2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE) | 978-1-6654-2314-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICMCCE51767.2020.00068

Detecting DNS Tunnel Based on Multidimensional Analysis

Kui Jiang[1,a]
[1]Information Center, Shenzhen University, Shenzhen, China
[a]e-mail: jiangkui@szu.edu.cn

Fei Wang[2,b*]
[2]College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China
[b]e-mail: 1810262052@email.szu.edu.cn
* Corresponding author: 1810262052@email.szu.edu.cn

***Abstract*—DNS (Domain Name System) is an important basic application of the Internet. The network management strategy has less restrictions on DNS protocol, which makes DNS become one of the means for attackers to establish covert channel for malicious activities. Taking the DNS tunnel as research object, through multidimensional analysis, this paper mined the characteristics of DNS tunnel in domain name, packet and traffic dimension, and proposed a DNS tunnel identification method which combined machine learning classification and anomaly detection. The experiments proved that this method has high accuracy and low false rate and it can effectively detect DNS tunnels.**

Keywords-DNS tunnel; multidimensional analysis; machine learning; anomaly detection

I. INTRODUCTION

DNS (Domain Name System) is an important infrastructure which is a key link to ensure the normal operation of the internet. Its main function is to translate domain names to IP addresses. Due to the importance of DNS in the internet, network management strategies usually impose a few or no restrictions on DNS, and DNS protocol has become a common means to establish covert channels [1].

DNS tunnel is a method which uses DNS query/response process to establish covert channel for data transmission. As shown in Fig.1, DNS tunnel consists of two parts, the client and the server. When the client can communicate with any DNS server, the DNS tunnel works in direct mode. At this time, the client transmits data with the server through UDP protocol. When the network policy is strict, the DNS tunnel can work in the relay mode. The client program encapsulates the encrypted or encoded communication data in the domain name field of the DNS query packet, and sends it to the DNS authorization server which is the server of DNS tunnel through the multi-level DNS server in the Internet. The server program resolves the DNS query, encapsulates the communication data in the response packet, and returns it to the client. Therefore, DNS tunnel work in relay mode requires a controllable domain name and an authorized server to ensure that the encapsulated DNS query can be distributed to the tunnel server [2].

The original intention of developing DNS tunnel tools was to bypass WIFI login authentication and achieve free Internet access. However, these tools are often used by network attackers for malicious activities. For example, a botnet user can use DNS tunnel for remote control [3]. The attacker encapsulates the instructions of the remote server in DNS packets. Some attackers also use DNS tunnel to exfiltrate sensitive data from internal networks [4]. In APT attacks, DNS tunnel is also one of the methods that attackers use for internal and external network communications [5].

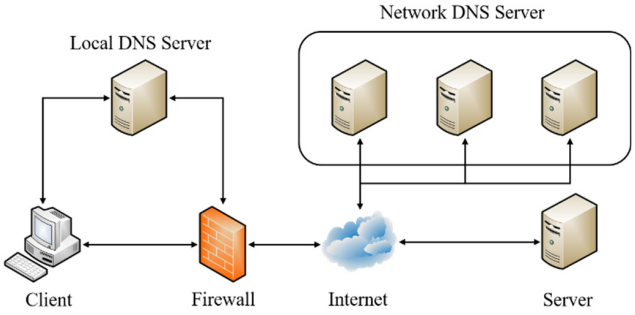

Figure 1. Architecture of DNS covert channel

Because of the great harm of DNS tunnel, the detection of DNS tunnel has become an important issue in the field of network security. Researching the detection of DNS tunnel and discovering covert channel in time can reduce or eliminate the harm it brings, which is of great significance to the safe and stable operation of the network system.

II. RELATED WORK

In the past research, various methods have been proposed to detect DNS covert channel. The detection methods of DNS tunnel can be divided into two types: payload analysis and traffic analysis.

The payload analysis mainly finds the characteristics of the tunnel from the DNS packet. Detecting DNS tunnel by analyzing the characteristics of domain name length, DNS query type, DNS packet length, etc. Born et al. [6] proposed a method based on character frequency analysis to detect the DNS tunnel. The normal domain name conforms to the Zipf's law, while the DNS tunnel's domain name conforms to the random distribution. Qi et al. [7] designed a DNS tunnel detection method based on bigram. Reference [8] used feedforward neural network to detect DNS tunnel through payload analysis. Zhang et al. [9] proposed a DNS tunnel detection mechanism based on deep learning, which can identify covert channels from a single DNS request.

