Hindawi
Security and Communication Networks
Volume 2022, Article ID 8511258, 10 pages
https://doi.org/10.1155/2022/8511258

Research Article

Detecting ShadowsocksR User Based on Intelligence of Cyber Entities

Jiancong Zhang, Ping Dong, Minyu Jin, and Yuting Tang

Huaxin Consulting Co., Ltd., Hangzhou 310000, China

Correspondence should be addressed to Ping Dong; dongping.hx@chinaccs.cn

Received 15 April 2022; Accepted 18 July 2022; Published 24 August 2022

Academic Editor: Jehad Ali

ShadowsocksR (SSR), as a typical emerging anonymous communication tool, may record user information on the SSR client or server, leading to the theft of the user's privacy, and may be used by attackers to anonymize their internal network environment and organization, which will cause serious damage to data security and bring severe challenges to security defense and threat assessment within organizations. To solve the problem of accurately and effectively discovering SSR users within an organization in a real traffic environment, in this paper, we propose an SSR user detection method based on network entity intelligence as follows: (1) According to the communication characteristics of SSR users, relevant network entity intelligence information from inside and outside the organization is obtained, such as the distribution of IP addresses within and outside the organization, and the differences between SSR and non-SSR users are analyzed to construct a feature space. (2) The communication behaviors of SSR and non-SSR users are further analyzed and features are extracted from the perspective of traffic behavior analysis, and the feature space of the SSR user detection model is expanded. (3) A data-driven machine-learning-based approach is designed and implemented to provide suggestions for the automatic identification of SSR users based on the extracted feature vectors. Results show that the detection method proposed in this paper has a detection accuracy of more than 95% for SSR users in the experimental environment, can accurately distinguish between SSR communication and normal communication, and can achieve accurate SSR user detection.

1. Introduction

With the increasing demands of network users for data security and access to websites outside an organization, anonymous communication proxy tools, such as ShadowsocksR (SSR), are used by most Internet users because of their simple operation and relatively safe performance. Their basic function is to bypass a firewall to access blocked content. To ensure the security and anti-identification of data, ShadowsocksR (SSR) has been upgraded and strengthened on the basis of SS. It can not only complete the basic functions of SS but can also help cover the network activities of clients and hide the real users' IPs. SSR is conducive to ensuring the privacy and security of network terminals and has become the first choice for increasingly more users to access the Internet. However, SSR helps users to break through the network supervision within an organization, which is bound to bring challenges to network supervision. On the one hand, it is impossible to identify the real identity of the SSR users and judge their trustworthiness, which may lead to the leakage of internal organizational information and pose serious security threats. On the other hand, if the SSR users inside an organization make malicious remarks or carry out network attacks by hiding their real identities, administrators will not be able to take timely measures to prevent the impact of such situations. To protect the internal assets and network environment of an organization, accurately identifying the SSR users within the organization to improve internal security supervision and defense capabilities have become an urgent goal to meet.

The network communication between SSR and non-SSR users is very similar because SSR will pretend to communicate as a common network user in the process of providing resource requests and forwarding so that the commonly used traffic identification technology cannot accurately identify SSR traffic, which brings great challenges to the detection of SSR users. At this stage, a significant amount of detection research on proxy communication has been conducted in academia, all of which have achieved remarkable results. According to the different analysis methods employed, agent communication detection can be divided into that based on behavior analysis and that based on network protocol analysis. Proxy communication detection based on network behavior analysis [1–3] extracts representative network behavior characteristics through behavior analysis of payload content from network data packets, network traffic files, and mixed log files, to establish a proxy detection system with rule matching as the core.

Proxy communication detection based on network protocol analysis [4–6] is mainly based on feature extraction of the attributes of static entities in network traffic to construct feature vectors, which are finally input into machine-learning classification algorithms for the automated proxy traffic identification process. Compared with the method of network behavior analysis, this method does not rely on packet-level features but obtains intelligence information extraction features of network entities (such as IP and URL) through network protocol analysis, by which the detection of web page proxies can be realized. However, there is still room for improvement in accuracy. Finding the differences between SSR and non-SSR users and improving the accuracy of detection results are the key issues that must be urgently solved in current SSR user detection research.

In the field of network traffic identification and threat detection, network entity intelligence has been widely used to expand the feature space of detection systems. Network entity intelligence is intelligence information provided by network entities, such as IP and domain, which is helpful for traffic analysis. To solve the abovementioned problem, namely that using the attribute information of traffic itself is not enough to accurately identify the communication traffic of SSR users, in this paper, we propose an SSR user detection method based on network entity intelligence. With the help of this method, network managers can effectively and comprehensively supervise the network behavior of the internal personnel of an organization and improve the overall defense ability. In this paper, we use the request and response traffic generated by SSR users to access resources as experimental data. Experimental results show that the detection method proposed herein can effectively detect SSR users in a network environment. The main contributions of this paper are the following:

(1) External network entity intelligence is introduced to construct a feature space for traffic identification. Collecting network entity intelligence adds useful information to discovering differences in communication behavior between SSR and non-SSR users.

(2) Combined with the network behavior analysis method, the expansion of the feature space constructed based on network entity intelligence by analyzing the network behavior characteristics based on the original network traffic data is realized, and the accuracy of SSR user communication traffic identification improved.

(3) A general SSR user detection method based on network entity intelligence is proposed that can achieve efficient detection of SSR users.

The rest of this paper is organized as follows: Section 2 introduces the related research, including the research status of SS detection and anomaly detection based on network intelligence. Section 3 introduces the working principle of the research object of this paper. Section 4 introduces the proposed method of detecting SSR users, including the detection framework and specific feature analysis process for SSR users. Section 5 details the experiments and analysis carried out in this paper, including evaluation criteria, model selection, and validity evaluation of the detection method under cross-validation. Section 6 concludes the paper.

