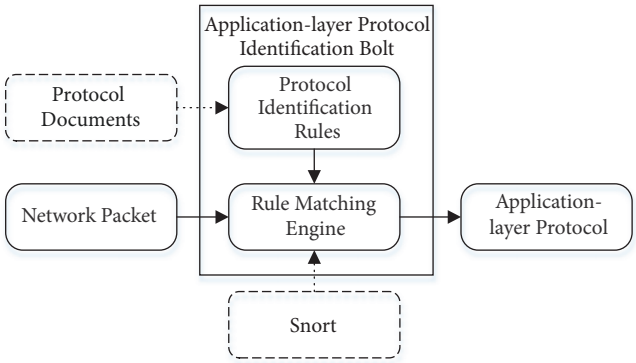


Figure 2: Application layer protocol identification bolt.

transmitting the processing results to the external system for storage or display.

4.2.1. Input of Flow Data. Data transmission between Spouts and Bolts, as well as that among Bolts, is in the form of Streams, while Spouts obtain data from external sources in different ways. In this paper, we use KafkaSpout to read packets from the Kafka message queue and transmit tuples to the packets parsing Bolts.

In the distributed processing environment, KafkaSpouts retrieve data from Kafka partitions in parallel on all slave nodes. And the degree of parallelism has a crucial influence on the throughput of the system. Meanwhile, it is affected by the number of Kafka partitions, because each Kafka partition can only be consumed by one KafkaSpout. Namely, the key to improving Storm throughput is to increase the number of Kafka partitions.

4.2.2. Packets Analyzing and Filtering. The inputting packet tuples are analyzed and filtered to obtain the content of messages, and the header information in each layer of the network protocol is extracted, including PFQ pkthdr header, Ethernet frame header, IP header, TCP header, and UDP header. In the process of packet analysis, several rules are set to filter packets unrelated to network device identification, such as the network control protocol packets and routing protocol packets to improve the processing efficiency of the system.

4.2.3. Application Protocol Identification. Traditionally, in order to identify application protocols, the port numbers of the captured packets are always matched with the well-known ones. Such method is deficient due to a high false positive rate. Therefore, we aim to improve the identification accuracy by employing DPI technology to analyze the packet payloads. Although such method is less efficient, its accuracy rate is significantly higher than that of the former. The module of application protocol identification is developed in two phases: (a) designing the rule matching engine and (b) writing protocol identification rules. The first step extracts the rule matching engine of the Snort core components and introduces multithreading. Then, based on referencing protocol documents, we summarize the characteristics of a typical protocol and write an appropriate rule according to the writing specification of Snort rules.

```
alert udp $EXTERNAL_NET any -> $HOME_NET any
(msg: "snmp"; content: "|30|"; offset:0; depth:1;
byte_test:1,<,0x80,1; content: "|02|"; offset:2; depth:1;
sid:70; rev:1;)
```

Box 1

The process of identifying application protocols is shown in Figure 2. The rule matching engine generates a rule tree according to the specified protocol identification rules and performs rule matching for network traffic. When a match occurs, it indicates that the packet is identified as a specific type of protocol.

Box 1 shows an identification rule for the SNMP protocol and corresponding explanation ensues.

This statement indicates that the rule is released as the first version, with the id of 70. All the UDP packets sent from any port or IP address are detected without exception. According to the identification rule, the first byte value of the application layer payload is 0x30. The second byte value must be less than 0x80, and the third is 0x02. When the match is successful, the alert action is triggered and the protocol type value SNMP is returned to the caller.

After a review of various typical protocol documents, we have completed the writing of recognition rules for 25 typical application layer protocols such as BITTORRENT, DNS, DROPBOX, HTTP, SMTP, and SSH.

4.2.4. Application Identification. After the identification of the application protocol, we further figure out the type of application that generates the packets. As we all know, apart from the traffic generated by the user interaction, the background process of the application communicates with the server periodically, thereby generating more traffic. We analyze the traffic and extract various features to identify the applications.

The data transmission between an application and the server is generally divided into two cases. First, the application uses a custom data transmission protocol, such as the OICQ protocol, which is designed by Tencent and is merely used as the data transmission protocol of QQ. In this case, as long as the application protocol has been identified, the application is sure to be identified. Second, multiple kinds of applications share an application layer data transmission protocol, such as the HTTP protocol, to encapsulate data transmitted between the client and the server. For this situation, we distinguish different applications by extracting multiple field values from the traffic. For example, in the HTTP protocol, the HOST field represents the combination of the domain name and the port number of the server address. Usually, the addresses and corresponding domain names of applications devised by different companies are not identical. Even though different applications provided by the same company share the same server, the addresses are still distinguished from each other with different HTTP request parameters. Therefore, the combination of the HOST field

of HTTP protocol and request parameters can be used to identify different applications.

This paper observes and analyzes the application traffic in the experimental network and summarizes a collection of 116 commonly used application identification rules under 21 categories, such as browser, e-mail, remote management, online game, instant messaging, social networking, Web disk, input tool, online video, P2P video, and stock software. These identification rules cover traffic generated from users' clicks, login activities, automatic updates, and background process communications.

4.2.5. HTTP User-Agent Detection. As the name suggests, the User-Agent field in HTTP traffic contains the user agent information. Generally, the User-Agent field generated by a browser contains the information like the types and versions of the browser and the operating system. As an important piece of information in the User-Agent field, a specific operating system has its own structure-mapping rules, diversifying the types of all the existing operating systems. For instance, the prefixes of the Windows operating systems are normally Windows NT, and suffixes represent specific operating system versions. The identification rules of the operating systems presented in this paper cover the mainstream operating systems such as Windows, Mac OS, OpenBSD, and Ubuntu.

4.2.6. DNS Resolution and Web Access Records. The user generates a large number of HTTP requests and DNS requests when he manipulates the applications or accesses websites using the browser. The destination IP address in an HTTP request together with the corresponding time information can reflect the behavioral characteristics of the user to some extent. And the DNS responses can help us to associate multiple IP addresses of the same domain name together. The resolution of DNS packets is mainly for the IPv4 protocol. From the response packets, we can extract the <domain name, address> pairs. Specifically, the Questions field indicates the number of requested domain names, and the number of corresponding addresses is contained in the Answer RRs field. There are many kinds of DNS response types, which are distinguished by the Type field. For example, the A record maps a domain name to the corresponding IP address while the CNAME record maps an alias name to a canonical domain name.

