Journal of Theoretical and Applied Information Technology
15 September 2012. Vol. 43 No.1

ISSN: **1992-8645** www.jatit.org E-ISSN: **1817-3195**

DETECTION METHOD FOR NETWORK PENETRATING BEHAVIOR BASED ON COMMUNICATION FINGERPRINT

[1]**ZHANGGUO TANG,** [2]**HUANZHOU LI,** [3]**MINGQUAN ZHONG,** [4]**JIAN ZHANG**
[1]Institute of Computer Network and Communication Technology,
Sichuan Normal University, Chengdu 610066, China
E-mail: [1]tangzhangguo@sicnu.edu.cn

ABSTRACT

In order to monitor the use of network transmission software, the network penetrating technique based on encrypted proxy is discussed. By comparing the behavior of related penetration software, the concept of communication fingerprint is introduced to expand the extension of the communication features. The fingerprints database of encrypted proxy software with specific characteristics is constructed, and a heuristic identification system for encrypted proxy software is designed and implemented. Test results indicate that the system runs efficiently and the results are accurate.

Keywords: *Proxy; Encrypted Proxy; Network Penetrating; Communication Fingerprint; Network*

1. INTRODUCTION

With the rapid development of Internet, It is particularly important to control Information and content on the Internet. However, in recent years a class of Internet penetration software appeared. Through dynamically proxy server, it can send the encrypted information, and thereby breakthrough Internet blockade so as to avoid supervision. Network penetration technique generally integrated with agent technology, encrypted tunnel technology, multi-hop technology, anonymous communication technology[1,2], it can not only break through the existing safety device to access illegal websites, but also sent attack code, secret data to the destination host. Therefore, the research of network penetration technology and the corresponding monitoring measures is of great realistic significance. In order to prevent the spread of harmful information, generally the firewall is arranged between in the trust and distrust network, which is used for sensitive content filtering such as websites, IP address, key words and URL. However, the weakness of these filtering techniques is powerless to the encrypted information, and the keyword of URL or webpage can be used to encrypt by different methods, so that such information filtering system is fundamentally out of action [3]. In this paper, we focus on the working principle of encrypted proxy. Through the analysis and summarization of communication process and behavior of such software, we mastered the working mechanism and communication fingerprint, on that basis, the generic detection scheme of such software was designed. The rest of this paper is organized as follows: Section 2 addresses related work of network penetrating behavior and encrypted proxy technology. Section 3 describes the background of communication fingerprint. Our detection approach and key technologies rare also explained in section 3. Experiments are explained in section4, followed by conclusions and future work in section 5.

2. RELATED WORK

2.1encrypted proxy technology

The network penetration process based on the encrypted proxy is shown in Figure 1. Encrypted proxy server was placed between the insecure network environment and safe network environment, only when the encrypted communication both use the same protocol, the encryption server port can be connected and accessed by client[4,5,6]. To implement the information transmitting and plaintext-ciphertext conversion between application program and encrypted server, the client needs to run the encrypted proxy software, and through the use of network cryptosystem in the secure channel it can provide customers with safe and reliable data.

2.2The comparison of existing encrypted proxy technology and tools

In order to evaluate the encryption software penetrating ability, we must first understand the different network content filtering principle [7].

Table 1 gives the three main kinds of filtering methods and the corresponding penetration technology.

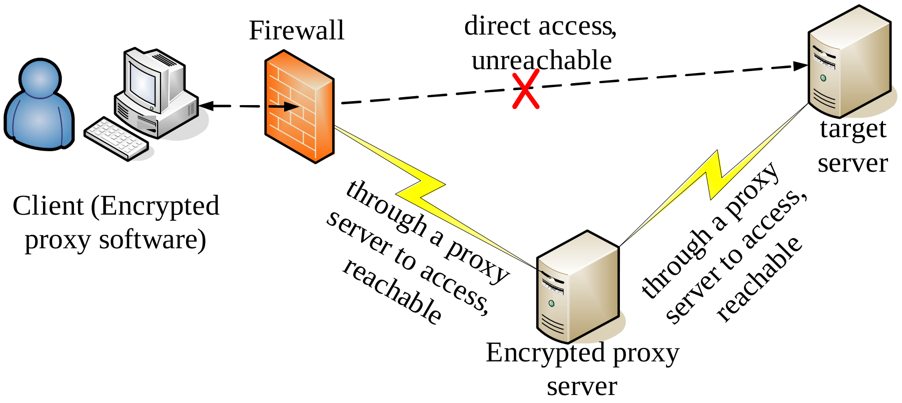
Firewall
direct access, unreachable
target server
Client (Encrypted proxy software)
through a proxy server to access, reachable
through a proxy server to access, reachable
Encrypted proxy server

