Automation Analytical Model of Familial Malware Based on Communication Features

Yiyang Wang, Shaoming Chen and Chongbin Wu
No. 4 Jianzhong Road, Tianhe District, Guangzhou City, Guangdong Province, China

***Abstract*—Up to date, the Internet industry has done a lot of research on malware analysis, which brings in effective malware detection. However, in terms of familial communication feature extraction and detection, a very mature product has yet to be seen. Given this situation, this paper tries to establish an automatic model for the analysis, extraction of familial communication features which is based on the family clustering results， replay the communication traffic of the same family, classify and analyze communication features of each malware, extract the relatively stable hexadecimal common packet payload value, as the family communication features for network detection. Experimental results show that the proposed analytical can effectively identify the familial malware and its new variants.**

Keywords-familial malware; communication features; automated analysis; feature extraction

I. INTRODUCTION

With the rapid growth of the computer network technology, the Internet is gradually penetrating into the government, industry, education and national defense fields. While the network brings a lot of information easily, it also brings viruses, Trojans, worms and lots of other malware, resulting in hacking attacks, information theft and other security incidents. Most malware attacks are highly hidden and dangerous, seriously threatening the information security of computers. [1]

The current routine malware analysis and detection methods are mainly based on two aspects: single sample host features and communication features. The former can't ensure that all Trojans can be discovered, and it's difficult to deploy on each computer in the network. As long as one single host in the network does not deploy this method, there may be malware and the whole network is in danger. Moreover, there is a certain lag in the detection of new malware. So, in recent years, the security industry began to focus on the signatures of the malware families. Network layer Intrusion Detection System (IDS) and Intrusion Protection System (IPS) can increase malware detection rates by adding special rules. But, IDS is not designed to detect malware's communication behaviors, it is difficult to ensure the detection rate. New malware variants pop up every day, but this single sample feature based method can do nothing to unknown malware. In addition, this method requires a large signature base, seriously affecting the detection efficiency.

This paper mainly focuses on familial malware communication feature extraction methods, and presents automation Analytical Model of familial malware Based on Communication Features (AMCF for short). The model consists of three parts. (1) Family clustering: analyze family clustering based on dynamic and static signatures [2]. (2) Communication feature extraction: analyze familial communication packets, extract communication features and the offsets to further revise and complement the previous family classification. (3) Communication feature Verification: verify the communication traffic, support YARA rules, regular expressions and other forms of matching detection, improving the processing performance and detection rate.

II. RELATED WORK

Malware family determination, i.e., to determine whether different malware samples are derived from the same set of malicious code, developed by the same author and team, and whether they have internal relevance and similarities. Malware analysis is based on its functions and hazards. At present, there are mainly static analysis and dynamic analysis method [3]. Static analysis method disassembles and decompiles the executables, instead of executing the malware, to understand the malware's processes and functions, and obtain static signatures for malware detection and elimination. Dynamic analysis method executes the malware in a controlled environment, analyzing the interactive behaviors between the malware and the environment. This method captures the environment's changes before and after malware execution, learns the malware's instructions or system calls at different layers, so as to approximate the actual malware functions [3].

At present, the research of malware families is mainly through the similarity of the code. Park [4] et al. dynamically capture malware behaviors, construct the system call graph to determine the code similarity. Kinable [5] et al. analyze the malware system call graph via static method, compare the similarity of malware and determine malware families via graph matching. M.Bailey and J.Oberheide [6] et al. optimize the accuracy of automated identification and malware analysis, but sacrifice efficiency to some extent. With a sharp increase in the number of malware, the previous manual analysis method can't cope with the rapid detection and disposal needs. So, the automatic analysis system, which integrates real-time capture, individual analysis, group clustering or familial feature extraction, comes to the stage. YE Y [7] et al. design and implement the AMCS system for automatic malware classification, generate family features based on classification results for host detection. This system uses static analysis method to extract the sample's instruction sequence and frequency, and through the integration of TF-IDF and k-medoids clustering method to classify and extract the family

features. XU Xiao-lin et al. [2] combine the dynamic and static method, study the malware behavior and code segments of massive samples, propose a malware feature clustering based real-time automatic analysis model.