2. Related Work

In this section, we mainly introduce the research status of SS/SSR communication detection and the application of network entity intelligence in the field of anomaly detection.

2.1. SS/SSR Detection. In recent years, the relevant research hotspots have mainly focused on the identification and detection of SS/SSR communication traffic. To analyze SS security [7], SS traffic detection can be divided into SS traffic detection based on feature extraction and web page fingerprinting according to the detection algorithm used. SS traffic detection based on feature extraction implements SS traffic detection by extracting traffic packet-level features combined with machine learning or deep learning to classify traffic.

Most studies rely on the statistical characteristics of data packets [8–11]. Among them, Wang et al. added the sliding window JS divergence feature based on the traditional statistical features based on traffic packet length and timestamp by analyzing the protocol principle of SS and the characteristics of traffic data of different applications and used the random-forest algorithm in the experiment. The SS traffic was identified with 94.5% accuracy in the environment. Detection methods based on the relational features of packets also exist, such as that reported by Zeng [12]. In the scenario of the web page and website identification for anonymous traffic, few studies apply web page fingerprinting to traffic identification. SS traffic detection based on network fingerprinting must be carried out in this scenario, e.g., as in the following.

Maohua [13] et al. focused on analyzing the homologous relationship between website fingerprints and designed a convolutional-neural-network-bi-line long short-term memory (CNN-BiLSTM) attack classification model that can classify the anonymous encrypted traffic of SS with an accuracy of 98.1% with small samples. Li et al. [14] identified traffic from three perspectives, i.e., based on bit-flow, host features, and hidden features, adopting the idea of layering,

combining multi-granularity heuristic traffic identification and website fingerprinting based on hybrid flow segmentation, to improve the accuracy of identification. According to the different classification methods used in the aforementioned detection schemes, the detection methods can be divided into SS traffic detection based on random forest [8, 9], CNN algorithms [11], and the XGboost boosting algorithm [10].

The confusion and camouflage of SSR make it more difficult to identify its traffic. To date, few detection schemes for SSR communication exist, most of which have been improved on the basis of the SS communication detection method. Ji et al. [15] proposed an SSR traffic identification algorithm based on the XGboost algorithm and K-means clustering algorithm for SS traffic disguised as HTTP traffic. They analyzed the differences between SSR and non-SSR communication traffic from the perspective of the HTTP protocol and encrypted traffic and extracted the statistical features of the traffic load, the information entropy of the first four packet payloads of single traffic, and other entropy obtained by cluster analysis as the features by which to identify the SSR traffic. Experiments show that this model can achieve a better recognition effect than others under the same conditions. However, it does not effectively utilize the intelligence of external network entities, resulting in a large false-positive rate for the model in the real environment. The authors of reference [16] proposed an SSR traffic detection method based on the DART algorithm the following year. Compared with the previous method, it can identify more types and ranges of SSR obfuscated traffic. However, the algorithm is improved from the HTTP and TLS protocols; so, it can only be used to detect the camouflaged traffic detection of HTTP and TLS. Starting from the SS (R) protocol, Luo et al. [17] characterized encrypted traffic with strong camouflage from the time, packet-length, payload, and packet-header dimensions separately. The identification of SS (R) traffic is realized with the help of the GOSS algorithm, and SS and SSR traffic are distinguished only based on port information. Table 1 shows the main features and disadvantages of the abovementioned detection schemes.

The aforementioned methods for SS communication traffic can all achieve a certain detection effect but compared with SS, SSR, and non-SSR users have a higher similarity in traffic, so detection methods that rely on network behavior analysis will not be enough to achieve ideal detection results. How to extract features from the traffic level to comprehensively describe the traffic generated by SSR users and realize the detection of SSR users is the main problem to be solved in this paper.

2.2. Anomaly Detection Based on Network Entity Intelligence. Network entity intelligence is derived from public information on the Internet and can provide a basis for tracking network threats [18]. It is necessary to effectively use entity intelligence data widely existing in the network; realize the detection of network anomalies through data analysis, mining, and modeling; and then generate security threat intelligence.

Anomaly detection based on network entity intelligence refers to the ability to make full use of threat intelligence information on the Internet to identify known or unknown attack behaviors and threat methods. Eshete et al. [19] introduced the known attacks of a system and the intelligence information provided by similar system entities and constructed a query graph of the attack, which transforms threat detection into a graph-pattern-matching problem, enabling reliable detection of network attacks. Yurekten et al. [20] integrated the concepts of cyber entity intelligence, network function virtualization (NFC), and business function chaining (SFC) into an automated defense system of software-defined networks, which can evaluate defense strategies based on intelligence, in which one can choose to apply one or more network-level automated defense solutions to ensure that the defense system is scalable while increasing the intensity of attack processing. Gao et al. [21] proposed a pipeline technology for extracting threat intelligence in entity intelligence and the correlation between intelligence and for drawing threat behavior maps for threat discovery. Rong et al. [22] collected and correlated entity intelligence related to existing network attack methods to predict the network security status of a current system and achieve effective defense against network attacks. Zhang et al. [23] gathered the intelligence of cyber threat entities inside and outside the system for situational awareness and established a situational awareness model based on the stochastic a. Experiments show that the network security situational awareness method with the help of threat intelligence can accurately reflect the changes in network security situations and predict attack behavior. Xin et al. [24] combined the entity intelligence and features of malicious URLs, which can not only quickly detect known malicious URLs but can also quickly identify unknown malicious URLs. Gang et al. [25], drawing on entity intelligence and user behavior analysis, designed a solution to integrate the network anomaly detection module with the security data platform, realizing real-time online anomaly detection of network data. Muhammad et al. [26, 27] also used cyber entity intelligence for authentication problems in the Internet of Things (IoT) and Internet of Vehicles. With the help of intelligence information, the identification of legal vehicles and the identification of authorized users of the IoT can be realized. It can thus be seen that network entity intelligence has been widely used to solve the classification problem in network security.