Traffic analysis continuously analyzes multiple DNS packets, and detects DNS tunnels through statistical characteristics of DNS traffic. Aiello et al. [10] used statistical characteristics such as packet size and packet arrival time as features for detection. Reference [11] dynamically monitored

978-1-6654-2314-4/20/$31.00 ©2020 IEEE
DOI 10.1109/ICMCCE51767.2020.00068

DNS data in mobile network, and detects DNS tunnels based on SVM classification. Luo et al. [12] extracted features from DNS session to detect DNS tunnels.

The detection of a single DNS packet may cause false positives to some legitimate domain names, and the detection based on traffic analysis also ignores anomaly DNS traffic may be caused by other reasons. In order to improve the detection accuracy of DNS tunnels and reduce the possibility of false positives, this paper combines the payload analysis and traffic analysis, and fully considers the nature of multiple dimensions such as DNS domain name composition, DNS packet and traffic characteristics, used machine learning classification and anomaly detection to detect DNS tunnel.

III. DETECTION METHOD BASED ON MULTIDIMENSIONAL ANALYSIS

A. Multidimensional analysis of DNS tunnel

In order to find the difference between normal DNS and DNS tunnel, it is necessary to analyze the behavior of DNS tunnel from multiple dimensions. This paper analyzes the difference between normal DNS and DNS tunnels from the three dimensions of domain name, packet, and traffic, extracts the features that can be used to detect DNS tunnels.

1) Domain name dimension

a) Domain name length: The normal domain name is usually short in length for the convenience of memory, while the DNS tunnel uses a longer domain name in order to transmit as much data as possible.

b) Domain name level: According to the DNS protocol, a domain name is a label sequence composed of English letters, numbers and "-". The highest-level domain name is written on the right. Each level of domain name is separated by "." (dot). The length of each level of domain name is not more than 63 characters. Normal DNS query domain names mostly use second-level, third-level, and fourth-level domain names, with fewer domain names at other levels. While the domain names in DNS tunnel are basically above fourth-level.

c) Character frequency: The DNS tunnel domain names are generated by encoding or encryption. The number of English letters and numbers in the domain name is relatively close, and some DNS tunnel tools use abnormal characters such as "+" and "=", which may not find in normal DNS.

d) Entropy: Information entropy is an indicator proposed by Shannon to measure the randomness of the sample. The randomness of domain name can be measured by calculating the entropy of domain name. $H(x)$ is the entropy of the sample and $p(x)$ is the probability of the occurrence of sample x.

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \quad (1)$$

2) Packet dimension

a) DNS Packet size: The normal DNS query packet only contains the information of the query domain name. In order to carry more information in one transmission, DNS tunnel will use a longer domain name for communication. As a result, its packet size will be significantly larger than normal DNS.

b) DNS Query type: In DNS query packet, different types of request records can be used to query different types of data related to the domain name. Most of the DNS request record types are A and AAAA records in normal network. DNS tunnel usually use uses TXT, CNAME, MX, NULL and other unusual request record types in DNS query packet.

TABLE I. FEATURES LIST

Feature Dimension	Feature Name	Description
Domain	Domain Length	Length of domain
	Domain Level	Level of domain
	Ratio of English Character	Ratio of English character in domain
	Ratio of digit Character	Ratio of digit character in domain
	Ratio of other Character	Ratio of other character in domain
	Entropy	Entropy of domain
Packet	Query Type	DNS query type
	Packet Size	DNS packet size
DNS session	Duration	Duration of DNS session
	Packet Number	Packet number of DNS session
	Total Data Size	Size of data transmitted in DNS session
	Avg Packet Size	Average packet size of DNS session

3) Traffic dimension

a) DNS Session duration: DNS session is a kind of UDP session. UDP is a connectionless transmission protocol, and the duration of UDP session cannot be strictly defined. Therefore, this paper defines the duration of DNS session as the time interval between the first DNS packet and the last DNS packet in a DNS session. During normal DNS resolution, the duration of the session is short. DNS tunnel usually closes the UDP socket after the end of the communication. During the communication process, a specific UDP port is occupied for a long time. Therefore, the DNS session duration of the DNS tunnel will be longer than the normal DNS session.

b) Packet number and data size: The amount of data that a single DNS query packet can transmit is very limited, so DNS tunnel will send a large number of DNS packet for data transmission. In a DNS session, the total number of data packets transmitted by the DNS tunnel is large, and the data size transmitted in the entire session is also large. However, a normal DNS session will end after the DNS resolution is completed. The packet number and data size in normal DNS session is shorter than DNS tunnel.

4) Feature Engineering

Through multidimensional analysis of DNS tunnel, we extracted twelve features in the three dimensions of domain name, packet, and traffic to identify DNS tunnel, as shown in Table I. The features in domain name and packet dimensions can be directly extracted or calculated from a single DNS query packet. The traffic features need to be obtained by counting the relevant packets of the 5-tuple of the DNS session.

