Advances in Engineering Research (AER), volume 124
2nd International Symposium on Advances in Electrical, Electronics and Computer Engineering
(ISAEECE 2017)

HTTP Botnet Detection Algorithm Based on Content Association Recommendation

Wang Jiajia [1, a] Chen Yu[1,b]

[1] Taizhou Pylotechnic College, Taizhou, China 225300

[a]wangjiajia_99@163.com [b]xiaoyueryuchen@163.com

Keywords: HTTP protocol，botnet, recommendation algorithm

Abstract. HTTP botnet is widely distributed and causes great harm. The traditional detection method is not analyze the attack data stream until the attack stop. In order to further reduce the harm, according to the characteristics of http/https botnet, an online detection method based on HTTP protocol is proposed, which is based on the content association recommendation algorithm. This method is able to distinguish between normal data and malicious data, and complete detection before the attack start, without increasing the burden of the network because of small complexity. The experiment proves the feasibility of this method.

Introduction

Botnet mainly consists of three components:the hosts which are infected by the bots, command and control(C&C) servers and botmaster. Through a botnet, the botmaster could use the C&C servers to control all infected hosts to attack any target in the absence of the owners. There are various types of attacks, include DDOS, spam, steal sensitive information and so on[1]. In October 21, 2016, the Dyn network attack was launched by the botnet controlled by Mirai. By now, China has become the country that has the most zombie hosts in the world[2]. Therefore, it is emergency to detect the existence of botnet and eliminate its adverse effects.

In the early Botnet, the IRC protocol was mainly used to distribute the botmaster's commands from the C&C server to the hosts. Although the IRC botnet is easy to implement and easy to manage, but it has some limitations, such as plaintext transmission, continuous connection, single point of failure, etc, there are many solutions to this kind of botnet up to present[3-4]. In order to overcome the above problems, the P2P botnet has emerged. To replace the traditional single C&C servers, the botmaster sends commands to one or more zombie hosts, then sends them to their neighbors by the associated zombie hosts. The P2P botnet has its advantages, eg. more concealment, but it is difficult to implement, with high complexity, so the application scope is not very wide. At present, there are some solutions for such botnets[5-6].

Now, the type of Botnet is back to the centralized C&C server mode, using HTTP or HTTPS as the communication protocol between the C&C server and the zombie hosts. Thanks to the HTTP/HTTPS protocol is widely used in the network, and the number of HTTP/HTTPS servers is quite large, the botnet based on HTTP/HTTPS protocol has better imperceptibility and stability, not paralyzed by a single server failure. For this type of Botnet detection has many works already.

According to the analysis of the behavior of HTTP/HTTPS Botnet, Eslahi[7] adopted the method of access rate and periodic access analysis to detect botnet. Cai[8] analyzed the characteristics of HTTP Botnet, such as the response time interval, the size of the packet, the number of packets, the content similarity of packets, and so on, and puts forward the relevant detection algorithm. But the Lin[9] told us that the bots access frequency, time interval, packet size, etc, these properties can be adjusted. Once adjusted, the original detection algorithm based on the analysis of these characteristic behavior will produce a large false positive rate and false negative rate. Therefore, the detection and prevention of botnet based on HTTP/HTTPS protocol is a long way to go.

In this paper, we propose a method for detection and prevention of botnet based on HTTP/HTTPS protocol, the detection algorithm based on content association recommendation is designed. The method only detects the communication data between the bots and the C&C server. We could detect the existence of the botnet before the attack start that greatly reduce the burden of network.

According to this algorithm, even if some of the bots communication data is not found, it could detect the hidden danger and trace back to the source. The algorithm has small complexity and can realize online detection.

Analysis of Botnet based on HTTP Protocol

Packet Acquisition. At present, the mainstream botnet uses HTTP/HTTPS protocol to send commands and control messages to the hosts, therefore, we only analysize HTTP/HTTPS protocol packets. When the network data from the switch mirror port is sent to the detection end, the capature tools (such as WireShark) should be set the appropriate rules, then capture the packets. In this way, we could filter the botnet independent traffic, reduce the burden of detection end, but also we could quicken to get detection results, realize online detection.

Packet Detection. When the attack occurs, the malicious traffic in the network (such as DDoS) increases rapidly, if this time we analysize and test the data, it will further aggravate the burden of network, and not conducive to dredge network congestion. Therefore, the packets' detection should not be placed on the attack duration. HTTP/HTTPS botnet with CaaS (Crime-as-a-Service) features, attack traffic is very subtle, difficult to detect. In order to nip in the bud, it is the best way to detect the existence of botnets before the attack start.

