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

P2P Botnet Detection Method Based on Data Flow

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: P2P, botnet, data stream, detection and prevention

Abstract. P2P data transmission is the mainstream of network data transmission. P2P botnet malicious data is hidden in normal transmit data, not only difficult to detect but also could cause great harm. This paper presents a method of P2P botnet detection based on data flow, first of all, extract the P2P data stream accurately, and then detect the P2P bots data stream, the small computational complexity does not affect the normal operation of the network. According to the characteristics of the P2P bots data stream , the detection method can detect the existence of the bots before the attack start and filter out illegal data. The experiment results show that this method has good detection efficiency and further maintains the security of the network.

Introduction

The ability to detect botnets is an important part of current network security system. The botnet includes a collection of computers that are infected by bots, botmasters through command and control(C&C) channel remotely manipulate these computers, provide infrastructure for viruses, denial of service attacks, identity theft, click fraud and other malicious acts. C&C Channel is an necessary part between the botmaster and the zombie hosts' communication. There are a number of different ways to build a C&C channel in a botnet. In the centralized architecture, all the C&C channels controlled by the bots are owned by the botmaster, so the single point failure becomes a problem that can not be ignored. To overcome this problem, botmasters have recently tended to build a more resilient network architecture-- Peer-to-Peer(P2P) botnet.

Related Work

Traditional botnets often adopt a centralized structure, as shown in Figure 1. Once the C&C server is found, the entire botnet will no longer be controlled. P2P botnet could form a relatively decentralized network environment, all bots are connected to each other and communicate, it is no longer exists a single point of failure since no central server, as shown in Figure 2.

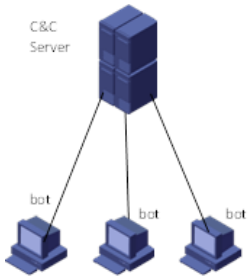

Figure 1 centralized botnet

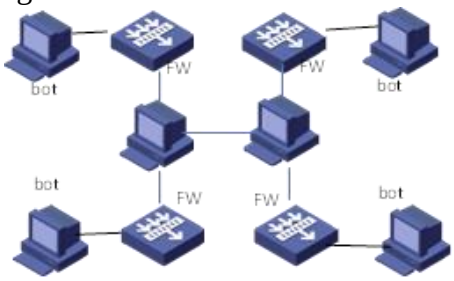

Figure 2 distributed botnet

In recent years, the P2P transmission mode developed rapidly, the current P2P transmission data in different countries accounted for about 50% to 70%[1], therefore, P2P traffic has become the mainstream of the network. To filter the malicious P2P data stream in the network and do not affect the normal data, prevent the improper use of P2P traffic is the most important factor in the current network security. The premise to distinguish between normal and malicious P2P data flow is to understand the characteristics of P2P data.

Since the development of P2P, there is a lot of technology to avoid the inspection of security software, such as the use of dynamic port number, disguised as other protocols (eg. HTTP) traffic,

data encryption, etc. This makes traditional detection methods[2-4] which rely on single character such as the port, load, simple behavior analysis and other can not identify all P2P data.

Traditional centralized botnet processing method is no longer suitable for P2P Botnet, P2P botnet already has some works by now. Zhigang[5] proposed a botnet detection tool, according to the similarities and differences between the user behavior and the social behavior of the bot nodes and the common nodes, high accuracy is obtained in his experiment. However, when the network traffic increases, the accuracy would decline. According to the user's behavior, Sherif[6] extracts the C&C phase feature of the botnet, and proposed a method which is tested by machine learning, the experiment results show that the proposed method can get the ideal results before the attack start. However, it is difficult to meet the requirements of online detection because of complex calculation. According to the relationship between the network behavior and the host behavior, Yin[7] proposed a detection method in six stages. The advantage is not only can detect known and unknown malicious behavior, but also can eliminate the threat, the disadvantage is that once the P2P bots using encryption transmission mode, the detection is invalid.