Based on the descriptions in this subsection, it is not difficult to find that cyber entity intelligence is a reliable source of information for threat detection. By making full use of the network entity intelligence information from inside and outside a network system, not only can a more adequate analysis of network traffic be supported but the feature space for traffic identification can also be expanded to achieve a more comprehensive characterization of abnormal traffic. In this way, normal and abnormal traffic can be more accurately distinguished, and a more efficient and accurate traffic detection solution can be generated. Therefore, to improve the accuracy of SSR user detection, we start from the working principle of SSR and introduce the

Table 1: Comparison of the related Shadowsocks (R) detection methods.

Reference	Year	Object	Main techniques	Defects
[8]	2022	SS	Added the sliding window JS divergence feature. Includes random forest.	Method to identify smartphone applications from the network traffic of SS proxy.
[9]	2020	Proxy	Uncertainty-based traffic sample selection strategy. Includes random forest.	Types of proxies that can be detected are limited.
[10]	2018	SS	Bit-flow features and XGboost algorithm.	Insufficient features for traffic characterization.
[11]	2020	SS	CNN. Includes random forest	Deep learning needs enough training samples.
[12]	2019	SS	Novel SS detection method based on flow context and host behavior.	Applicability of the method is affected by scenario.
[13]	2021	SSH + SS	CNN-BiLSTM algorithm.	Poor applicability in real network environments.
[14]	2017	SS	Multi-granularity heuristic traffic detection algorithm and mixed stream division-based website fingerprint detection algorithm.	Not applicable to multi-obfuscation SSR traffic detection.
[15]	2020	SSR	XGboost algorithm.	Only focused on identification of traffic camouflaged with HTTP protocol in SSR.
[16]	2021	SSR	DART algorithm.	Only focused on identification of traffic camouflaged with HTTP and TLS protocols in SSR.
[17]	2021	SS and SSR	Protocol analysis and GOSS algorithm.	No effective distinction between SS and SSR.

features needed to characterize SSR user traffic for the external intelligence information enrichment detection model. Thereby, high-accuracy SSR user identification is realized.

3. Working Principle of SSR

In this section, we briefly introduce the working principle of SSR and analyze and illustrate the network entity intelligence used in detail. Table 2 summarizes the basic definitions and symbols used.

3.1. ShadowsocksR. SSR is a proxy software based on the Socks5 protocol. The communication principle of SSR is shown in Figure 1. SSR splits the Socks5 protocol into two parts, SSR-Local and SSR-Server, which are located on both sides of the firewall. SSR-Local is generally the local machine or router or another machine on the local area network, without crossing the firewall. The request sent by the user communicates with SSR-Local based on the Socks protocol, and the two ends of SSR-Local and SSR-Server communicate through a variety of optional encryption methods. SSR-Server decrypts the received encrypted data, restores the data to the original request, and then sends the data to the service that the user needs to access, and returns along the original route after obtaining the response.

SSR uses encryption to encapsulate requests and forwarding, disguises traffic as regular protocol traffic, and adds "protocol" and "obfuscation" plug-in options based on encryption. The "protocol" is used to encapsulate the data in a certain format before the network application data are encrypted, thereby increasing the concealment of the data; and "obfuscation" disguises the encrypted network traffic data as regular network traffic such as HTTP or TLS traffic, thereby reducing the identifiability of the data.

The similarity of the communication traffic between SSR and non-SSR users greatly increases the difficulty of detecting SSR traffic based on traffic characteristics. Therefore, in this paper, we construct a feature vector for SSR traffic by introducing relevant network entity intelligence, to achieve the purpose of improving the detection accuracy of SSR users.

3.2. Communication Behavior of SSR Users. As anonymous proxy software, SSR can decide the proxy mode according to the user's choice, including direct connection mode, PAC, and global mode. Users, who choose direct connection mode will not be able to use the proxy service of SSR; users who choose the PAC mode will generate a whitelist based on the automatically updated rule document, and the system will provide users with proxy services when accessing the domain name addresses in the whitelist; and, in the global mode, users will proxy clients through SSR, regardless of whether they access the local network or the external network.

According to the operation mechanism and proxy mode of the SSR proxy service, the network communication requests initiated by the SSR users will be directly forwarded by SSR-Local, and there is no direct connection between the user host and SSR-Server. A comparison of network communication between SSR and non-SSR users is shown in Figure 2.

4. Model of Identifying SSR Users

According to the analysis of the working principle and communication behavior of SSR presented in Section 3, we found that the communication between the user and SSR client behaves as a normal TCP connection, and SSR adds a confusion mechanism on the basis of SS, which leads to a further strengthening of the communication pattern similarity with non-SSR users and also increases the difficulty of detecting SSR users greatly. How to select

Table 2: Notation table.