Figure 1 Schematic diagram of encrypted proxy firewall penetration

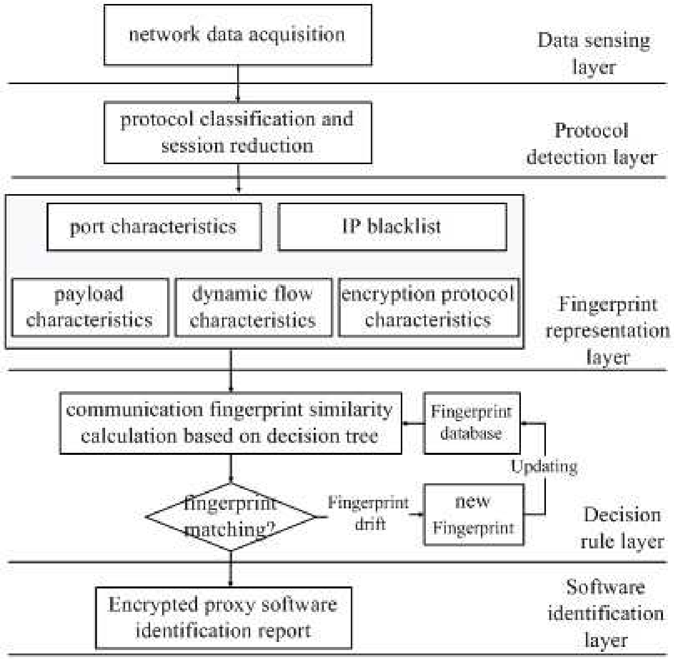
network data acquisition
Data sensing layer
protocol classification and session reduction
Protocol detection layer
port characteristics
IP blacklist
payload characteristics
dynamic flow characteristics
encryption protocol characteristics
Fingerprint representation layer
communication fingerprint similarity calculation based on decision tree
Fingerprint database
Updating
fingerprint matching?
Fingerprint drift
new Fingerprint
Decision rule layer
Encrypted proxy software identification report
Software identification layer

Figure 2 the model's logic diagram

Table 1 Existing content filtering technology and the corresponding penetration technology

content filtering technology	penetration technology
DNS request filtering	Transmitted DNS request packet through an encrypted channel or covert channel
Webpage content filtering	Transmitted Web traffic through an encrypted channel or covert channel
IP address filtering	Transmitted all traffic through an encrypted channel or covert channel

Table 2 comparison of commonly used tools for penetration

	Ultra Surf	Gaps	Frigate	Garden
Transmitted DNS request packet through an encrypted channel	supported	supported	unsupported	supported
Transmitted Web traffic through an encrypted channel	supported	supported	supported	supported
Transmitted all traffic through an encrypted channel	supported	supported	supported	supported

The breakthrough effect of network blocking software by different encryption methods is different. In view of the existing mainstream network encrypted proxy software, we presented the comparative test results of the network penetration method and capability, as shown in Table 2. As can be seen, not all of the tool will encrypt all of the content, for example, the Free Gate software does not encrypt the DNS request packet, and this approach provides the appropriate means of detection.