P2P Bots Data Features

Local area network(LAN) usually uses NAT, firewall or proxy server technology, Internet data is difficult to connect directly to LAN, but does not affect the LAN's data access directly to the Internet. Therefore, it is very difficult to communicate directly between the hosts in different LAN, P2P network still needs the server to some extent. The host can accept the data connection is known as server, the server is usually located in the Internet rather than the LAN, the host can not accept the data connection can initiate the connection to the server, convenient for receiving commands. Although some bot nodes can be used as servers technically, they are not fixed, the botmaster has a number of bot nodes that temporarily act as servers and allow them to switch to each other. Once a server is exposed, other servers will immediately replace it.

Each bots node maintains a subset of IP address lists that contain several other bots. Botmaster holds the sets of all the IP address lists, once a bot node exposured, IP address list would be updated in a timely manner. There are a lot of smaller botnets in the botnet, the connections between them are different, in order for the command to be communicated throughout the entire botnet, the host must be connected to multiple nodes simultaneously and transmit any commands issued by other nodes.

In order to ensure the control of entire botnet, botmaster issues command usually use asymmetric encryption and digital signature technology. It makes the network security device not only can not identify the data from the botmaster, but also can not forge. Therefore, it is difficult to detect botnet data from the characteristics of network packets, still further, can not prevent the botnet attacks.

From the above, we can conclude that the P2P botnet has the following characteristics:

1)P2P botnet uses dynamic port number, source port of the packet sented by a P2P bot is unchanged at the same time, but the destination port of different hosts is dynamic;

2)There are a lot of server hosts in P2P botnet, they can switch to each other, so there is no single point failure problem;

3)The hosts in P2P botnet must be connected to multiple nodes at the same time;

4)P2P botnet data adopts encryption and signature technology, we can not extract the packet characteristics.

Coarse-grained Detection Technology

According to the above characteristics, this paper designs a method of P2P botnet prevention based on data flow, the first step is to extract the P2P data stream, the P2P data stream in the network should be separated from other data. As shown in Figure 3, the following steps are as follows:

1)Extracting a network packet, record its source IP address, destionation IP address, source port number, destionation port number, and packet delivery time, packet size. For packet i, denoted as six tuple P_i={IP_S，IP_D，POR_S，POR_D，T，Payload};

2)In the unit time t(such as 1 hours), calculate the number of couple{IP_S, POR_S} connect to the different couple{ IP_{Dj}, POR_{Dj}}, j=1,2,…n, denoted as n;

Taking packet i as an example:

If($P_i.IP_S$， $P_i.POR_S$ does not exist in the record)

Then {$P_i.n=1$; in current data stream ($P_i.IP_S$， $P_i.POR_S$) add record ($P_i.IP_D$， $P_i.POR_D$);}

Else if(in current data stream ($P_i.IP_S$， $P_i.POR_S$)exist record ($P_i.IP_D$， $P_i.POR_D$))

Then receive next packet P_{i+1};

Else{$P_i.n=P_i.n+1$; in current data stream ($P_i.IP_S$， $P_i.POR_S$) add record ($P_i.IP_D$， $P_i.POR_D$)}

3)If the value of n is greater than or equal to 2, the current source IP address, port number may be used for P2P botnet data transmission.

Table 1 Schematic diagram of P2P data stream

Source IP source Port	Destination IP，Destination Port	Destination IP，Destination Port	Destination IP，Destination Port	Destination IP，Destination Port	Destination IP，Destination Port
SourceA IP sourceA Port	Destination A_1IP DestinationA_1Port	Destination A_2IP DestinationA_2Port	Destination A_3IP DestinationA_3Port	Destination A_4IP DestinationA_4 Port	Destination A_5IP DestinationA_5 Port
SourceB IP sourceB Port	Destination B_1IP DestinationB_1Port	Destination B_2IP DestinationB_2Port	Destination B_2IP DestinationB_3Port		

In Table 1, the host A couple(SourceA IP, sourceA Port) and the host B couple(SourceB IP, sourceB Port) are likely to be used to transmit data from the botnet. In this paper, we define in the unit time t, the packets with the same source address, the same source port, and the same protocol belong to the same data flow.