4.3. Distributed Storage of Network Traffic Data. Based on the processing and extraction of the packets, the information of the extracted applications, operating systems, HTTP User-Agent, DNS, and Web access records is stored in the corresponding columns of the distributed database HBase. By reading data from HBase, the device identification and user identification are implemented.

5. Device Identification Based on Runtime Environment Analysis

5.1. Basic Ideas. For the first device identification scenario, we propose a novel identification method based on runtime environment analysis. Its basic idea is to realize the unique identification of devices based on the combination of the device operating system, the HTTP User-Agent information, and especially the installed applications.

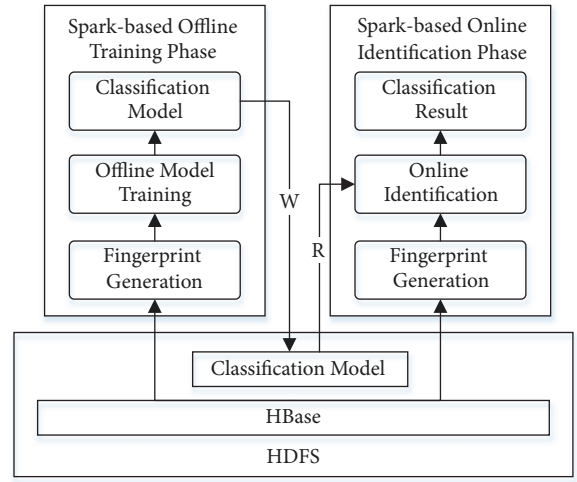


Figure 3: Two-stage fingerprint recognition.

As shown in Figure 3, the identification process can be divided into two phases, namely, the Spark-based offline training phase and the Spark-based online training phase. In the offline training phase, Spark distributedly reads the relevant features from the HBase, generates the corresponding device fingerprint denoted as a vector, and accordingly learns an appropriate classification model by offline training. The dataset needed for offline training and verification can be labeled with IP addresses (MAC addresses can be utilized for labeling in LANs instead). In the online identification stage, distributed analysis and feature extraction are performed on the real-time traffic. And the generated fingerprint vector is classified by the offline training model. Finally, the classification results indicate the identity of the device.

5.2. Feature Selection and Fingerprint Generation. This section focuses on the operating environment of user devices and derives a device identification technology based on its characteristics. The operating environment mainly includes two types of features, which are the operating system type and application type (version information is included). To generate the device fingerprints for identification of the user device, we collect the type and version information of applications from the results of application identification, and extract the attributes such as the browser type and version and operating system type from the results of HTTP User-Agent detection.

Specifically, the device fingerprints are generated by analyzing the traffic per unit time. If the traffic of an application is detected within the time period, the corresponding feature attribute of the application is set to 1 or the corresponding frequency. The dimension of the device fingerprint feature vector is 961. According to the value types of the extracted feature attributes, we design two types of device fingerprints: boolean type device fingerprint and numerical type

device fingerprint, where the Boolean type device fingerprint indicates whether features such as applications or operating systems appear in the network traffic, and the numeric device fingerprint indicates how often these features appear.

Note that all the identifiable application type sets are $S = \{S_1, S_2, \cdots, S_N\}$ and the ith application's version set is $V_i = \{V_i^1, V_i^2, \cdots, V_i^{C_i}\}$. C_i refers to the total number of recognizable versions of the ith application and OS represents the operating system type. The feature vector of the device fingerprint can be presented by formula (1).

$$\overrightarrow{FP_{dev}} = \left\{ \underbrace{S_1, S_1V_1^1, \cdots, S_1V_1^{C_1}}_{S_1}, \underbrace{S_2, S_2V_2^1, \cdots, S_2V_2^{C_2}}_{S_2}, \cdots, \underbrace{S_N, S_NV_N^1, \cdots, S_NV_N^{C_N}}_{S_N}, OS \right\} \tag{1}$$

When the value of the attribute in fingerprint $\overrightarrow{FP_{dev}}$ is numerical, it is a numerical type device fingerprint. When the value of the attribute is only 0 or 1, it is a Boolean type one. Then the device identification problem can be modeled as a multiclassification problem in machine learning.

5.3. Offline Model Training and Verification. Since the efficiency of identification depends greatly on the classification algorithm and the dimension of the device fingerprint vector is relatively small, the multiclassification algorithm can generally be used to train the identification model. We compare the classification effects of Multinomial Naive Bayes algorithm, random forest algorithm, and the logistic regression algorithm. After that, the best performed algorithm is selected for online identification.

We collect network traffic of 118 user devices for 53 days from June 1st to July 23rd, 2016. The network traffic produced on each device is examined per hour. Based on the examination, the features are extracted to form a fingerprint (all zero-vector fingerprints are discarded). Then we get 50,305 valid device fingerprints in total. The data collected in the first 30 days is used to train and verify the offline model, including 30,148 records, while the remaining 20,157 records gathered in the following days are used to evaluate the accuracy of the classification model.

During the offline training process, the device fingerprints in the data set are randomly divided into two subsets, one being the training set containing 70% fingerprints and the other being the verification set containing the remaining 30%. Above all, the classification model of Boolean type device fingerprinting is trained and verified as follows.

5.3.1. Training and Verification of the Classification Model of Boolean Type Device Fingerprinting. First, the random forest classification model is trained. Different from Multinomial Naive Bayesian and logistic regression, the random forest classification model has two parameters that need to be tuned, i.e., the number of decision trees (*nums*) and the maximal depth of the decision tree (*depth*). The parameter of *nums* affects the accuracy of the overall classification, while *depth* affects the classification accuracy of each decision tree. Training and testing are performed under different values of *nums* and *depth*, and the obtained classification accuracy is shown in Figure 4(a).

As can be seen from Figure 4(a), *depth* generally has a greater impact on the classification accuracy. With the increase of *depth*, the accuracy is significantly improved. When the value of *depth* is 30, the classification turns to be optimal. The effect of *nums* on the classification accuracy correlates positively to *depth*: when *depth* is small, the accuracy rate rises with the increase of *nums*; when the *depth* is larger, the classification accuracy rate first goes up with the increase of *nums*, and then remains stable when *nums* is greater than 20. When the value of *nums* is 150, the classification accuracy is the highest. Therefore, we set the value of *nums* as 150 and the value of *depth* as 30, respectively, for optimization.