Before the C&C server sends attack commands to zombie hosts, these hosts are in the sleep state, and periodically polling the server for new control commands. In the same botnet, these polling messages are very similar. If we can detect and analyze these polling messages, we could protect the security of the network before the C&C servers send out the attack commands.

Packet Analysis. The attacker uses fast-flux technology makes it difficult to find the C&C server[10]. In order to maintain the availability and concealment of botnets, fast-flux's IP address would constant change. A zombie host usually keeps in touch with multiple C&C servers, it may not be able to find all C&C servers from a single or a few packets, satisfactory results could be obtained by using the content association recommendation algorithm, and has a small computational complexity.

Online Detection Model and Algorithm based on Recommendation Algorithm

Figure 1 is a botnet detection model based on HTTP/HTTPS protocol proposed in this paper.

Mutual communication data between network to be detected and Internet is mapped to detection server through the mirror port of the core switch. The host in the network will be connected to the C&C server through the POST packet after being infected by the bots, take configuration information and perform related operations, the server returns the information to the host through a simple encryption code, the purpose is to prevent safty software killing. At this point the POST packet has a fixed format and content, and transmit with plaintext. Therefore, this paper detects the existence of botnets through these packets. Detection server running WireShark grabs POST packets on HTTP/HTTPS protocol. After the Log format, the packet to be detected contains the basic information: source ip address, destination ip address, source port, destionation port and protocol, etc. Due to the diversity of network services, it is very difficult to send high similarity data packets to the same host repeatedly. So if you detect the existence of such packets in the network, the current host may have been infected by the bots.

There are tens of thousands of zombie hosts in Internet. In this paper, the classical edit distance algorithm proposed by Levenshtein[11] is used to match the string similarity.

Classical Lenvenshtein Algorithm. Suppose there are 2 strings str1, str2, each contains m and n elements, mapping m*n elements to a matrix$\{d_{ij}\}(0 \le i \le m, 0 \le j \le n)$, the way to fill the matrix is as follows:

$$d_{ij} = \begin{cases} i (j = 0) \\ j (i = 0) \\ \min\left(d_{i-1j-1}, d_{i-1j}, d_{ij-1}\right) + a_{ij} (i, j > 0) \end{cases}, \quad a_{ij} = \begin{cases} 0 & str1_i = str2_j \\ 1 & str1_i \neq str2_j \end{cases} (i = 1,2 \ldots m; j = 1,2 \ldots n)$$

The distance between the string str1 and str2 can be defined as: $LD(str1, str2) = d_{mn}$.

It can be argued that, when LD(str1, str2)=0, the two strings are completely matched, when LD(str1, str2)=1, the two strings do not match at all. The Lenvenshtein distance, also known as edit distance.

In order to better detect the network data flow, this paper proposes that if the edit distance between two strings is less than 0.3, we think that the two strings are highly similar. Taking BlackEnergy as an example, by default, the zombie hosts and the C&C servers contact every 30s and packets for each communication is not more than 100K, the purpose is not to cause the attention of safety equipment. Therefore, the classic editing algorithm could do online detection.

The hot information in the network may cause multiple hosts to repeated access, but the same host will not repeated access the same content in a short time. Therefore suspected botnet data should have the following characteristics: 1) source IP address exactly the same, 2) packet data highly similar. After finding out the botnet data in the network, it is important to prevent the server which has being infected by the botnet from harming the network further. But there are more than one server, fast-flux technology makes the situation more complicated, how to accurately locate the server is the top priority.

Recommendation Algorithm based on Content Association. Recommendation algorithms are often used to detect activity related behavior. In this paper, we propose a recommendation algorithm based on content association, which can accurately locate the hosts and servers infected by the botnet.

N_A： All HTTP/HTTPS POST packets collection sended by host A.

S_A： According to the content and the source IP address of the packets in the N_A, the set of packets that highly similar

S_{AE}： According to the packet in S_A, the set of packets with the destination IP address as host E.

L_{AE}： The preference degree of host A to host E in a suspected botnet. $L_{AE} = {}^{S_{AE}}/_{S_A}$

Specific detection algorithm is as follows:

Step 1: Check whether host A sends suspected botnet data packets. If host A does not send any suspected data, it turns to other hosts in the network. If there is a possibility that host A sends data to the Botnet, then the relevant attributes are counted,eg, N_A 、S_A 、S_{AE} 、L_{AE}, etc.