Use the above method we can ruled out a number of non P2P packets (such as web game data packets, etc.), not only further reduce the scope of next detection, but also improve the detection rate and reduce false alarm rate.

It is now known that the data to be detected are P2P protocols. How to separate the normal P2P data and P2P botnet data is the content of the next analysis.

Fine Grain Detection Technology

1)There are a lot of P2P programs on a host. The user may open a P2P program when it is in use, download a limited number of files and then turn off (e.g. Emule). P2P bots for the botmaster is the opposite, when the attack occurs, botmaster need a sufficient number of bot nodes online, therefore, the P2P bots would be active during the entire boot time.

2)In order to avoid the attention of security tools, the number of packets sent by the bots during the whole activity is small, and the packet size is small.

3)Once a host is infected with bots, the host with which it communicates has a higher chance of infection.

Based on the above inference, this paper designs a method to accurately detect P2P botnet.

In the coarse grain detection phase, the time of deliverying each packet has been recorded, extract the earliest time for all packets $T_{min}$and the latest time T_{max}, then alculate system working time $T=T_{max}-T_{min}$.

The system working time T is divided into a number of unit time t (such as 1 hours). At present, the number of packets per P2P flow in the t time is known, and the size of each packet is known, so we can get the median packet size for each P2P data flow. Here the median rather than the mean value is determined by the characteristics of the P2P bot, the packets sended by P2P bot is generally small, in order to prevent the bot at some stage deliberately sent a larger size of the packet to confuse security

tools, we use the median. If the data flow is out of range in multiple units time of the detection, the host is set up as a bot node, and all packets belong to the data flow should be deleted.

Taking into account the infect characteristics of bot and to further prevent the hazards of bots, other hosts that communicate with this host must also be set up as a bot node and filter the data stream.

Experiment and Conclusion

The experiment uses three P2P software: Skype, Emule, Bittorrent and two P2P bots: Storm,Waledac, these data as the detect content of our method. The experiment host with the same configuration, the whole experiment lasted for 3 hours, the specific topology shown in Figure 4:

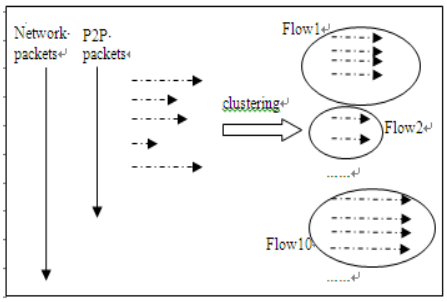

Figure 3 Extracting P2P data stream

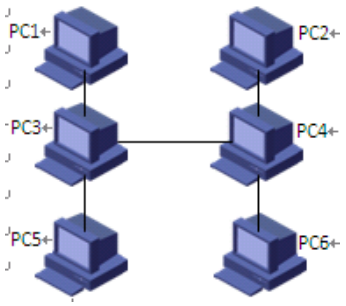

Figure 4 Experimental topology

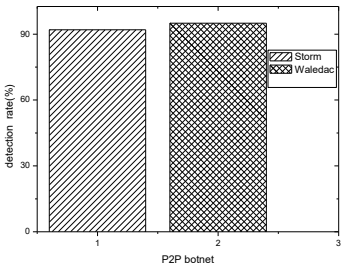

Figure 5 Two kinds of bot packet detection rate

In PC1, PC3, PC5, the 3 kinds of P2P software were installed in the 3 hosts and used normally, then we use WireShark to capture packets, after unit time t, filtering in the corresponding protocol data and analyzing the size of the P2P packet, we can draw the conclusion as shown in Table 2.

In PC2, PC4, PC6, the 2 kinds of P2P bots were installed in the 3 hosts and used normally, then we use WireShark to capture packets, after unit time t, filtering in the corresponding protocol data and analyzing the size of the P2P packet, we can draw the conclusion as shown in Table 3.