Then, the Multinomial Naive Bayesian and logistic regression classification models are trained separately, and the classification accuracy of the models is evaluated by the verification set and test set, respectively. Figure 5(a) shows the classification accuracy of the three models, where MNB refers to Multinomial Naive Bayes, RF denotes random forest, and LR represents logistic regression. From Figure 5(a), it can be seen that the classification accuracy of the verification set by performing the logistic regression algorithm is considerably higher than that of other algorithms. For the same algorithm, the classification effect of the test set is significantly lower than that of the verification set. This is because the data in the training and verification set is randomly segmented, and data in the test and training set has a time-series relationship. Moreover, the operating environment of the device is likely to change, so the accuracy of the device identification may gradually decrease over time.

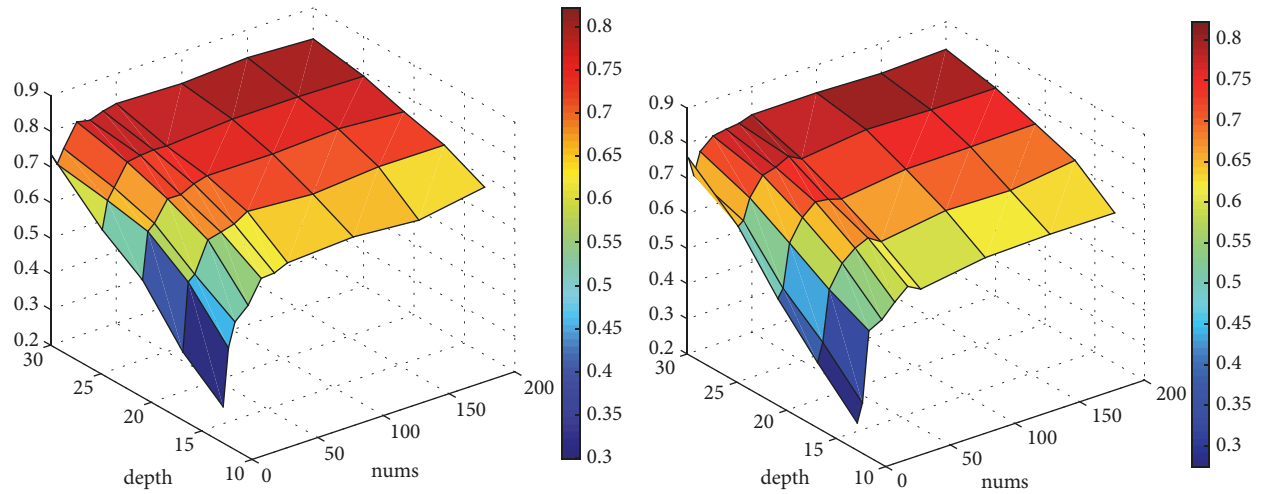
Further analysis of the data reveals that a portion of records in the device fingerprint vectors stay close to the full value of 0. This is due to the fact that not all device traffic is generated by the identifiable applications involved in this paper. Such traffic cannot be identified as valid device information. To deal with it, the following definitions are given.

Definition 1 (effective dimension). Given a fingerprint vector, denote the number of feature columns with a nonzero value as effective dimension.

Definition 2 (the minimal effective dimension). For the set of fingerprint vectors to be identified, the minimal effective dimension is defined as the threshold, below which the fingerprints are deemed as invalid ones and filtered out due to a lack of information.

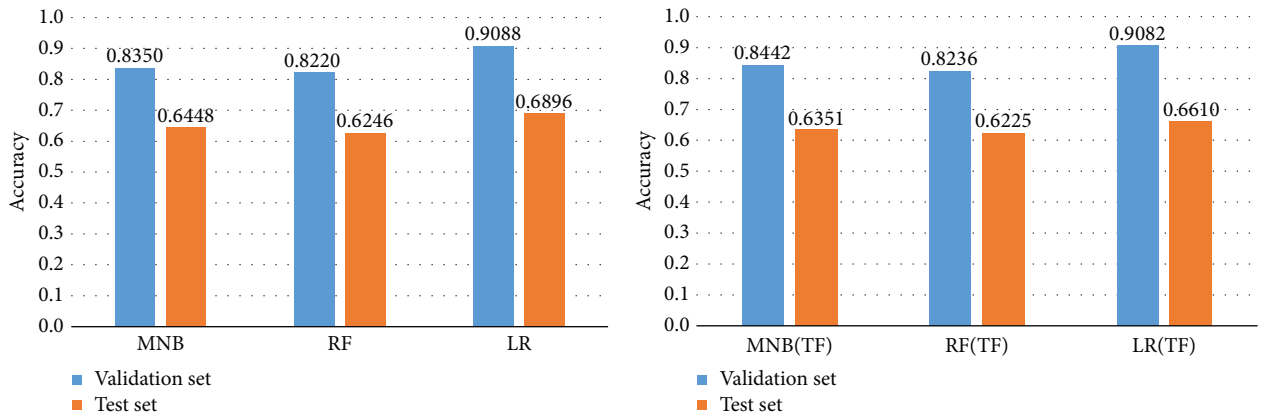
Figure 6(a) shows the impact of this threshold on the classification accuracy. And Table 1 shows the ratio of valid device fingerprints to the total number with different values of the minimal effective dimension, which shows the traffic coverage rate of device identification.

From Figure 6(a) and Table 1, we can see that the classification accuracy of the validation set and the test set increases gradually as the minimal effective dimension



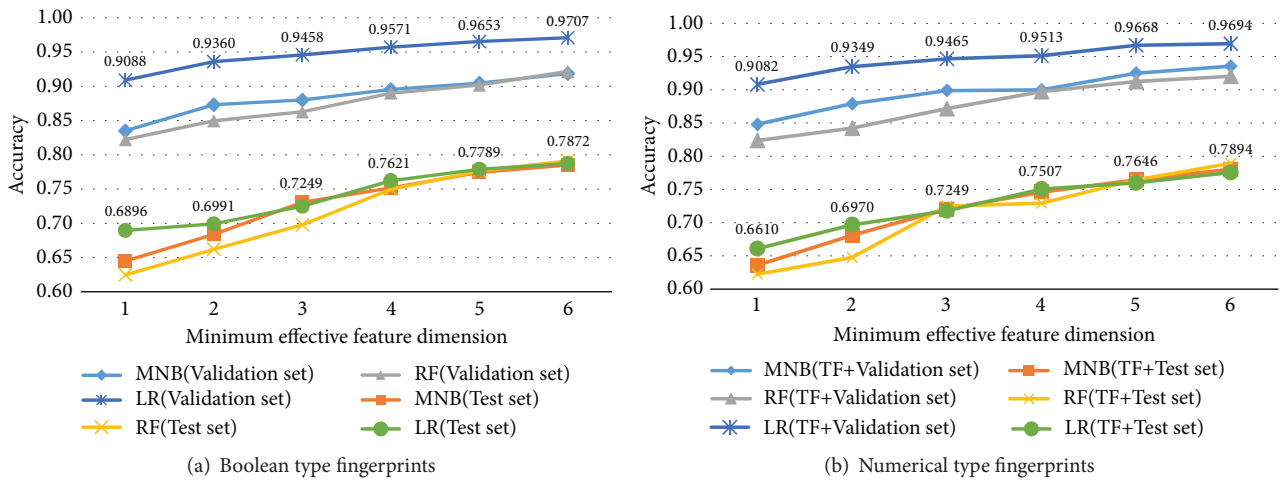
(a) Boolean type fingerprints (b) Numerical type fingerprints