The above methods or systems use different feature extraction methods, analyze and determine unknown malware through classification or clustering methods. But there are some problems, mainly in: 1) The extracted features rely heavily on human experience, have certain limitations in adapting to massive samples and unknown types. 2) The extraction methods and processes are too complex to be automatic, can't realize the automatic feature extraction of large-scale samples rapidly. 3) They are not very effective for malware families with stable communication features but dispersed static features. To solve these problems, we propose an optimized analysis and detection method after analyzing a lot of malware: obtain dump file or unpacked sample files, extract string vector features and execute clustering analysis, to form the family clusters. Use communication feature and algorithm analysis method to obtain the familial features. Finally, use the detection model deployed the LAN gateway to verify the features. Verify and optimize the communication features of each malware family, so as to realize the analysis and detection ability for the whole family [2].

III. MODEL DESIGN

A. System Architecture

The AMCF model is mainly composed of three parts, which are family clustering, communication feature extraction and communication feature detection. The framework of the model is shown in Figure I.

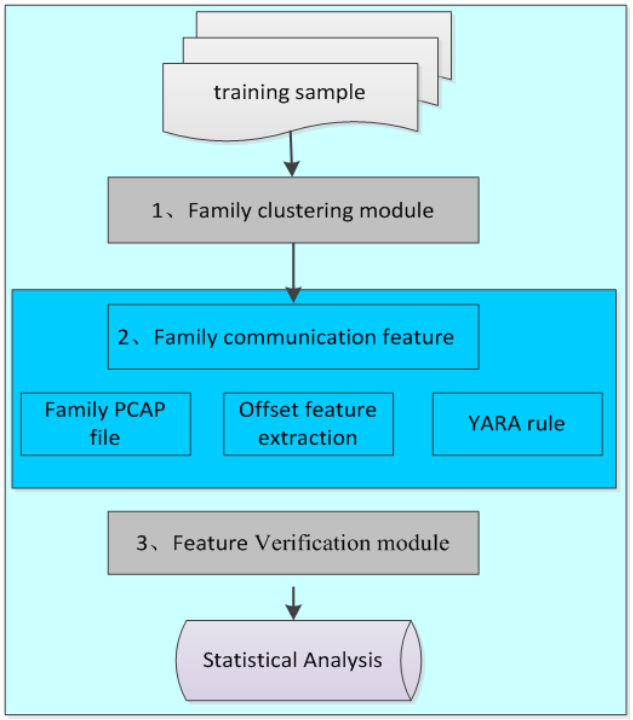

FIGURE I. SYSTEM ARCHITECTURE

B. Family Clustering Module

Deploy sandbox system based on the open source code, implement static format analysis, dynamic execution analysis, memory DUMP file analysis, and construct a mature feature vector set [2]. Use LSH nearest neighbor clustering algorithm to classify feature vectors. Extract familial PCAP files based on the captured PCAP files.

C. Family Communication Feature Extraction

While the extraction carrier of dynamic and static behavior features is the operating system, the extraction carrier of communication features is the network traffic. The vast majority of malware, especially the Trojans, Botnets, their code execution, control, or file transfer requires network connection. Use traffic analysis tools to classify and extract communication features. Based on the family clustering results， replay the communication traffic of the same family, classify and analyze communication features of each malware. Execute samples in various system environments to obtain relevant network communication packets, filter out the differences caused by system environment and put forward the common features. From three stages, connection establishment, connection communication and connection persistence, extract four types of packets: the first packet, the heartbeat packet, the control packet and the file transmission packet. And then, extract the relatively stable hexadecimal common packet payload value, as the family communication features for network detection. [8]

D. Feature Verification Module

Verify familial communication features in virtual environments. Deploy the communication feature detection system in a high performance host. Use the network card to monitor traffic and verify the accuracy of family communication features. In virtual environments with multiple virtual machine and operating systems, execute rule pattern matching through the execution of the family malware, fine-tune and optimize familial communication features based on the verification results.

IV. EXTRACTION ALGORITHM

A. Communication Feature Extraction

According to the principle of network communication protocol, extract various types of packets from PCAP files. The specific description of each packet type is as follows:

1) Extract the first packet: According to the three handshakes of the communication protocol, determine the first packet upon communication connection and extract it.