Symbol	Definition
SS (R)	ShadowsocksR (SSR)
$N(extra - ip)$	Number of extra-area IPs communicating with the users' host within the specified time window
$N(\text{intra} - ip)$	Number of intra-area IPs communicating with the users' host within the specified time window
$N(extra - domain)$	Number of extra-domains accessed by the users' host within the specified time window
$N(\text{intra-domain})$	Number of intra-domains accessed by communicating with the users' host within the specified time window
$D(ip)$	Average ratio of extra_ip and intra_ip per minute
$D(Domain)$	Average ratio of extra_domain and intra_ip per minute
s	Statistical characteristics of the sender
r	Statistical characteristics of the receiver
t	Statistical characteristics of the bidirectional conversation flow

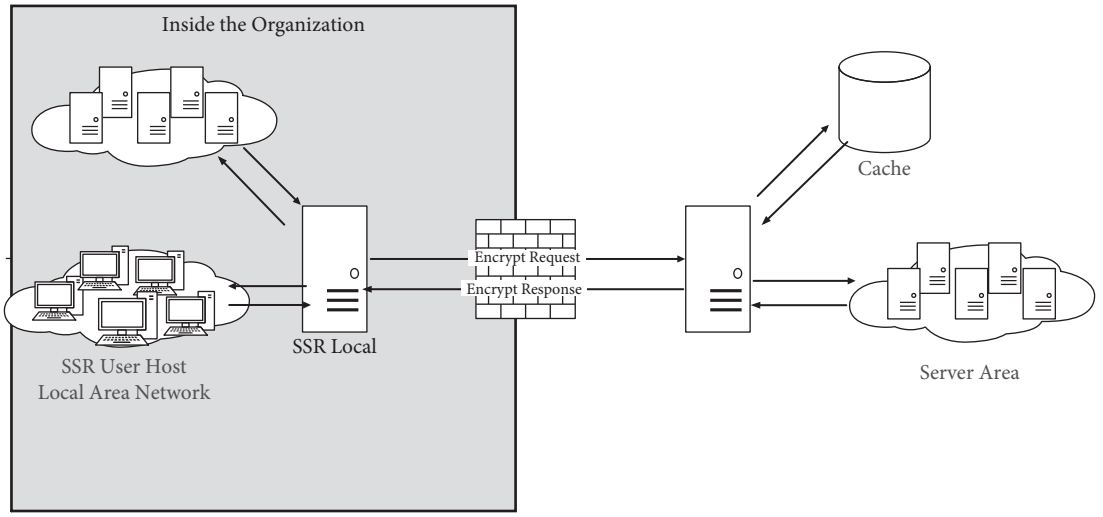

Figure 1: Schematic of SSR.

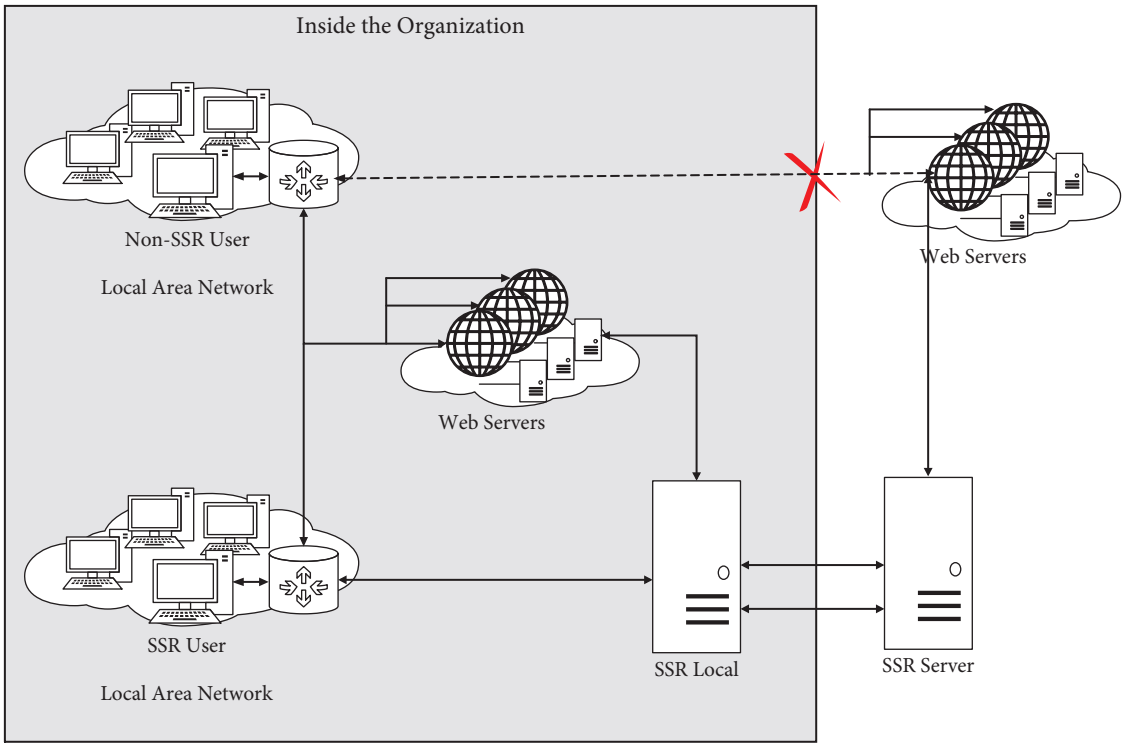

Figure 2: Comparison of network access between SSR and non-SSR users.

external intelligence information to expand the feature space that characterizes the network traffic of SSR users is the main problem to be solved.

4.1. Available Cyber Entity Intelligence Analysis. To achieve effective SSR user identification, we introduce the following network entity intelligence to analyze the communication behavior of SSR and non-SSR users from the perspective of entity intelligence.

Comparing the communication objects of SSR and non-SSR users, we select the intelligence information composed of the IP and domains of the communication objects as follows:

(a) Location characteristics of communication objects: each organization maintains a range of IP addresses within its network isolation area that determines the geographic location with which users within the organization can communicate. We compare the geographical distribution of communication objects between SSR and non-SSR users within the organization by collecting IP whitelist information inside the organization's isolation area (or outside the organization's isolation area). According to the IP address range whitelist comparison, the geographic location information of communication objects is divided into two categories: session traffic with communication objects located in the organization isolation area and session traffic with communication objects located outside the organization isolation area. The whitelist on which the feature is divided must be updated and maintained according to the IP address ranges within different organizations.