Figure 4: The classification accuracy of device fingerprints using random forest model.



(a) Boolean type fingerprints (b) Numerical type fingerprints

Figure 5: The accuracy of device recognition using MNB, RF, and LR.



(a) Boolean type fingerprints (b) Numerical type fingerprints

Figure 6: The effect of the minimum effective feature dimension on the device classification accuracy.

Table 1: The effect of the minimum effective feature dimension on the number of device fingerprints.

Minimum Effective Feature Dimension	Validation Set	Test Set
1	100.00%	100.00%
2	89.58%	90.43%
3	82.98%	81.66%
4	75.94%	74.85%
5	70.57%	68.99%
6	63.73%	62.08%

climbs. The classification accuracy values of three models all plateau close to 80% for the test set when the minimal effective dimension is 6. However, despite the top accuracy, only 62.08% of device fingerprints in the test set are retained. In comparison, when the minimal effective dimension is lowered to 4, the classification accuracy of Multinomial Naive Bayes and logistic regression for the test set is higher than 75%, and 74.85% of device fingerprints in the test set are retained. Considering the effect of the minimal effective dimension on fingerprint classification accuracy and traffic coverage, the minimal effective dimension 4 is determined as the threshold for filtering fingerprint data. Since the logistic regression model performs comparatively better for both the validation set and the test set, the logistic regression model is chosen as the online identification model for the Boolean type device fingerprinting.

5.3.2. Training and Verification of Classification Model for Numerical Type Device Fingerprinting. For numerical type device fingerprinting, the same random forest parameters are first trained. The results are shown in Figure 4(b). When *nums* is 100 and depth is 30, the random forest model has the best classification effect. Since the specific value of each feature has an important influence on the classification result of the Multinomial Naive Bayesian classification model, we need to perform the term frequency (TF) transform as shown in formula (2) for each feature value.

$$f_x^{tf} = 1 + \log\left(f_x\right) \tag{2}$$

To further implement numerical type device fingerprinting, we train the Multinomial Naive Bayes classification model and the logistic regression classification model, respectively, and calculate the classification accuracy on the verification set and the test set. The results are shown in Figure 5(b). We also verify the impact of the minimum effective dimension on the classification accuracy, as shown in Figure 6(b). By comparing Figures 5(a) and 5(b), Figure 6(a) and Figure 6(b), respectively, we can find that the performances of Boolean and numerical type device fingerprinting are basically the same and can both achieve a comparatively high device identification accuracy. However, because the numerical type fingerprints may fluctuate on feature values, we only leverage the Boolean type device fingerprinting to test the online identification accuracy of devices.

5.4. Online Identification of User Devices. The online identification of user devices employs the Boolean type device fingerprinting, and a logistic regression model is taken as the classification model. The experiment is based on the sliding window mechanism, which simulates the online identification process by replaying network traffic in the test set for 23 days. The sliding window has two important parameters: the windows slide and the windows size.

A prediction is made iteratively when the sliding window slides backward over a distance of the window slide. The windows size is the range that fully covers flow data. When we want to make a prediction, we need to read feature data within the time range of the previous windows size from the current moment. The online identification accuracy rate of the user devices is counted as the ratio of the total number of device fingerprints correctly classified to the total number of device fingerprints in all the windows.

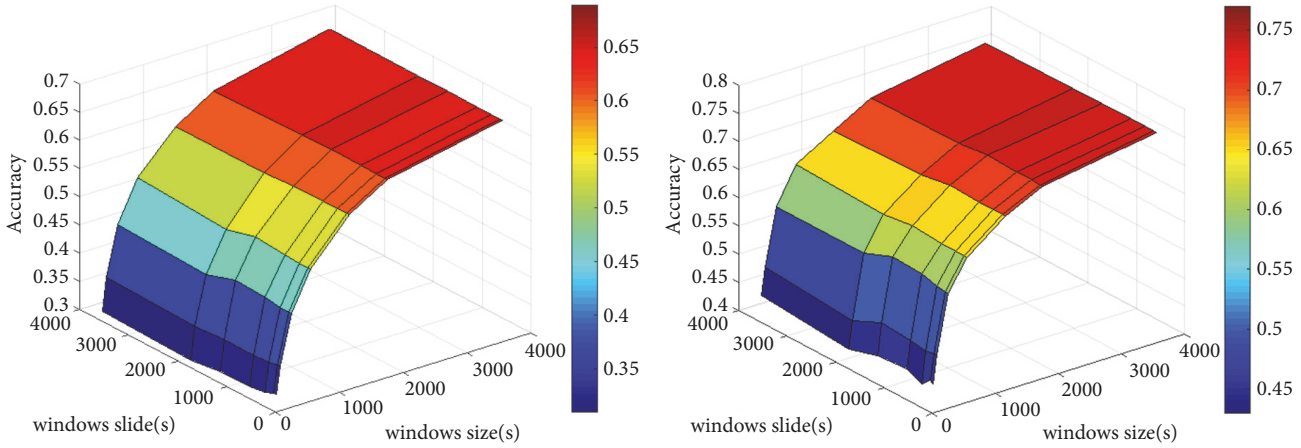
In this paper, the values of the windows slide and the windows size are set to be 1 minute, 2 minutes, 5 minutes, 10 minutes, 20 minutes, 30 minutes, and 60 minutes, respectively. By adjusting the values, we analyze how the two parameters influence online identification accuracy of user devices. Figure 7 shows the results when the minimum effective dimensions are 1 and 4, respectively.