2) Extract the heartbeat packet: Heartbeat packets is characterized by the regular transmission of fixed packet data, based on which, the heartbeat packets are extracted. The DATA sections of most heartbeat packet files are short, which are less generic than the first packet. There is a certain false alarm rate.

3) Extract the transmission packet: The length of a single packet in network communication behavior is fixed, and the

number of bytes of the file is greater than the maximum length. So, the file is divided into multiple packets for transmission. Based on this, extract the transmission packets, but this kind of packet needs manual analysis and feature extraction.

4) Extract the command packet: It requires capturing in the environment that the server side is active and is sending commands, then you can access to the PCAP file package contains the instruction. Therefore, the PCAP file containing the instruction packet is obtained, which is usually assisted by manual assistance or virtual environment.

5) Give an example: Take the first packet from the sandbox as example, read the corresponding data and compare word-by-word, extract the hexadecimal common feature.

B. Offset Feature Extraction Algorithm

1) Identify the family based on malware name or YARA rules. Use LSH algorithm to aggregate the first packet, heartbeat packets, control packets as well as transmission packet. For example, 50 sample belong to Storm family.

2) Randomly select two samples, cutting them byte by byte, the length m = 4 (optional parameters, the start cutting position and data), such as: 0:1234 (where 0 represents the offset, 1234 represent hexadecimal features), 1:2345, 2:3458.

3) Excludes null data, such as 5:0000.

4) Check the intersection points of two sets. If the number is greater than n = 4 (optional parameter), then generate a set of features to detect the remaining samples.

5) Repeat 2~4 for undetected samples.

6) Merge short features. If the latter feature offset lands in the previous range, 1:2345 0:1234, then merge it into 0:12345.

7) Extract offset family features.

8) Verify the family communication results through network traffic. Data sources: sandbox's data and captured PCAP files.

C. Algorithm Realization

1) The communication features are extracted from the PCAP file and the third party report. After extracting the DATA section from the packet, use the code for data feature extraction, as is shown below:

```
datas[n] as Array
i=0
for x in pcaps do
    datas[i++]=call getPcapData(x)
end
for x in (other third party data) do
     datas[i++]=call getJsonData(x)
End
```

2) Take the first packet as example, get the DATA section from the sandbox system, as shown below:

("author" represents the source, "PCAP" represents the "DATA" segment data, "name' represents the virus name, "MD5" represents the sample hash, "time" represents the analysis time):

{"author": "threatexpert", "pcap": "436F6E9366DA17000000010000001ABA49D92E963F28DB", "name": ["Virus.Win32.Nimnul.a [Kaspersky Lab]", "Virus:Win32/Ramnit.A [Microsoft]"], "md5": ["86bfa2714cf51eb2d128212b3d60c6da"], "time": "2016-01-22 02:35:25"}

3) The feature set is extracted from a hash sample in two different environments (e.g., Winxp, Win7). Compare them; the intersection is the common feature set. Implementation code is shown below:

```
x as struct buf//data in xp
y as struct buf//data in win7
res as struct buf
for j=0,i from 0 to n do
    if x[i].bits!=0 then
            if x[i].bits==y[i].bits then
                res[j++]=x[i]
            end if
    end if
end
if len(res)>=4 then
   call next(res,next_data)
end if
```

V. EXPERIMENTAL ANALYSIS AND VERIFICATION

A. Experimental Procedure

1) Build simulation environment. Realize the automatic malware analysis and familial clustering ability;

2) Focus on the recent popular NITOL family for hash samples collected;

3) Feature analysis and extraction of family communication. According to the algorithm in section 4, the common communication features of NITOL family malware are extracted.

4) Verify the family's detection capability. Use the extracted communication features, the package data of 317 hash samples of NITOL family were detected, objective to detect all family samples.

5) Verify the familial massive data detection capability. Use the extracted family communication features, the data of 1000 hash samples of NITOL family are detected, and the purpose was to detect the majority of family samples.

6) Verify the detection capability against unknown malware families. Random access to 4000 communication PCAP file in the malware library, which contains the NITOL family of samples of packet data, but also contains other family of packet data, for testing the ability to detect unknown malware.