(b) Domain intelligence: similar to the location characteristics of communication objects, to manage the internal network within each organization, the websites that users within the organization can access will be restricted, and the list of domain names and addresses accessed by users will also be maintained. The domain address information accessed in the DNS request of the SSR users or the "host" information in the HTTP traffic may also come from outside the organization's network isolation domain. We determine whether the geographic location of the domain is within the organization isolation area based on the type of domain and the organization to which it belongs. We use the top 500 domain addresses outside the organization's isolation domain to be selected from the ALexa Top data to form a blacklist of domain addresses.

4.2. Feature Extraction. To accurately identify SSR users, network entity intelligence and network behavior analysis are used to jointly construct the feature space of the detection model.

4.2.1. Feature Analysis Based on Network Entity Intelligence. To accurately identify SSR users, network entity intelligence and network behavior analysis are used to jointly construct the feature space of the detection model.

(1) Communication Object Location Features. The purpose of SSR users is to access IP addresses outside the organization's isolation area. Therefore, the location feature of communication objects of flow data is constructed by using the IP address whitelist of objects that can communicate within the organization. The specific representation of the feature is as follows: extract the IP information of the communication object in the network traffic files, and compare the whitelist information of the IP range list of the isolated domain of the organization to determine whether the IP is marked as intra-area or out-of-area. Aggregate the average number of out-of-area communication objects of intra-area communication objects of SSR and non-SSR users with time windows of 1 min, 5 min, 15 min, 30 min, and 1 h. The result is shown in Figure 3.

It can be seen in Figure 3 that regardless of the time granularity, SSR users have significantly more access to IP addresses outside the organization area than non-SSR users. The corresponding feature vector is extracted based on the position label of the communication object, which is composed of the distributed features inside and outside the area of the IP.

$$D(ip) = \frac{N(extra - ip)}{N(\text{intra} - ip)}, \tag{1}$$

where $D(ip)$ represents the intra- and extra-area distributions of the communication object, $N(extra - ip)$ represents the number of extra-area IPs communicating with the users' host within the specified time window, and $N(intra - ip)$ represents the number of intra-area IPs communicating with the users' host within the specified time window.

(2) Domain Location Features. Similar to the location characteristics of communication objects, the blacklist of domain name addresses outside the isolation area is used to collect and filter traffic data to generate the location features of user access domains. The specific description of the feature is as follows: the domain information extracted from the network traffic file is matched against the domain blacklist to determine whether the domains are marked as intra- or extra-area. Similarly, the average number of SSR and non-SSR users accessing domain servers inside and outside the area is aggregated in time windows of 1 min, 5 min, 15 min, 30 min, and 1 h.

As shown in Figure 4, regardless of the time granularity, the number of SSR users accessing extra-area domains is significantly higher than that of non-SSR users. The corresponding feature vector is extracted based on the location label of the domain visited by the users, which is composed of the distribution features of the domain visited inside and outside the area, which is defined as the ratio of the numbers of extra- and intra-domains accessed by users within a specified time window. The formula is as follows:

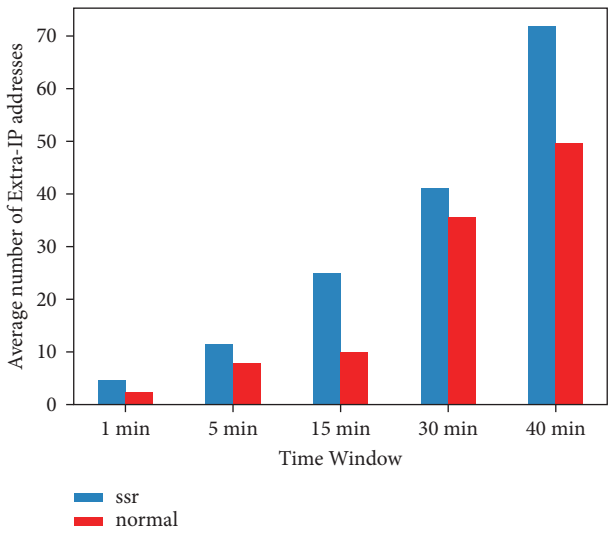

Figure 3: IP distribution result.

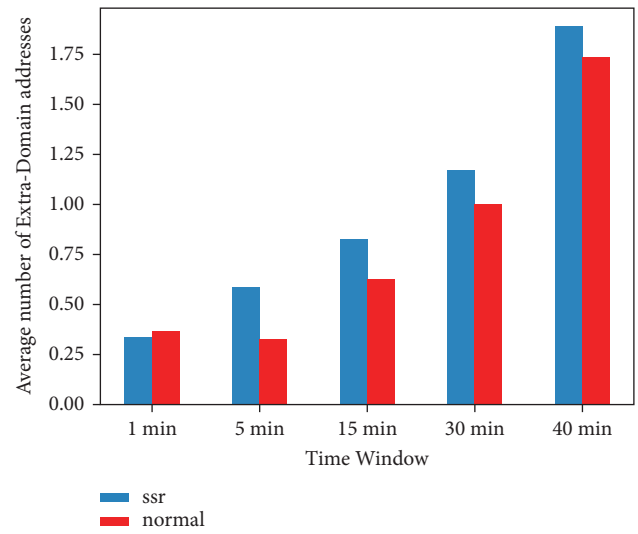

Figure 4: Domain distribution result.

$$D(domain) = \frac{N(extra - domain)}{N(intra - domain)}, \quad (2)$$

where $D(domain)$ represents the intra- and extra-area distributions of the domain, $N(extra - domain)$ represents the number of extra-domains accessed by the users' host within the specified time window, and $N(intra - domain)$ represents the number of intra-domains accessed by communicating with the users' host within the specified time window.