As can be seen from the figure, the online identification accuracy rate is barely influenced by the change of windows slide, while it is in a positive correlated response to the increase of windows size. That is, the bigger the windows size, the more accurate the identification. When the windows size is 60 minutes, the identification accuracy rate reaches a maximum value of 68.93%. If we filter the data with low information content by setting the minimal effective dimension to 4, the maximal online identification accuracy will increase to 77.03%.

6. User Identification Based on Network Behavioral Fingerprints

6.1. Basic Ideas. In the scenario of intrusion detection, if some intruder occupies an authorized device, we cannot detect the intrusion by using the device identification technologies. So it is of great practical importance to identify user across multiple devices. To this end, we try to analyze the user's behavior habits and generate fingerprints, which are constituted by the device-independent Web access records, DNS domain name information, and HTTP User-Agent field. Except for the feature selection, other steps are similar to those of device identification. The specific identification procedures and verification steps go as follows.

6.2. Feature Selection and Fingerprint Generation. The user's network behavior habits are mainly reflected by his Web access records, and the attributes such as operating system and browser in HTTP User-Agent can also reflect the user's preferences to some extent. Therefore, we use the Web access records extracted from the application protocol identification unit, the mapping relationship between IP addresses and domain names obtained from the DNS analysis results, and



(a) Setting the minimum effective feature dimension to 1

(b) Setting the minimum effective feature dimension to 4

Figure 7: The Effect of sliding window on the accuracy of online device identification.

the information about the types of browsers, versions, and operating system types achieved through HTTP User-Agent detection, to generate user's online behavioral fingerprints for the identification of network users.

The behavioral fingerprint vector is generated by extracting features from captured traffic in a unit time. For the target IP address in the Web access record, we associate it with the domain name based on the DNS response records and treat all IP addresses and subdomains under the same domain name as one attribute of the vector.

After the features of domain names are determined, combined with the information contained in the HTTP User-Agent field, a network user's behavioral fingerprint is generated, which constitutes the fingerprint vector of the user behavior in a unit time.

The dimension of the user behavioral fingerprint vector is 57,593. According to the value type of feature attributes, behavioral fingerprints can also be divided into two types: boolean and numerical type. However, we can see from the fingerprint classification results of the device that there is no obvious difference between the classification accuracy of the two types of fingerprints, and the classification accuracy of Boolean type device fingerprinting is slightly better than that of numerical type fingerprinting. Therefore, we will test the identification accuracy of network user with Boolean type behavioral fingerprinting.

6.3. Training and Verification of Offline Model. Since the overall dimension of the behavioral fingerprint vector is large, we select Multinomial Naive Bayes and random forest to perform and compare their performance to select the better one for online identification of network users.

The network traffic collected in this paper contains data of 118 users. The data collection procedure lasts for 53 days. Each fingerprint is generated through extraction of each user's network data per hour. Note that the fingerprints with a full list of zero feature values are discarded. Altogether, we get a total of 54107 fingerprints. The overall fingerprints are categorized into two groups. One group contains 32,217

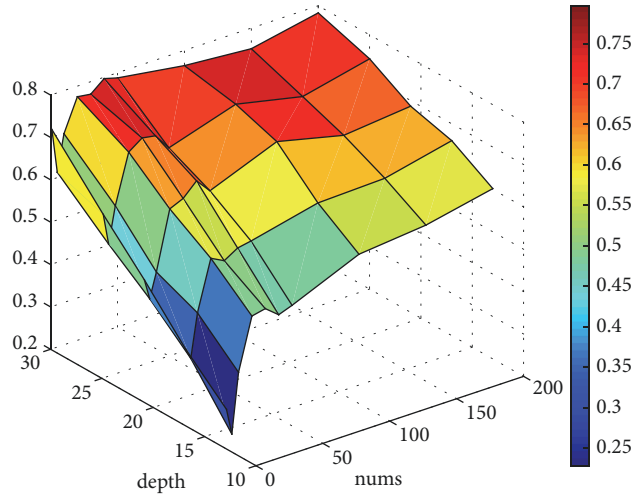


Figure 8: The classification accuracy of behavioral fingerprints using random forest model.

behavioral fingerprints collected in the first 30 days, used for the training and verification of the offline models. The other group includes the remaining 21,890 behavioral fingerprints collected in the following 23 days, used for testing the identification accuracy of network users.

Moreover, when training an offline model, the first group of the behavioral fingerprints are randomly divided into the training and verification sets, of which 70% of the behavioral fingerprints are used as a training set for model training, and the remaining 30% are used for verifying. The rest of the behavioral fingerprints are gathered as a test set for evaluating the classification accuracy. Two offline models, Naive Bayes and Random Forest, are trained with the same allocation of data sets.

First, the random forest model is trained. Its *nums* and *depth* parameters are taken into account. The obtained classification accuracy from training and testing under different values of *nums* and *depth* is shown in Figure 8. As can be seen

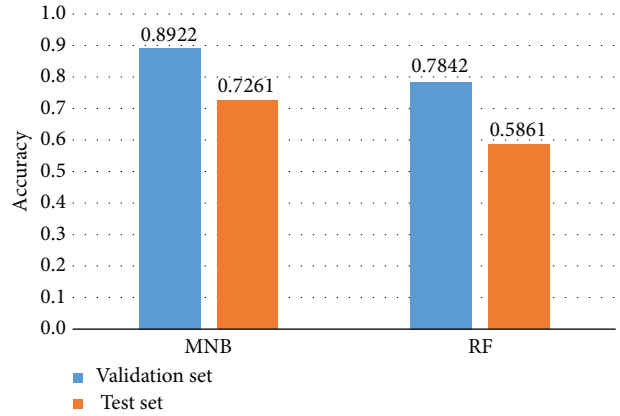


Figure 9: Behavioral fingerprints classification accuracy using MNB and RF.

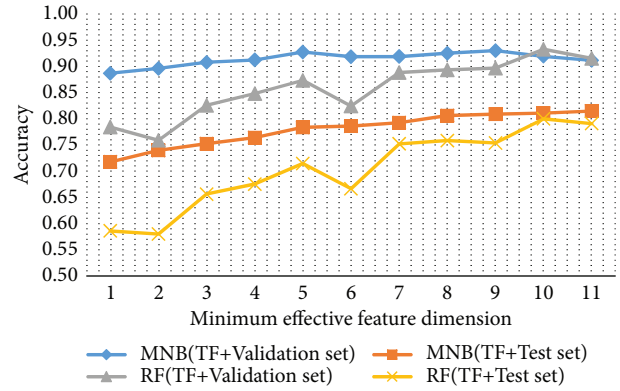


Figure 10: The effect of the minimum effective feature dimension on the user classification accuracy.

from the figure, when the *nums* is 40 and the *depth* is 30, the random forest model achieves the best classification accuracy.