Step 2: To host E, its importance in a suspected botnet would be calculated. Specific calculation methods are as follows: $F_E = \frac{\sum L_{iE}}{\sum i}$,, i is one host in the network, and $L_{AE} \neq 0$； $R_E = {}^{num(S_E)}/_{num(N_E)}$,

that is to calculate the proportion of suspicious packets in all packets.

Step 3: Any suspected zombie host i in the network, has the F_i attribute. Through the above steps, it is easy to see that the value of F_i attribute is between 0 and 1. The larger the value of F_i attribute, the more suspected botnet data was received by host i in the current network, it should focus on whether host i is used as a server by the botnet. Any suspected zombie host i in the network, has the R_i attribute. Through the above steps, it is easy to see that the value of R_i attribute is between 0 and 1. The larger the value of R_i attribute, the more suspected botnet data was sented by host i in the current network, it should focus on whether host i is used as a zombie host by the botnet.

The algorithm we proposed is related to the network size and the communication among zombie hosts in the network. Suppose the number of zombie hosts in the network is *N*, while the complexity of the algorithm is $O(N^2)$.

Experiments and results

The experimental environment is shown in Figure 2: 2 C&C servers, 4 hosts have been infected with the bots(PC1-4), 1 normal host(PC5), 1 detection server. Bots(BlackEnergy) have been deployed, the normal data set using the DARPA data set of the Lincoln Laboratory, DARPA data set is playing while BlackEnergy is enabled, in order to mix the zombie data and normal data. The whole experiment lasted about 3 hours.

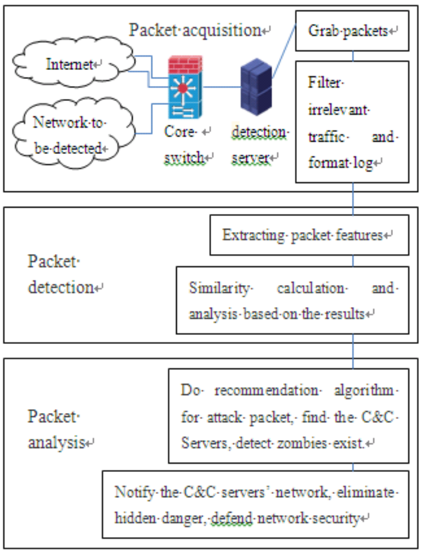
Packet acquisition
Internet
Network to be detected
Core switch
detection server
Grab packets
Filter irrelevant traffic and format log
Packet detection
Extracting packet features
Similarity calculation and analysis based on the results
Packet analysis
Do recommendation algorithm for attack packet, find the C&C Servers, detect zombies exist
Notify the C&C servers' network, eliminate hidden danger, defend network security