Table 3 shows the feature set constructed based on network entity intelligence information.

4.2.2. Network Behavior Analysis. To improve the accuracy of the classification results, it is not sufficient to rely only on the feature vector constructed by the network entity intelligence information. Network behavior analysis is a network traffic analysis method that relies on the periodic behavior features of network communication of different encryption protocols. At present, network behavior analysis has been widely used in the research of proxy traffic discovery. In this paper, the network behavior analysis method is used from the two aspects of communication target and communication process. The extracted features are shown in Table 4, in which s represents the statistical characteristics of the sender, r represents the statistical characteristics of the receiver, and t represents the statistical characteristics of the bidirectional conversation flow.

4.3. Detection Framework. To provide network managers with information about suspected SSR users to facilitate network supervision within an organization and enhance the defense capability of the network within that organization, we propose a detection method for SSR users based on network entity intelligence. The model constructs feature vectors by introducing network entity intelligence and network behavior analysis and then uses machine-learning classification algorithms to automatically identify SSR users.

Table 3: Communication stream features.

Portrait feature	Feature description
$D(ip)$	Average ratio of extra_ip and intra_ip per minute
$D(Domain)$	Average ratio of extra_domain and intra_ip per minute

According to the feature analysis process, the detection framework for SSR is as shown in Figure 5, which should include four modules: data preprocessing, feature extraction, machine-learning classification detection, and result alerting.

The final detection target in the present work is SSR users. The author's campus network was the test environment used in the experiments. Traffic collection was performed on the host running the SSR software and the host without the proxy to obtain the experimental data.

The locally collected network data packets were divided into flows through data-packet restoration to form flow data (NetFlow, including application layer protocol information, such as host information in HTTP traffic and domain in DNS traffic), as the input data of the feature-extraction module.

The feature-extraction module uses two aspects of network entity intelligence and network behavior analysis to extract features and jointly construct feature vectors. Features extracted relying on cyber entity intelligence and cyber behavior include the distribution features of communication objects, domain distribution features, and network behavior analysis characteristics, e.g., the packet length, quantities of packets, and session duration.

The machine-learning classification module takes the feature vector constructed by the feature-extraction module as input and selects the appropriate machine-learning

Table 4: SSR user detection portrait feature set.

Feature name				
s_packet_count	s_packet_length	s_duration	s_avg_packet_count	s_avg_packet_length
s_min_packet_length	s_max_packet_length	s_std_packet_length	s_std_packet_count	s_packet_tilt
s_min_packet_count	s_max_packet_count	$s\_Count(co)$	$s\_D(ip)$	$D(Do\ main)$
r_packet_count	r_packet_length	r_duration	r_avg_packet_count	r_avg_packet_length
r_min_packet_length	r_max_packet_length	r_std_packet_length	r_std_packet_count	r_packet_tilt
r_min_packet_count	r_max_packet_count	$r\_Count(co)$	$r\_D(ip)$	
t_packet_count	t_packet_length	t_duration	t_avg_packet_count	t_avg_packet_length
t_min_packet_length	t_max_packet_length	t_std_packet_length	t_std_packet_count	t_packet_tilt
t_min_packet_count	t_max_packet_count			

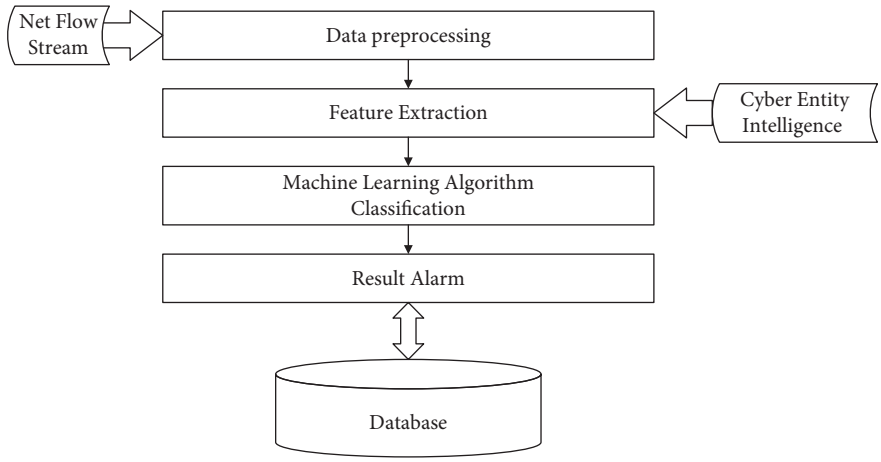

Figure 5: SSR user detection framework.

algorithm as the classifier to complete the classification of the traffic of SSR users. We determine the final classification model by comparing the classification effects of gradient boosting tree (GBT), decision tree (C4.5), random forest (RF), naive Bayes algorithm (NB), and support vector machine (SVM). Regarding the results, the alarm module traces the online authentication information of the suspicious users according to the classification result (records the user's online time, unique identification number, assigned IP, and other information). The user identification code information of the SSR user can be associated with the SSR user traffic information, Internet access time, and IP information. Finally, the obtained suspicious SSR user detection results are stored in the database and an alarm is implemented.

5. Experiments and Analysis

5.1. Experimental Environment and Dataset. The experiments in this paper were carried out on a Big Data platform provided by the author's team. The traffic data used in the experiment were the pCap file collected by running WireShark on the host in the laboratory, including 24 GB of normal traffic data and 10 GB of SSR user traffic data.