3. DETECTION METHODS FOR NETWORK PENETRATING BEHAVIOR BASED ON COMMUNICATION FINGERPRINT

3.1 Working mechanism and process

The detection model works on the network gateway, aiming to detect the use behavior of encrypted proxy software through the analysis of network data flow. The working mechanism and the process are shown in Figure 2. In order to accurately identify different types of encrypted proxy software, here we introduce the concept of communication fingerprints, which contains five attributes, such as port features, IP features, protocol encryption features, dynamic flow characteristic and the packet payload characteristics. To take into account the efficiency and accuracy, the detection mechanism should be stratified. In this paper the detection engine is divided into three levels, among the port, IP is a wide range of blind scan layer, followed by the intention detection layer composed by protocol encryption behavior and dynamic flow behavior. On the basis of the two above detection layers, exact matching layer uses the packet's payload characteristics for contenting detection. Blind scan layer, intention detection layer and accurate matching layer investigate step by step from three levels that is the communication protocol, communication behavior and the communication content. Generally speaking, with the detection level layer depth the performance cost increases. However, this hierarchical heuristic detection architecture can not only ensure a very high detection accuracy, versatility and adaptability[8], but also significantly reduces the collect quantity and analysis, at the same time make it possible for rapid detection and accurate recognition under the complex background noise.

3.2 Part of the key technologies

3.2.1 Formal description method of communication fingerprints characteristics

In order to determine what kind of network characteristics can be used as network fingerprint of specific software [9], we defined the concept of network fingerprint model and gave the general formal description.

[Definition] Network fingerprint model: network fingerprint model can be abstracted as a 5-tuple array, expressed as $FP = < P_I, P_O, C, O, F >$. Among them, P_I represents a set of input data packages, P_O represents the set of output or response packages, $C = \{c_1, c_2, \cdots, c_n\}$ represents network character set, $O = \{o_1, o_2, \cdots, o_m\}$ represents the detecting objects, such as, $O = \{\text{OS type}, \text{URL}, \mathit{Payload}\}$. The fingerprint detection function F is defined as: $P = P_I \times P_O = \{(x, y) | x \in P_I, y \in P_O\}$ is a directed ordered pair, $Z = \rho(P)$, then F is mappings of the power set of the Cartesian product on the P_I and P_O, that is, $F : Z \to C$. Accordingly, we got the necessary and sufficient conditions whether network feature can be used as a fingerprint model, such as:

Testability: For the detection of object O, $\forall z \in Z$, there existed $c \in C$,which make $F(z) = C$;

Uniqueness: For the detection of object O, to arbitrary $Z_1 \in Z$, $Z_2 \in Z$, and $z_1 \neq z_2$,then we got $F(z_1) \neq F(z_2)$, and vice versa.

Stability: For the detection of object O, to arbitrary $Z_1 \in Z$, $Z_2 \in Z$, and $z_1 \neq z_2$, there existed $c \in C$,which make $F(z_1) = c$, $F(z_2) = c$;or there existed $C_1 \in C$ 、 $C_2 \in C$, and $c_1 \neq c_2$, which make $F(z) = c_1$ and $F(z) = c_2$ occurred simultaneously.

Separability: For any two different network communication software i and j, exerted the same detection function on the same object, that is, $FP_i = < P_{Ii}, P_{Oi}, C_i, O, F >$, $FP_j = < P_{Ij}, P_{Oj}, C_j, O, F >$, $i \neq j$,then $c_i \neq c_j$

In order to capture the communication behaviors of encrypted proxy software, we can use a kind of reverse direction analysis method, or black box testing method to analysis communication packets. Figure 3 gives the operation and communication mechanism of the Free Gate software when it runs on a computer the first time. First, the Free Gate started, and then tested environment during the startup, then the specific data were generated, through a variety of queries the Free Gate obtained proxy server information, after accessing to information of proxy server it began to establish the process of encryption communication, if it can successfully start its communication. In order to ensure that data will not intercepted by third parties, the proxy server and client devices need to communicate securely, which is divided into two parts involved by secret key establishment and secure communications. Thus, it can be summed up that the network activity of network penetrating software is divided into three stages [10, 11]:

Detecting. Access to domestic and foreign well-known sites is used to detect whether the internet can be access.

Access. Access to full-time DNS server provides updates information support.

Encrypting. Access to the encrypted proxy server usually uses SSL encryption.