Table 2 Common P2P software packet size

P2P software	Packet size range(byte)	Packet size median(byte)
Skype	18-170	130
Emule	6-1452	1440
Bittorrent	5-1452	1440

Table 3 Common P2P bot packet size

P2P bot	Packet size range(byte)	Packet size median(byte)
Storm	0-160	79
Waledac	0-160	78

Based on the above data, this paper sets the threshold of the packet median size in the unit time t is 100 bytes. In the entire process of the experiment, once the data stream has been detected to transmit packets which median size is less than 100 bytes always, it is considered that the host that sends the data flow has been infected by P2P bots.

Firstly, we capture packets with WireShark, and then aggregate all P2P packets to data streams. At this stage, it can be seen that almost all data streams include small packets, this is because in TCP transport protocol, the initial control packets, reply packets, reset packets is indispensable, but common software is used more for data download and delivery, so the large size packets are the mainstream, would not cause false positives. Skype uses small packets, but compared to the bot, users use Skype is in a relatively short time, only a small portion of the runtime, not throughout the whole runtime, and not cause false positives.

Then we need to detect the bot packets. The whole experiment lasts for 3 hours, including the 3 units time, that is, the data flow must continue meet the condition of the packet median size is less than the threshold value duration of 3 units, then can be determined as the host has been infected with the bots. Because the bot is contagious, so the data flow corresponding to a number of destination hosts should also be identified to bot nodes. The data packets corresponding to this data stream should be filtered out.

As can be seen from Figure 5, the method for the detection of two kinds of P2P Zombie: Storm, Waledac has reached the rate of 92% and 97%, the detection rate of the bots is 100%. In general, the detection rate of the proposed method has reached 95%, and in the whole detection process we can not only filter out the zombie data flow but also not affect the use of normal P2P software.

Summary and Prospect

According to the characteristics of P2P Botnet, this paper puts forward a kind of prevention method, which only detects the P2P data flow, detect and filter out the zombie data in the case of less computational complexity, and has good detection rate. To prevent further impact to the network, the detection can be completed before the attack start. We are aim to the data flow, both the known and the unknown zombies can be detected, due to the characteristic of P2P bot is easy to infect, both sides of the suspicious data flow should be kept watch, this is further improve the efficiency and reduce the omission.

References

[1] “Ipoque internet study 2008/2009,” http://www.ipoque.com/en/resources/internet-studies, accessed on 4 January 2014.

[2] H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, “Internet Traffic Classification Demystified: Myths, Caveats, and The Best Practices,” Proceedings of the ACM CoNEXT Conference, Dec. 2008.

[3] W. Yu, X. Yang, and S. Z. Yu, “Automatic Application Signature Construction from Unknown Traffic,” 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), pp. 1115-1120, Apr. 2010..

[4] M. M. Hassan and M. N. Marsono, “A Three-Class Heuristics Technique: Generating Training Corpus for Peer-to-Peer Traffic Classification,” IEEE 4th International Conference on Internet Multimedia Services Architecture and Application, Dec. 2011.

[5] Zhigang, J., W. Ying and B. Wei, 2012. P2P Botnets detection based on user behavior sociality and traffic entropy function. Proceedings of the 2nd International Conference on Consumer Electronics, Communications and Networks, Apr. 21-23, IEEE Xplore Press, Yichang, pp: 1953-1955.

[6] Sherif, S., I. Traore, A. Ghorbani, B. Sayed and D. Zhao et al., 2011. Detecting P2P botnets through network behavior analysis and machine learning. Proceedings of the 9th Annual International Conference on Privacy, Security and Trust, Jul. 19-21, IEEE Xplore Press, Montreal, QC, pp: 174-180.

[7] Yin, C. and A.A. Ghorbani, 2011. P2P botnet detection based on association between common network behaviors and host behaviors. Proceedings of the International Conference on Multimedia Technology, Jul. 26-28, IEEE Xplore Press, Hangzhou, pp: 5010-5012.