7) Analysis and detection of malware in cryptographic protocol family. Due to the communication data encryption malware, communication features a large change, the extraction of common features is more difficult, only for each type of variants of the artificial protocol to carry out targeted detection, Therefore, it is not in the scope of this study, which will be used as the next step of the research plan to expand the paper

B. Experimental Results of Feature Extraction

Feature extraction is divided into two steps. The first step, for the total sample set (a total of 317) for the 10% sample, Feature extraction algorithm for the NITOL family of 30 PCAP hash data extraction common feature set. The second step, for the failure to detect the package file. Five new common features are extracted from the remain samples. Therefore, the common features of the NITOL family is initially locked into 8, an instance of the common feature is extracted as shown in the following figure:

{"rule": [{"offset":0,"sign":"\\xb0\\x00\\x00\\x00\\x77"}, {"offset":12,"sign":"\\x57\\x69\\x6e\\x20"}]}

An instance of package data hit as shown below:

{"author":"threatexpert","pcap":"B00000007700000009040 00057696E205850205","name": ["DDoS:Win32/NITOL.A[Microsoft]"],"md5":["d2d4a47d9b5f 3bc9b168da46b4c8c1c5"], "time": "2016-04-29 08:29:58"}

C. Verification Result Analysis

1) After the two common feature extraction, the 8 common characteristics of NITOL family were detected and the results were found out, as shown in the following table. As you can see, 8 feature vectors were extracted from the two common features, and 312 samples were successfully detected. The part can't be detected includes 4 encrypted data packets (not the scope of this study). The remaining one sample fails to detect the package. After further analysis, although the static features and the NITOL family are very similar, the network behaviors are quite different. The classification is more suitable for the new family. Therefore, the general detection rate reaches 98.4%, which proves that the feature of Microsoft named NITOL family of samples with better detection results.

2) Randomly extract the first packet of NITOL family through the sandbox system, carry out the detection test. The detection results are as follows:

TABLE I. DETECTION RATE

samples number	100(packages)	1000(packages)
Detection rate	98%	96%

3) The packet data of 4000 hash samples are detected, scan out of 1240 hash family samples. Among them, 468 NITOL samples are found.

4) Detect 329 samples belonging to other families These samples are verified to be NITOL variants or infected samples. It can be seen that although the samples were modified or infected, but also can detect found in the network communication layer.

5) Detect 443 packed families. These samples are packed, proving to be NITOL variant. This shows that even if the malware samples for camouflage and packing etc. But most of its network communication behavior is still consistent, can be detection from the network layer..

VI. SUMMARY

In this paper, based on in-depth study of the current routine malware analysis methods, comprehensive utilization of existing sandbox system and family clustering algorithm, the AMCF model is proposed. Use the offset common feature extraction algorithm, and family communication feature analysis extraction method, automated family communication feature extraction, the virtual environment and the actual network environment, to verify the accuracy of the features, improving the effectiveness of the model. Experimental results show that the proposed analytical and detection method can identify the familial malware and its new variants effectively.

ACKNOWLEDGMENT

This work is one of the projects supported by: National 242 Information Security Program. (2014A005) (2015Z102)

REFERENCES

[1] China Internet network security annual report 2015,

[2] http://www.cert.org.cn/publish/main/upload/File/2015 annual report (1).pdf, 2015.

[3] XU Xiao-lin, YUN Xiao-chun.ZHOU, Yong-lin, KANG Xue-bin, Online analytical model of massive malware based on feature clustering [J]- Journal on Communications 2013(8)

[4] QIAN Yucun, PENG Guojun, WANG Ying, et al. Homology analysis of malicious code and family clustering. Computer Engineering and Applications, 2015, 51(18): 76-81.

[5] Park, Younghee, Fast malware classification by automated behavioral graph matching[C]//Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research, 2010.

[6] Kinable J, Kostakis O, Malware classification based on call graph clustering[J].Journal in Computer Virology,2011,7 (4): 233-245.

[7] Bailey M, Oberheide J, Andersen J Automated classification and analysis of internet malware 2007.

[8] YE Y, LI T, CHEN Y Automatic malware categorization using cluster ensemble 2010.

[9] LI W, LI L H, LI J, et al. Characteristics Analysis of Traffic Behavior of Remote Access Trojan in Three Communication Phases [J]. Netinfo Security, 2015.