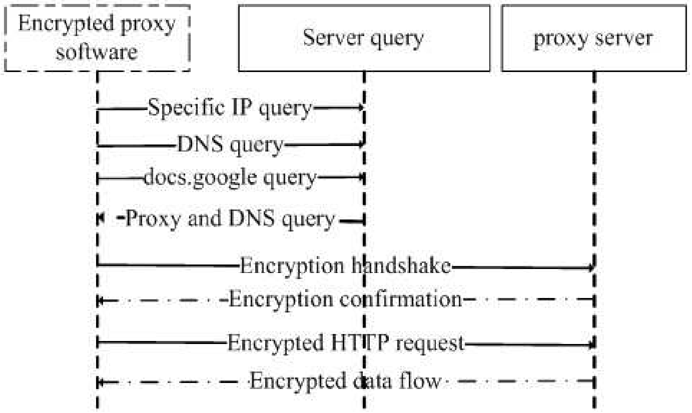

Figure3. Communication sequence diagram of encrypted proxy software

Through the above analysis we can find that the communication packets of encrypted proxy software have the two categories of features, which are payload characteristic code and dynamic flow anomalies. Therefore, we used both deep packet inspection (DPI) and dynamic flow inspection (DFI). DPI not only analyzed the 5-tuple of IP header, but also added the application layer payload segment analysis, thus it can not only accurately recognize encryption agent content, but also deal with the escape behaviors such as port-reuse, random port or even the encrypted transmission. DFI identifies encryption agent traffic via average flow rate, flow duration, number of bytes, packet length and other characteristic information. Through a large number of data packet analysis experiment, we established the flow characteristic model. By capturing the series of flow behavior and comparing with the flow model, thus the use behavior of the encrypted proxy software can be detected. As an example, the key features of Free Gate's communication stream were extracted as follows:

DNS query packet length is generally 558.

Access to most of the IP does not belong to the mainland (among them, China Taiwan and the United States account for a larger proportion).

DNS requests use the same port for 2-3 rounds of queries.

DNS is divided into 5 sections.

Each round of query interval may be equal; each wheel is equal to the number of queries.

HTTP and DNS use serial ports, and DNS request ports greater than 1024.

The DNS name servers and proxy servers end in16.mjuyh.com, such as:

9c932c843ccc51d8ef0d40b371818075016705f.a0ff830ba9a8ddc6834624598e47271d55a31c7.16.mjuyh.com.

Use SSL to encrypt the communication.

3.2.2 Identifying SSL-encrypted data

Almost all encrypted proxy software adopted SSL to encrypt the contents of application layer, as shown in Table 3. Therefore, the use of SSL protocol can be used as an assist feature on the communication behaviors of such software.

SSL protocol stack was shown in Figure 4[12]. Because the sequence encryption in Record Protocol didn’t have any filling mechanism, the protocol ciphertext length exposed the plaintext length, so that flow analysis can be done. On the other hand, the handshake messages transmitted in plaintext before the "Finished" message in the handshake protocol layer, thus it can be used for payload characteristic code analysis [13, 14]. The method is to first filter out the session flow by 443 ports, and then examines the session payload contents to identify whether there exists some corresponding feature strings.

3.2.3 decision trees algorithm

The detecting process of encrypted proxy software based on the network data package is in fact a classification problem to distinguish the encrypted proxy activities from legitimate communications. As an important data mining Technology, classification is designed according to the characteristics of the data set to construct the classification function or model, by which the unknown samples can be mapped in a certain to a given category. Structural model was generally divided into two stages, that is, training and testing, and accordingly the model data sets were randomly divided into training set and test data sets. In the training phase, the training data set was used to construct the classification model by analyzing the data described by the attributes. When in the testing phase, the test data set was used to evaluate the accuracy of the classification model, if that model accuracy is acceptable, then it can be used to classify for other data[15].

In this paper, the decision tree model to detect encrypted proxy software used communication fingerprints as input vector, as shown in Figure 5. Considering that the tree structure was too complex and will produce over-fitting, to avoid this, we should consider the partial ordering relation existing in the importance of properties used to identify encryption agent software, which is we must take the important as principal component and the less important for clipping. In addition, in order to improve the efficiency of algorithm model, we adopted the Hash table instead of a database to store communication fingerprints. The node element of Hash table indicates the communication behaviors of one host, communication behavior data types and Hash types are as follows:

```
Typedef struct    // Defining the node in Hash table
{
Int USER_ID;   // Using IP-MAC as primary key
Char ComFP;   // attribute name of each communication fingerprint
Int TreeNet;    // the level of behavior tree which the communication fingerprint lying on
ElemType* next
}ElemType
Typedef struct      //defying Hash table using HashTab
{
ElemType* next;
Int count;       // key value
Int typeindex;   // key value
} Hash Tab
```