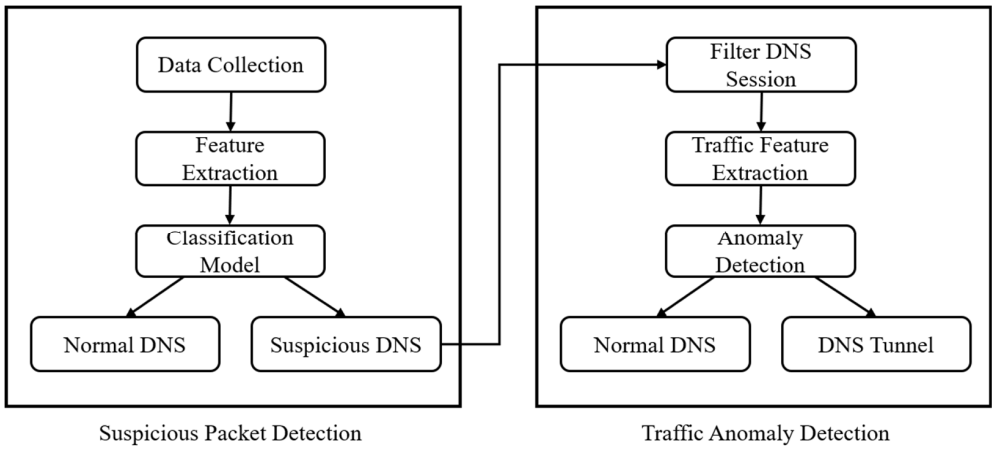

Figure 2. The architecture of detection model

B. Detection method

The DNS tunnel detection method based on multidimensional analysis is mainly composed of two modules: DNS suspicious packet detection and DNS session traffic anomaly detection.

The packet detection module is responsible for screening out suspicious DNS packet. First, collect the DNS traffic from the network outlet, extract the eight features of the domain name and packet dimensions from a single DNS packet to form a feature vector. Then, input the feature vector into the trained classification model. The classification model outputs the DNS query packet that may belong to the DNS tunnel. Detection of a single DNS packet may cause false positives for some legitimate long domain names. However, normal DNS and DNS tunnel have obvious differences in traffic. Therefore, we can use session traffic to further distinguish the normal DNS and DNS tunnel.

The anomaly detection module is based on One Class Support Vector Machine (OCSVM) algorithm. Firstly, the anomaly detection module selects DNS session according to the five-tuple information of suspicious DNS packet, including source IP address, source port, transport layer protocol, destination IP address and destination port. Then, Calculate the four traffic characteristics of the corresponding DNS session duration, number of packets, data size and average packet size. Using anomaly detection model to determine the abnormality of the DNS session traffic. If the DNS session traffic is abnormal, the DNS packets in the five-tuple belong to the DNS tunnel.

In the experiment, we will use the common machine learning algorithm to train the classification model in the packet detection module, and select the model for packet detection by comparing the detection effect.

IV. Experiment and Analysis

A. Assumption

After analyzing the existing DNS tunnel traffic, the detection method proposed in this paper will be based on the following assumptions:

1) The DNS tunnel uses a specific single domain name in the communication process.

2) During communication, the UDP socket created by DNS tunnel will not be closed until the end of communication.

B. Dataset

The normal DNS traffic used in the experiment was obtained by collecting one day's DNS traffic at the network outlet of our lab. The DNS tunnel traffic samples were generated by DNS tunnel tools (dns2tcp, iodine and dnscat2). The dataset has a total of 89210 DNS query packet samples which contains both normal DNS packet and DNS tunnel packet. Among the dataset, 80% of DNS query packets are used for classification model training, and 20% of DNS query packets are used for model performance testing. Select 15000 normal DNS session samples from normal DNS traffic and extract session features for the training of anomaly detection model.

TABLE II. DATASET

Data Source	Sample Size
lab DNS	85442
dns2tcp	53722
iodine	15178
dnscat2	22738

C. Classification model comparison

In this paper, we used decision tree (DT), logistic regression (LR), support vector machine (SVM) and random forest (RF) to construct classification models and compared these model's performance. We use accuracy, precision, recall, F1 score four indicators for evaluation. The confusion matrix is shown in Table III.

TABLE III. CONFUSION MATRIX

Prediction / Actual	DNS tunnel	Normal DNS
DNS tunnel	TP	FN
Normal DNS	FP	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

As shown in Table IV, the classification model based on random forest is higher than other models in accuracy, precision, recall and F1 score, and better than other classification models in performance. The decision tree can often adapt to the situation of less features, and the features used in this paper are relatively less, as a result, the decision tree has better performance than logistic regression and support vector machine. Random forest is composed of multiple decision trees, which is a kind of ensemble learning. Compared with decision tree model, random forest model has strong anti-interference ability and can reduce the impact of outliers. Due to its randomness, it also reduces the risk of model over fitting. Therefore, this paper used random forest to train classification model for DNS suspicious packet detection.