The experiment started from the original pCap file from the parsing program, which constitutes a unidirectional data-packet flow from the source IP address to the destination IP address. The data of each data flow include source IP, source port, destination IP, destination port, number of packets, packet length, session duration, protocol number, and domain.

5.2. Evaluation System. At this stage, the evaluation index of classification technology is close to maturity, and there is a specific and recognized index system. To fully represent the accuracy of classification techniques, three metrics are usually used to quantify the performance of a classifier: precision, accuracy, and true positive rate (TPR), which is also known as recall and precision. Their formulas are as follows:

$$\begin{aligned} \text{precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ \text{FPR} &= \frac{\text{FP}}{\text{FP} + \text{TN}}, \end{aligned} \tag{3}$$

where TP is the number of instances correctly classified as positive by the classifier, TN is the number of instances correctly classified as negative by the classifier, FP is the number of instances incorrectly classified as positive by the classifier, and FN is the number of instances incorrectly classified as negative by the classifier. Precision represents the proportion of in-class flows (the flows of interest in the

Table 5: Algorithm comparison.

Algorithm	Precision (%)	Accuracy (%)	FPR (%)
GBT	98.3	94.8	1.23
DT	90.4	91.1	8.1
RF	92.7	93.7	5.0
NB	97.9	36.7	0.4
SVM	75.0	54.3	23.8

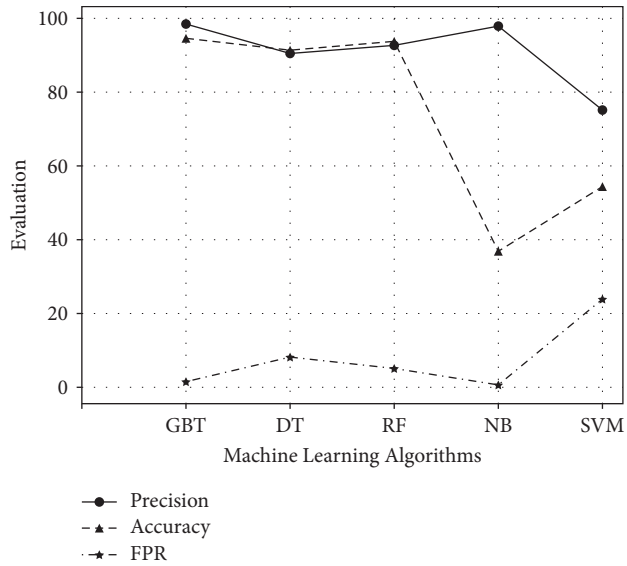

Figure 6: Algorithm comparison result.

present paper) that are correctly classified, accuracy indicates the percentage of correct predictions in the total sample, and FPR rate reflects the proportion of out-of-class (any non-in-class) flows that are misclassified as in-class flows. It is not difficult to find that a high DR and a low FPR are the best cases for classification results.

The accuracy indicators in the existing research are all based on flow (in this paper, we mean SSR flow), that is, the percentage of correctly classified flows to all flows in the network.

5.3. Comparison of Classification Algorithms. The ultimate goal of this paper is to accurately distinguish SSR traffic from non-SSR traffic and then trace the source of suspicious SSR users based on the classification results. The classification algorithm used is a supervised learning algorithm and includes the process of data preprocessing and training of the classification model.

To choose the most suitable classifier for the portrait feature set proposed in this paper, five supervised learning classification algorithms were adopted, i.e., GBT, C4.5, RF, NB, and SVM, to evaluate the traffic dataset; the results are listed in Table 5.

As shown in Figure 6, it can be found from the detection results that, in the experimental environment of the work reported in this paper, the GBT algorithm shows an outstanding detection effect in SSR traffic identification. Although the accuracy of the naive Bayes algorithm (NB) is second only to that of the GBT algorithm, its accuracy cannot show the desired effect. Based on the above-mentioned classification effects, we finally chose the GBT algorithm as the machine-learning classification algorithm for SSR user traffic identification. Moreover, it is found from the analysis that the features constructed based on network entity intelligence can effectively identify and distinguish SSR communication traffic.

6. Conclusions

This paper is oriented to the scenario in which internal users of an organization use SSR software for anonymous communication. To prevent users from leaking the company's confidential information when using such software, and to discover potential threats within the organization in a timely manner, a method of detecting SSR users based on network entity intelligence is proposed. Given the similarity of the communication between SSR and non-SSR users, we innovatively introduce network entity intelligence information. Based on not abandoning the behavioral features of the traffic itself, the feature vector is formed by introducing network entity intelligence to identify the difference between proxy communication and non-SSR user communication. Furthermore, the network behavior analysis method is used to extract and differentiate the communication behavior of SSR and non-SSR users, to expand the feature space of the SSR user detection model. Finally, a machine-learning classification algorithm is applied to realize automatic SSR communication detection. To choose the most suitable classifier, a comparison experiment of the detection rate of five classifiers was carried out in a real network traffic environment, and finally, the GBT algorithm was selected as the most suitable machine-learning classifier. Experiments show that the detection method proposed in this paper is a universal detection method that can accurately detect SSR users. Experimental results show that the detection method proposed in this paper has a detection accuracy of more than 95% for SSR users in the experimental environment, can accurately distinguish between SSR communication and normal communication, and achieves accurate SSR user detection.

However, since there are no label data collected from the communication traffic of SSR users in the real network, subsequent research will focus on expanding the scale of experimental data and improving the accuracy of the model to identify SSR users in the real network environment.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Huaxin Consulting Co., Ltd. The authors thank all the participants who took part in this study and enabled this research to be possible.

References

[1] V. Aghaei-Foroushani and A. N. Zincir-Heywood, "A proxy identifier based on patterns in traffic flows," in *Proceedings of the 2015 IEEE 16th International Symposium on High Assurance Systems Engineering*, pp. 118–125, Daytona Beach Shores, FL, USA, 08-10 January 2015.