Table 3 test results of common penetrating tool for application-layer encryption

	Ultra Surf	Gaps	Free Gate	Garden
HTTP	ciphertext	ciphertext	ciphertext	ciphertext
HTTPS	ciphertext	ciphertext	ciphertext	ciphertext
Mail	ciphertext	ciphertext	ciphertext	ciphertext
FTP	ciphertext	ciphertext	ciphertext	ciphertext
Google Talk	ciphertext	ciphertext	ciphertext	plaintext
Windows Live Messenger	ciphertext	ciphertext	ciphertext	plaintext

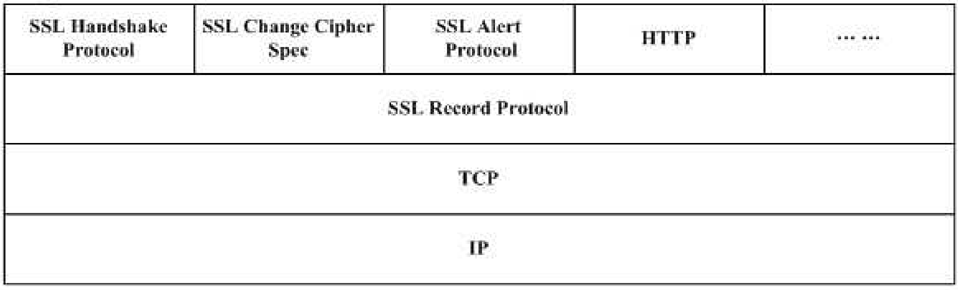

Figure 4 SSL protocol stack

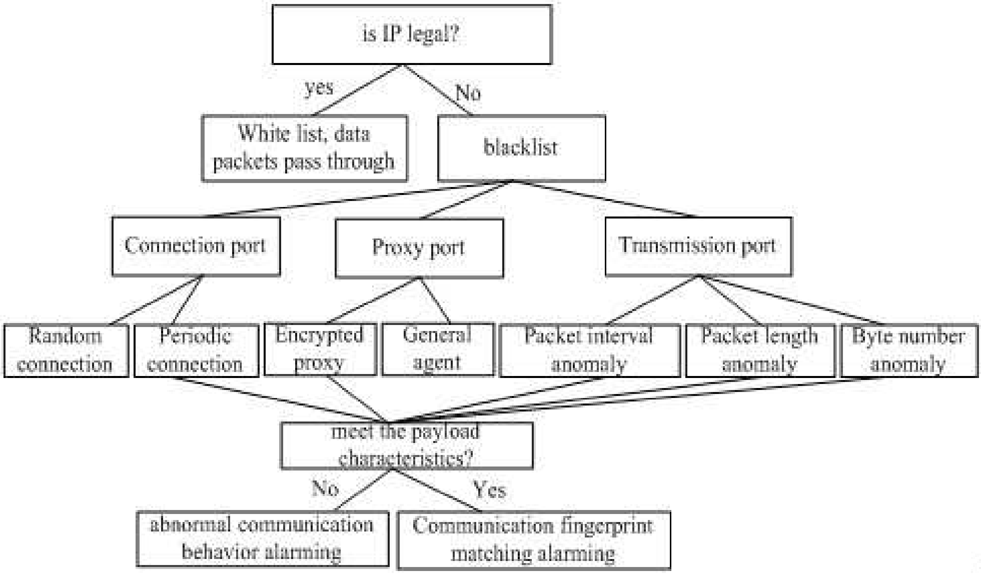

Figure 5 Communication behavior trees and filtering algorithm

4. EXPERIMENT AND RESULTS ANALYSIS

In the detection experiments, we build the network communication environment needed by Ultra Surf browser, deployed network packet capture software on the network export, and then loaded the packets into our filtering algorithm model. In order to simulate the real traffic environment, we used the campus network communication flow as background noise, by importing experimental data captured by Campus Network Center on that day; our model successfully located a plurality of suspicious packets file. One of our tests shows that our model correctly identified the Ultra Surf browser and successfully associated its communication behavior process: host 192.168.1.29 used Ultra Surf browser, after starting, the Ultra Surf browser opened the 2224 port, and then launched frequent periodic connection to its encrypted proxy server, after a successful connection it used HTTPS encryption protocol to communicate with a network host which IP is 118.161.221.171, the host's location was Taiwan.