Then the Multinomial Naive Bayes classification model is trained, and the classification accuracy is evaluated by the validation set and the test set, respectively. Figure 9 shows the classification accuracy of the validation set and the test set by performing the two models. These results show that the random forest model is far worse than the Multinomial Naive Bayes model for both data sets in terms of classification accuracy.

Finally, the effect of the minimum effective dimension on the classification effect and the ratio of valid behavioral fingerprints are tested. The results are shown in Figure 10 and Table 2, respectively.

As can be seen from Figure 10, the positive effect of the Multinomial Naive Bayes model on the test set is gradually enlarged as the minimum effective dimension increases. When the minimum effective dimension is set to 3, the classification accuracy rate of the test set is already higher than 75%. And when the minimum effective dimension is 9, the classification accuracy of the test set is the highest, reaching 80.70%, which can cover 74.87% of behavioral fingerprints.

Table 2: The effect of the minimum effective feature dimension on the number of behavioral fingerprints.

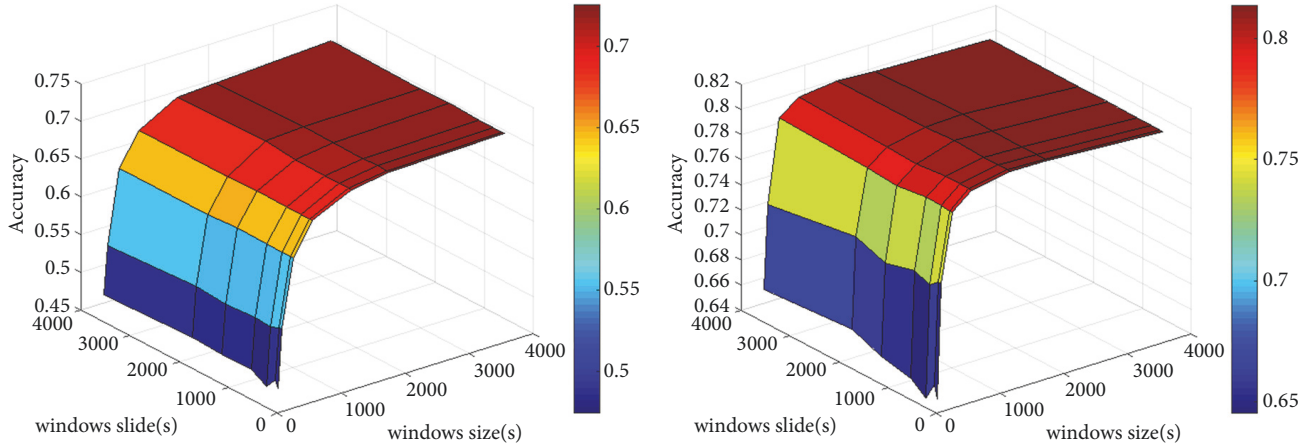
Minimum Effective Feature Dimension	Validation Set	Test Set
1	100.00%	100.00%
2	92.39%	93.70%
3	91.77%	90.54%
4	89.34%	86.72%
5	87.98%	84.22%
6	86.53%	81.03%
7	83.18%	78.36%
8	80.45%	76.42%
9	79.03%	74.87%
10	77.07%	73.43%
11	73.67%	72.16%

In comparison, the classification effect of the random forest model is not only relatively poorer, but also not stable enough. Therefore, we use the Multinomial Naive Bayes classification algorithm to implement the online user identification. Moreover, after comprehensively considering the effect of the minimum effective dimension on the classification accuracy and the coverage of the behavioral fingerprints, the value of 9 is selected as a condition to filter the invalid behavioral fingerprints.

6.4. Online User Identification. The online identification of network users is also performed in the sliding window manner, and the online process is simulated by replaying the real network traffic in the test set for 23 days. The step size and the window size of the sliding window are varied to figure out their effects on the user identification accuracy. In this paper, the values of step size are set to be 1 minute, 2 minutes, 5 minutes, 10 minutes, 20 minutes, 30 minutes, and 60 minutes and so are the values of window size. Figure 11 shows the results when the minimum effective dimensions are 1 and 9, respectively.

As can be seen from the figure, the step size of the sliding window has very little effect on the accuracy of online user identification. When the minimum effective dimension is 1, the accuracy goes up as the size of sliding window increases. When the window size is 60 minutes, the identification accuracy rate reaches a maximum of 72.58%. And the experimental results also show that when the window size of the sliding window is 20 minutes, the identification accuracy rate has already reached 71.42%. Thus, the time window size of the online user identification can be controlled within 20 minutes.

When the minimum effective dimension is 9, the identification accuracy rate firstly increases with the increase of the sliding window. When the window size increases to 5 minutes, the accuracy remains basically unchanged at 79.51%. When the window size is 20 minutes, the accuracy reaches 81.37%. And when the window size is 60 minutes, the rate



(a) Setting the minimum effective feature dimension to 1

(b) Setting the minimum effective feature dimension to 9

Figure 11: The effect of sliding window on the accuracy of online user identification.

Table 3: The results of high-speed packets capture (pps).

Packet length (Byte)	Number of physical cores				
	1	2	4	8	12
100	1329465	1566371	3206383	5936376	6272835
200	1130613	1734582	2917599	5352853	5586074
500	922524	1155344	2384589	2385555	2402127
1000	852867	1219714	1219109	1230260	1229943

is 80.74%. Therefore, the time window size of online user identification can be further shortened to 5 minutes.

7. Performance Tests and Results

7.1. Test Environment. In the above we have evaluated and proved the effectiveness of device identification and user identification with different algorithms and parameters, respectively. This section mainly tests the performance of the identification methods. The test environment is as follows:

7.1.1. Hardware

(i) 1 master node: Dell PowerEdge R730 (CPU: 2 6-core E5-2620V2, 2.1GHz, memory: 96 GB, external storage: 3.6TB).

(ii) 14 slave nodes: Dell PowerEdge C6220 II (CPU: 2 6-core E5-2620V2, 2.1GHz, memory: 64 GB, external: 8TB).

(iii) NIC: Intel 82599ES 10-Gigabit, supports up to 64 hardware queues.

7.1.2. Operating System

(i) Operating system: Red Hat Enterprise Linux 7.