TABLE IV. COMPARISON OF CLASSIFICATION MODEL

Classification model	Accuracy%	Precision%	Recall%	F1
DT	99.00%	99.41%	98.64%	0.9902
LR	93.20%	91.69%	95.47%	0.9354
SVM	98.26%	99.36%	97.25%	0.9829
RF	99.39%	99.43%	99.38%	0.9941

D. Analysis of DNS session anomaly detection

After DNS suspicious packet detection, 52 normal DNS query packets were misjudged as DNS tunnel packets. After anomaly detection based on one-class support vector machine, the results shown that these 52 packets belonged to normal DNS session traffic, while the five-tuple session traffic corresponding to the other DNS tunnel packets were all abnormal. Therefore, the anomaly detection model based on OCSVM can effectively eliminate the false positive of normal DNS packet, and the entire detection method has a false positive rate of 0.

TABLE V. THE RESULT OF ANOMALY DETECTION

Method	Accuracy%	Precision%	False Positive
RF	99.39%	99.43%	52
RF+OCSVM	99.68%	100%	0

Compared with a single random forest detection method, the detection method combined with random forest and one-class support vector machine has higher accuracy and lower false positives, and it can accurately identify the DNS tunnel in DNS traffic.

V. CONCLUSION

This paper introduces a method for detecting DNS tunnel based on multidimensional analysis. The random forest algorithm is used to fit classification model, and the suspicious packets that may belong to the DNS tunnel will be screened out from the DNS query packets. Then we combined with an anomaly detection model based on one-class support vector machine to identify abnormal DNS session traffic. Experimental results show that our detection method has a high accuracy rate, can accurately detect DNS tunnel with low false positive.

REFERENCES

[1] S. Zhang, F. Zou, L. Wang, M. Chen, "Detecting DNS-based covert channel on live traffic," Journal of Communications. vol. 34, pp. 143-151, May 2013.

[2] Jiang kui, Wang fei, Zhang wei, Hu huazhou. "Design and implementation of a reverse intelligent DNS system," China Education Network, 2019(11):65-67.

[3] Jin, Yong, Hikaru Ichise, and Katsuyoshi Iida. "Design of Detecting Botnet Communication by Monitoring Direct Outbound DNS Queries," International Conference on Cyber Security and Cloud Computing, New York, USA, 2015, pp. 37-41.

[4] Das A, Shen M Y, Shashanka M. "Detection of Exfiltration and Tunneling over DNS," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp.737-742.

[5] Nuojua Viivi, Gil David, Timo Hamalainen. "DNS Tunneling Detection Techniques – Classification, and Theoretical Comparison in Case of a Real APT Campaign," Internet of Things, Smart Spaces, and Next Generation Networks and Systems. Springer, Cham, 2017: 280-291.

[6] Born, Kenton, David Gustafson, "Detecting DNS Tunnels Using Character Frequency Analysis," arXiv preprint arXiv:1004.4358, 2010.

[7] Qi C, Chen X, Xu C. "A Bigram based Real Time DNS Tunnel Detection Approach," Procedia Computer Science, 2013(17), pp. 852-860.

[8] Yakov Bubnov. "DNS Tunneling Detection Using Feedforward Neural Network," European Journal of Engineering Research and Science, England, vol. 3, pp. 16-19, November 2018.

[9] J Zhang, L Yang, S Yu, J Ma. "A DNS Tunneling Detection Method Based on Deep Learning Models to Prevent Data Exfiltration," Network and System Security, 13th International Conference, NSS 2019, Sapporo, Japan, 2019, Proceedings pp.520-535.

[10] Maurizio Aiello, Maurizio Mongelli, Gianluca Papaleo. "Basic classifiers for DNS tunneling detection," 2013 IEEE Symposium on Computers and Communications ISCC, Split, Croatia, 2013, pp. 880-885.

[11] Van Thuan Do, Paal Engelstad, Boning Feng, Thanh van Do. "Detection of DNS Tunneling in Mobile Networks Using Machine Learning," International Conference on Information Science and Applications(ICISA), Macau, China, 2017, pp. 221-230.

[12] Luo youqiang, Liu shengli, Yan meng. "DNS tunnel Trojan detection method based on communication behavior analysis," Journal of ZheJiang University (Engineering Science), 2017, 51(9): 1780-1787.