5. CONCLUSION AND FUTURE WORK

In this paper, we have presented the heuristic detection with multi-feature, introduced the concept of communication fingerprint to extend the range of communication features, and developed decision tree methods to distinguish the encrypted proxy activities from normal traffic. Being expanded slightly, the method can be used for detection and recognition of the other network communication software.

Future work will include:

The expansion of training sets, test sets and the experiments for various kernels which can be use for performance improvement and some of its constraint parameters;

According to the different characteristics of the communication software, the unified, standardized, and formalized description of communication feature will be studied;

Communication fingerprint automated extraction technology is a key research direction of the next step.

6. ACKNOWLEDGMENT

This work was support by the Project of Applied Basic Research of Sichuan Province（No.07JY029-011）, the Project of Department of Education of Sichuan Province（No.12ZB105）.

REFERENCES

[1] L.Yan, H.Z. Li and Z.G.Tang, "Mingquan ZHONG. Techniques of Trojans penetrating personal firewall", *Netinfo Security*, Vol.9, No.20 , 2010 ,pp:48-66.

[2] L. Peng, "Research on Network penetration technique", *Beijing University of Posts and Telecommunications*, 2008.

[3] D.X. Qu, X.T. Tang, L.C. Xu and L.Shi, "Overview of Research on Network Information

Filting System", *Journal of Shandong Normal University*, Natural Science, Vol.22, No.2, 2006, pp 23-26.

[4] Z. Zhou, X.M. Han and B. Wen, "Analysis of Encrypted Proxy Servers Techniques". *Journal of Information Engineering University*, Vol.7, No.4, 2006, pp:336-340.

[5] D.F. Liu and P.C Tan, "Efficient Approach for Searching Data Package of Encrypted Proxy ", *Command Information System and Technology*, Vol.1, No.4, 2010, pp:55-58.

[6] C.L Wang. "Research on the penetrate technology of Boundary network safety protection device". *Network & Computer Security*, Vol.2, 2007, pp:17-20.

[7] J. Smart, K. Tedeschi and D. Meakins, "Peter Hannay & Christopher Bolan.", Subverting National Internet Censorship-An Investigation into existing Tools and Techniques, http://scissec.scis.ecu.edu.au/publications/2008/forensics/Smart%20et%20al%20Bypassing%20Internet%20Censorship.pdf.

[8] L. Martignoni, E. Stinson, M. Fredrikson, S. Jha and J.C. Mitchell, "A Layered Architecture for Detecting Malicious Behaviors", *RAID* 2008 pp:78-97.

[9] Z.G. Tang, H.Z. Li, M.Q. Zhong, J. Zhang. "Study of Remote Computer network Fingerprint model". *Computer Engineering and Design*, Vol.32, No.8, 2011, pp:2592-2595.

[10] Z. Zhou, "Efficient Approach for Searching Data Package of Encrypted Proxy". *Computer Engineering*, Vol.33, No.21, 2007, pp:142-146.

[11] H.T Gao, "Research on investigation and evidence collection of network penetrating software". *Police Technology*, Vol.11, No.6, 2010, pp:43-46.

[12] L. Zhang. "Research and application of SSL-based VPN encrypted tunnel". *Information Technology*, Vol. 8, 2010, pp :200-203.

[13] Y. Luo and Z. Huang, "Principle and Prevention of SSL Attack In Gateway Mode". *Information Security and Communications Privacy*, Vol.4, 2011, pp:50-52.

[14] S.M. Yang and S.D. He, "The SSL protocol connection process and safety performance analysis", *Software Guide*, Vol.10, No.3, 2011, pp:151-153.

[15] N.N. Xie, Y.X Liu. "Improvement of attribute selection criterion of decision trees". *Computer Engineering and Applications*, Vol.46, No.34, 2010, pp: 115-118.