(ii) Kernel version: 3.10.0-123.20.1.el7.x86_64.

7.2. Test Results. This section tests the performance of three modules: packets capture, distributed packets processing, and online identification modules. The test results are illustrated as follows:

7.2.1. Packets Capture Rate. First, network traffic is generated by the tcpreplay tool [34] and the packets capture rate is tested on the Intel 82599ES 10-Gigabit NIC. The NIC supports a maximum of 64 hardware queues and the number of hardware queues can be configured freely as required. However, due to the fact the CPU in our experimental computing node only has 12 physical cores and that the packet capture speed cannot be greatly enhanced if multiple packet capture threads are located in the same physical core, at most 12 hardware queues are enabled in this experiment.

In this paper, the speed of traffic capture is tested for the packets with different lengths and number of NIC hardware queues, separately. The results obtained are shown in Table 3. The results are acquired by calculating the average total numbers of packets that multiple NIC hardware queues capture within 10 seconds. The corresponding packet capture rate is shown in Table 4. From Tables 3 and 4 we can conclude that the packet length has a great influence on the capture rate. When the packet length is 1000 bytes, the ultimate speed of the NIC (9.76Gbps) can be reached by just consuming two hardware queues. If the packet length is reduced to 100 bytes, the packet capture rate rises as the number of NIC hardware queues increases till reaching a peak speed

Table 4: The results of high-speed packets capture (Gbps).

Packet length (Byte)	Number of physical cores				
	1	2	4	8	12
100	1.06	1.25	2.57	4.75	5.02
200	1.81	2.78	4.67	8.56	8.94
500	3.69	4.62	9.54	9.54	9.61
1000	6.82	9.76	9.75	9.84	9.84

Table 5: The speed test results of distributed processing framework.

The Number of Kafka Partitions	Maximum Processing Speed
1	3.76Gbps
2	6.68Gbps
3	10.01Gbps
4	13.35Gbps

Table 6: The time consuming of online identification.

#	Time consuming of online identification (ms)
1	9074
2	8256
3	7480
4	8188
5	6566
6	6780
7	6643
8	6503
9	8047
10	6080

of 5.02Gbps. When the length is changed to 200 bytes, the maximum capture rate is 8.94Gbps. When the packet length is 500 bytes, the maximum packet capture rate is 9.61Gbps.

The experimental results above show that the use of Linux PFQ kernel module can capture packets with a high speed and has robust systematic extensibility.

7.2.2. Speed of Distributed Processing of Packets. When testing the speed of the distributed processing framework, this paper uses KafkaProducer to write the network traffic captured by the packet capture module into each Kafka partition and then calculate the framework's processing speed of reading and analyzing data from Kafka partitions.

The number of KafkaSpouts is consistent with the number of Kafka partitions, which has a crucial influence on the speed of the distributed processing framework. Table 5 shows the speed of the distributed processing framework under different Kafka partition numbers. As we can see, the maximum processing speed is basically proportional to the number of Kafka partitions. It is noteworthy that it exceeds 10Gbps when the Kafka partition number is 3.

7.2.3. Response Speed of Online Application Identification. This paper uses the maximum window size of 60 minutes to test the response speed of the online identification module. This module contains two parts: online device identification based on the runtime environment and online user identification based on network behavioral fingerprints. Through the statistics of the time consumption of online identification modules for 10 trials, the average value is calculated as the response speed of the online identification module. The time consumption of the 10 online identifications is collected in Table 6. By averaging them, the response speed is derived as 7362 ms. This value is much smaller than the minimum step size of the recognition window (1 minute), so it can be considered that the online identification processing speed can meet the performance requirement.

8. Conclusion

In the intrusion detection area, it is increasingly important to detect the suspicious entities and potential threats. In this paper, we introduce the identification technologies of network entities to detect the potential intruders. In order to achieve network entities identification in high-speed environment, we use PFQ kernel module to capture high-speed network packets and employ Storm distributed real-time streaming data processing technology to realize online analysis of network traffic.

For the unauthorized devices in the monitored network, we design an online device identification technology based on runtime environment analysis. 961 features, such as application program, operating system, and HTTP User-Agent field, are selected to constitute the device fingerprints. And then the logistic regression algorithm is applied in a sliding window manner. For the case that the intruder occupies an authorized device and disguises himself as an authorized user, we extract the Web access record, DNS domain name, and HTTP User-Agent field to constitute user behavioral fingerprints. And then users are identified online in a sliding window manner using the Multinomial Naive Bayes model. The experimental results show that the traffic analysis framework and identification methods proposed in this paper have a high practicality as they can achieve satisfying identification accuracy rates in a short time. For future research, we intend to design an automated application identification tool in order to identify a large scale of applications and enhance the identification accuracy.

Data Availability

The network traffic data used to support the findings of this study have not been made available because they contain a lot of privacy information.

Disclosure

Any opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by National Key R&D Program of China 2018YFB0803400 and 2017YFB1003000, National Natural Science Foundation of China under grants 61572130, 61502100, 61532013, and 61632008, by Jiangsu Provincial Scientific and Technological Achievements Transfer Fund BA2016052, by Jiangsu Provincial Key Laboratory of Network and Information Security under grants BM2003201, by Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under grants 93K-9, and by Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] G. Pedro T, D. Jes·sE V, M. Gabriel F, and V. Enrique, "Anomaly-based network intrusion detection: Techniques, systems and challenges," in *Computers Security*, pp. 18–28, 18–28, 28(1–2, 2009.

[2] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: methods, systems and tools," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.

[3] J. Pang, B. Greenstein, R. Gummadi, S. Seshan, and D. Wetherall, "802.11 User fingerprinting," in *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking*, pp. 99–110, ACM, September 2007.

[4] Y. Zhang, M. Yang, X. Gu, P. Pan, and Z. Ling, *Proceedings of the 2018 International Conference on Advanced Cloud and Big Data*, LanZhou, China, 2018.

[5] B. Danev, D. Zanetti, and S. Capkun, "On physical-layer identification of wireless devices," *ACM Computing Surveys*, vol. 45, no. 1, article 6, 2012.

[6] T. Kohno, A. Broido, and C. Claffy K, "Remote physical device fingerprinting," in *Proceedings of the 26th IEEE Symposium on Security and Privacy (SP'05)*, Berkeley, CA, USA, 2005.