Figure 1 Botnet online detection model based on HTTP/HTTPS protocol

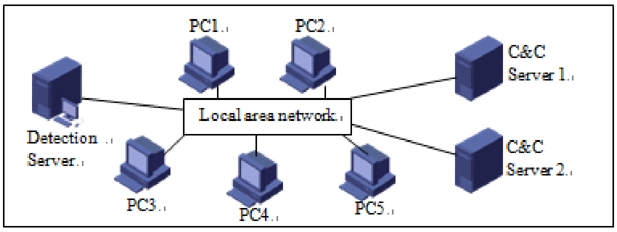

Figure 2 Experimental topology

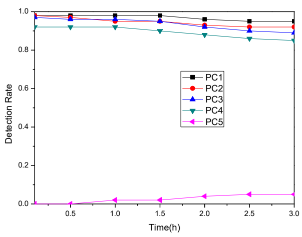

Figure 3 Detection rate of suspicious packets in different hosts

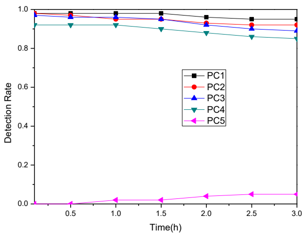

Figure 4 F_i change with time

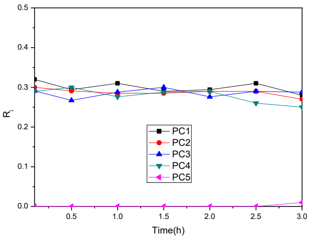

Figure 5 R_i change with time

Figure 3 shows that this method has high detection rate. This method can detect potential malicious traffic more than 90%, while bot packets are very few and completely hidden in legitimate traffic. The administrator could prepare in advance to avoid the attack occurred by surprise. For normal data, the false alarm rate is less than 6%. Figure 4 is the result to detect the F_i attribute, Figure 5 is the result to detect the R_i attribute, we can see that with the changes of network topology and changes of normal network data flow, as times goes on, there are some fluctuations in the property value itself, but malicious traffic can still be reflected in the chart. As long as the zombies exists in the network, this method can detect them and continue to show them in the detection results. Figure 4 shows that, the F_i property of two botnet related C&C servers appeared obvious fluctuation, because in the HTTP protocol Botnet, the hosts only communicate with server, there is no false positives on other hosts. We can see from Figure 5, infected bots hosts(PC1-4) and no infected bots (PC5), R_i attributes are not same. The infected bots hosts have obvious fluctuation in the whole test period, and no infected bots host are relatively stable, maintained at a low level. Therefore, if a host or server in the F_i and R_i attributes detection have a sustained abnormal performance (test value >0), it should be caused the administrators' attention.

Experiments show that this method has a good detection effect. This method is able to independently accomplish the detection, does not depend on early deployment of hosts. According to the detected content, not just look at the packet size, the packet number, etc, this method can avoid misinformation of packets that have similar properties. We use the content association recommendation algorithm, even more than one server control a host, all can be detected successfully. This method can prevent the attack from the server end, and provide the effective protection to the network.

Conclusions

This paper mainly discusses the characteristics of Botnet based on HTTP protocol, and designs a detection algorithm which is based on the content association recommendation. This method can through the transmission characteristics between C&C server and bots, detect botnet exist before the attack start. Experiments show that the proposed method can have good detection result, and the longer the more effective. The future work is to further research on general botnet detection algorithm, and reduce the false alarm rate and false negative rate.

References

[1] JIANG Jian, ZHUGE Jian-Wei, DUAN Hai-Xin, WU Jian-Ping. Research on Botnet Mechanisms and Defenses [J]，Journal of Software，2012，23(1)：82-96.

[2] HUAWEI Technologies Co. Ltd. 2015 Botnet and DDoS attacks special reports [OB/EL]，[2016.4][2016.11]http://e.huawei.com/cn/marketing-material/cn/products/enterprise_network/security/anti-ddos/20160406102046

[3] G. Fedynyshyn, M. C. Chuah and G. Tan, "Detection and Classificationof Different Botnet C&C Channels," in Proceedings of the 8thInternational Conference on Autonomic and Trusted Computing, 2011,pp. 228-242

[4] Goebel J and Holz T. Rishi: identify bot contaminated hostsby irc nickname evaluation[C]. Proceedings of USENIXHotBots'07, Berkeley, CA, USA, 2007: 163-174.

[5] Nagaraja S, Mittal P, Hong C, et al.. BotGrep: finding P2Pbots with structured graph analysis[C]. Proceedings of the19th USENIX Conference on Security, Washington, USA,2010, 7: 1-16

[6] M. Bailey, E. Cooke, F. Jahanian, X. Yunjing and M. Karir, "A Survey ofBotnet Technology and Defenses," in Proceedings of the CybersecurityApplications & Technology Conference for Homeland Security(CATCH), 2009, pp. 299-304

[7] Meisam Eslahi, H. Hashim, N.M. Tahir. An efficient false alarm reduction approach in http-based botnet detection[C], Computers & Informatics (ISCI), 2013 IEEE Symposium on, pp: 201-205.

[8] Tao Cai, Futai Zou. Detecting HTTP Botnet with Clustering Network Traffic[J], Wireless Communications, Networking and Mobile Computing (WiCOM), 2012 8th International Conference on, 2013, P:1-7.

[9] Lin Zhiwei. Network traffic analysis and detection for HTTP-Based botnet, Taiwan: Ming Chuan University, 2014, http://handle.ncl.edu.tw/11296/ndltd/74848926432490581243

[10] Riden J. Know your Enemy: Fast-Flux service networks. The Honeynet Project, 2008.

[11] Levenshtein V. Binary Codes Capable of Correcting Deletions,Insertions, and Reversals[J]. Soviet Physics Doklady, 1966,10(8): 707-710.