[2] P. Luo, F. Wang, S. Chen, and Z. Li, "Behavior-based method for real-time identification of encrypted proxy traffic," in *Proceedings of the 2021 13th International Conference on Communication Software and Networks (ICCSN)*, pp. 289–295, Chongqing, China, 04-07 June 2021.

[3] Z.-H. Han, X.-S. Chen, X.-M. Zeng, Y. Zhu, and M.-Y. Yin, "Detecting proxy user based on communication behavior portrait," *The Computer Journal*, vol. 62, no. 12, pp. 1777–1792, 2019.

[4] S. Miller, K. Curran, and T. Lunney, "Detection of anonymising proxies using machine learning," *International Journal of Digital Crime and Forensics*, vol. 13, no. 6, pp. 1–17, 2021.

[5] P. Fu, "Towards aggregated features: a novel proxy detection method using NetFlow data," in *Proceedings of the 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 409–416, Yanuca Island, Cuvu, Fiji, 14-16 December 2020.

[6] Z. Chen, P. Zhang, C. Huang, Q. Liu, and L. Xing, "A web proxy detection method based on multiple feature analysis," *Journal of Cyber Security*, vol. 3, no. 4, pp. 41–53, 2018.

[7] Y. Zhang, J. Chen, and K. Chen, "Network traffic identification of several open source secure proxy protocols," *International Journal of Network Management*, vol. 243, p. 987, 2019.

[8] Z. Deng, Z. Liu, Z. Chen, and Y. Guo, "The random forest based detection of shadowsock's traffic," in *Proceedings of the 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, pp. 75–78, Hangzhou, China, 26-27 August 2017.

[9] X. Zhang, X. Ma, and X. Han, "An uncertainty-based traffic training approach to efficiently identifying encrypted proxies," in *Proceedings of the 2020 12th International Conference on Advanced Infocomm Technology (ICAIT)*, Macao, China, 23-25 November 2020.

[10] H. E. Hang-song, "Research on Shadowsocks traffic identification based on Xgboost algorithm," *Software Guide*, vol. 17, no. 12, pp. 200–203, 2018.

[11] N. Zliang, T. Wu, Y. Zhang, and M. Xiao, "Shadowsocks traffic identification based on convolutional neural network," in *Proceedings of the 2020 International Conference on Information Science and Education (ICISE-IE)*, Sanya, China, 04-06 December 2020.

[12] X. Zeng, X. Chen, G. Shao et al., "Flow context and host behavior based shadowsocks's traffic identification," *IEEE Access*, vol. 7, pp. 41017–41032, 2019.

[13] Z. Zhuo, Y. Zhang, X. Zhang, and J. Zhang, "Website fingerprinting attack on anonymity networks based on profile hidden markov model," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1081–1095, May 2018.

[14] R. Li, *The Identification of Anonymous Traffic for Shadowsocks Based on Website Fingerprinting*, University of Electronic Science and Technology of China, Chengdu, China, 2016.

[15] Q. Ji, X. Deng, and L. Ni, "Research on ShadowsocksR traffic identification based on Xgboost algorithm," *Emerging Trends in Intelligent and Interactive Systems and Applications*, pp. 53–61, 2021.

[16] Q. Ji, X. Deng, and L. Ni, "Research on ShadowsocksR traffic identification based on DART algorithm," in *Proceedings of the 2021 7th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, pp. 666–672, Guiyang, China, 23-25 July 2021.

[17] J. Luo, L. Bao, and L. Ni, "A method of Shadowsocks (R) traffic identification based on protocol analysis," in *Proceedings of the 2021 IEEE 21st International Conference on Communication Technology (ICCT)*, pp. 6–10, Tianjin, China, 13-16 October 2021.

[18] H. Wang and T.-C. Yang, "Network threat intelligence correlation analysis technology," *Information & Technology*, vol. 45, no. 2, p. 7, 2018.

[19] S. M. Milajerdi, B. Eshete, R. Gjomemo, and V. N. Venkatakrishnan, "POIROT: aligning attack behavior with kernel audit records for cyber threat hunting," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS'19)*, pp. 1795–1812, Association for Computing Machinery, New York, NY, USA.

[20] O. Yurekten and M. Demirci, "Citadel: cyber threat intelligence assisted defense system for software-defined networks," *Computer Networks*, vol. 191, no. 2, Article ID 108013, 2021.

[21] P. Gao, "Enabling efficient cyber threat hunting with cyber threat intelligence," in *Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 193–204, 19-22 April 2021.

[22] X. Rong and D. Song, "Research on cyber-attack defense system based on Big data and threat intelligence," *Journal of Information Security Research*, vol. 5, no. 5, pp. 383–387, 2019.

[23] H. Zhang, Y. I. N. Yan, Z. H. A. O. Dongmei, and B. Liu, "Network security situational awareness model based on threat intelligence," *Journal on Communications*, vol. 42, no. 6, pp. 182–194, 2021.

[24] X. Wang, Y. Wu, and Z.-G. Lu, "Study on malicious URL detection based on threat intelligence platform," *Computer Science*, vol. 45, no. 3, pp. 124–130, 2018.

[25] G. Zhang, *Research and Application of Network Anomaly Detection*, Beijing University of Posts and Telecommunications, Beijing, China, 2019.

[26] M. Adil, M. Attique, J. Ali, A. Farouk, and H. Song, "Hopctp: a robust channel categorization data preservation scheme for industrial healthcare Internet of Things," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 10, pp. 7151–7161, 2022.

[27] M. Adil, "Three byte-based mutual authentication scheme for autonomous Internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9358–9369, 2019.