[7] R. Gerdes, T. Daniels, M. Mina, and S. Russell, "Device Identification via Analog Signal Fingerprinting: A Matched Filter Approach," in *Proceedings of the 13th Annual Network and Distributed System Security Symposium Conference (NDSS'06)*, San Diego, CA, USA, 2006.

[8] E. D. Thomas, J. A. Van Randwyk, E. J. Lee et al., "Passive data link layer 802.11 wireless device driver fingerprinting," in *Proceedings of the 15th conference on USENIX Security Symposium*, Vancouver, B.C., Canada.

[9] T. Yen, Y. Xie, F. Yu, R. Yu, and M. Abadi, "Host Fingerprinting and Tracking on the Web: Privacy and Security Implications," in *Proceedings of the 19th Annual Network and Distributed System Security Symposium (NDSS'12)*, 2012.

[10] T. Stöber, M. Frank, J. Schmitt, and I. Martinovic, "Who do you sync you are?" in *Proceedings of the the sixth ACM conference*, p. 7, Budapest, Hungary, April 2013.

[11] P. Eckersley, "How unique is your web browser?" in *Privacy Enhancing Technologies: 10th International Symposium, PETS 2010, Berlin, Germany, July 21–23, 2010. Proceedings*, vol. 6205 of *Lecture Notes in Computer Science*, pp. 1–18, Springer, Berlin, Germany, 2010.

[12] K. Mowery, D. Bogenreif, S. Yilek, and H. Shacham, "Fingerprinting information in JavaScript implementations," in *Proceedings of 2011 Web 2.0 Security and Privacy (W2SP'11*, Oakland, California, 2011.

[13] J. Mayer and J. Mitchell, "Third-party web tracking: Policy and technology," in *Proceedings of the 33rd IEEE Symposium on Security and Privacy (SP'12)*, San Francisco, CA, USA, 2012.

[14] G. Acar, M. Juarez, N. Nikiforakis et al., "FPDetective: Dusting the web for fingerprinters," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security, CCS 2013*, pp. 1129–1140, Germany, November 2013.

[15] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "Cookieless monster: Exploring the ecosystem of web-based device fingerprinting," in *Proceedings of the 34th IEEE Symposium on Security and Privacy, SP 2013*, pp. 541–555, USA, May 2013.

[16] N. Nikiforakis, A. Kapravelos, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "On the workings and current practices of web-based device fingerprinting," *IEEE Security & Privacy*, vol. 12, no. 3, pp. 28–36, 2014.

[17] P. Laperdrix, W. Rudametkin, and B. Baudry, "Beauty and the Beast: Diverting Modern Web Browsers to Build Unique Browser Fingerprints," in *Proceedings of the 2016 IEEE Symposium on Security and Privacy, SP 2016*, pp. 878–894, USA, May 2016.

[18] A. Kurtz, H. Gascon, T. Becker, K. Rieck, and F. Freiling, "Fingerprinting Mobile Devices Using Personalized Configurations," *Proceedings on Privacy Enhancing Technologies*, vol. 2016, no. 1, pp. 4–19, 2016.

[19] W. Wu, J. Wu, Y. Wang, Z. Ling, and M. Yang, "Efficient Fingerprinting-Based Android Device Identification with Zero-Permission Identifiers," *IEEE Access*, vol. 4, pp. 8073–8083, 2016.

[20] Y. Cao, S. Li, and E. Wijmans, "(cross-)browser fingerprinting via os and hardware level features," in *Proceedings of the 24th Annual Network and Distributed System Security Symposium, NDSS*, San Diego, CA.

[21] R. V. Yampolskiy and V. Govindaraju, "Behavioural biometrics: a survey and classification," *International Journal of Biometrics*, vol. 1, no. 1, pp. 81–113, 2008.

[22] N. Zheng, A. Paloski, and H. Wang, "An efficient user verification system via mouse movements," in *Proceedings of the 18th ACM Conference on Computer and Communications Security*, pp. 139–150, ACM, October 2011.

[23] S. Douhou and J. R. Magnus, "The reliability of user authentication through keystroke dynamics," *Statistica Neerlandica. Journal of the Netherlands Society for Statistics and Operations Research*, vol. 63, no. 4, pp. 432–449, 2009.

[24] N. V. Verde, G. Ateniese, E. Gabrielli, L. V. Mancini, and A. Spognardi, "No NAT'd user left behind: Fingerprinting users behind NAT from NetFlow records alone," in *Proceedings of the 2014 IEEE 34th International Conference on Distributed Computing Systems, ICDCS 2014*, pp. 218–227, Spain, July 2014.

[25] B. Padmanabhan and Y. Yang, *Clickprints on the web: Are there signatures in web browsing data*, 2006, http://knowledge.wharton.upenn.edu/papers/1323.pdf.

[26] Y. Yang, "Web user behavioral profiling for user identification," *Decision Support Systems*, vol. 49, no. 3, pp. 261–271, 2010.

[27] M. Kumpošt and V. Matyáš, "User Profiling and Re-identification: Case of University-Wide Network Analysis," in *Trust, Privacy and Security in Digital Business*, vol. 5695 of *Lecture Notes in Computer Science*, pp. 1–10, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[28] D. Herrmann, C. Gerber, C. Banse, and H. Federrath, "Analyzing Characteristic Host Access Patterns for Re-identification of Web User Sessions," in *Information Security Technology for Applications*, vol. 7127 of *Lecture Notes in Computer Science*, pp. 136–154, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[29] C. Banse, D. Herrmann, and H. Federrath, "Tracking Users on the Internet with Behavioral Patterns: Evaluation of Its Practical Feasibility," in *Information Security and Privacy Research*, vol. 376 of *IFIP Advances in Information and Communication Technology*, pp. 235–248, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[30] D. Herrmann, C. Banse, and H. Federrath, "Behavior-based tracking: Exploiting characteristic patterns in DNS traffic," *Computers & Security*, vol. 39, pp. 17–33, 2013.

[31] D. W. Kim and J. Zhang, "You Are How You Query: Deriving Behavioral Fingerprints from DNS Traffic," in *Security and Privacy in Communication Networks*, vol. 164 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pp. 348–366, Springer International Publishing, Cham, 2015.

[32] X. Gu, M. Yang, C. Shi, Z. Ling, and J. Luo, "A novel attack to track users based on behavior patterns," *Concurrency and Computation: Practice and Experience*, 2016.

[33] Blockmon, "cnplab/blockmon," https://github.com/cnplab/blockmon.

[34] Tcpreplay, "Tcpreplay development is now being done by AppNeta," URL http, http://tcpreplay.synfin.net.