Botnet Host Detection Based on Heartbeat Association

Ding Wei, Hua Zidong, Li Panhui, Gong Qiushi and Cheng Yuxi
School of Cyber Science and Engineering, Southeast University,
Jiangning District, Nanjing, Jiangsu Province, P.R.China
wding@njnet.edu.cn, zdhua@njnet.edu.cn, phli@njnet.edu.cn, qsgong@@njnet.edu.cn,
yxcheng@njnet.edu.cn

ABSTRACT
As a common means of communication, heartbeat is often used by the network applications. Hosts with the same heartbeat tend to have the same applications and thus share the homogenous vulnerabilities. Based on the detected heartbeat, the paper designs the heartbeat network, the heartbeat associated graph and an attribute propagation algorithm based on the heartbeat associated graph. The paper takes the distributed denial of service (DDoS) malicious host information provided by the intrusion detection system (IDS) deployed on the boundary of China education and research network (CERNET) Nanjing master node network as attribution, and constructs the associated graph based on the user datagram protocol (UDP) heartbeat detection result at the same location. The attribute propagation algorithm was tested for 17 days. And The result shows that the method can effectively detect DDoS malicious hosts that are not located by IDS.

CCS Concepts
• **Networks → Network monitoring**.

Keywords
Heartbeat network; heartbeat association; botnet host; attribute propagation algorithm;

1. INTRODUCTION
A botnet [1] is a collection of hosts that are infected by bot programs and controlled by a malicious attacker. A botnet is used by an attacker as a platform for malicious activities like DDoS attacks, spam, and identity theft. Botnets have the properties of complex network environment, large number of members and variable membership status, which calls for a reliable communication method to maintain the botnet, such as heartbeat.

Heartbeat [2] is a common method for maintaining high availability and high reliability of network devices and network applications. The basic principle of heartbeat is that the network devices or network applications follow the heartbeat protocol and cyclically sends a specific heartbeat packet for the receiver to determine the activity of the sender. Heartbeat is simple and cost-effective, hence it is widely used in the Internet, such as instant messaging software, Internet-of-things (IoT) devices and distributed system programs for system diagnostics, network fault detection, etc. Botnets also use heartbeat to take control of the activity of the botnet hosts.

ICCSP 2020, January 10–12, 2020, Nanjing, China.
ACM ISBN 978-1-4503-7744-7/20/01…$15.00.
DOI: https://doi.org/10.1145/3377644.3377653

The core of this paper is to build the heartbeat network based on the detected heartbeat, construct the heartbeat associated graph, and try to locate those botnet hosts that have been infected but not located yet through the attribute propagation by using the heartbeat association graph and the infected host information.

The rest of this paper is organized as follows: Section 2 summarizes some previous research which is relevant to this paper. Section 3 introduces the heartbeat network and heartbeat associated graph. Section 4 presents the botnet host detection method based on heartbeat association. Section 5 shows the tests and result. Finally, section 6 concludes this paper.

2. RELATED WORK
Early botnet detection was generally done based on feature analysis. [3] used a honeypot to capture bots, and then extract features for rule matching. And [4] relied on IRC trackers to obtain botnet for research. However, these methods cannot be applied to unknown botnets. Later, the academia began to detect botnets based on the abnormal behavior. [5] identified botnet hosts by monitoring abnormal activities at the host level (such as binary file downloading, sending spam, etc.) and abnormal behavior of network traffic. With the popularity of machine learning, there are also applications in the detection of botnets. [6] introduced a method to detect botnets by using deep learning based on HTTP botnet traffic features. Although detection based on abnormality can achieve high accuracy, this detection is available after the botnet's abnormal behavior has occurred.

The research on heartbeat detection has gradually emerged in recent years, mainly about the detection of Trojans: [7] designed a Trojan rapid detection system by analyzing the heartbeat behavior of Trojan communication. [8] proposed a method based on wavelet transform for detecting heartbeat behavior of Trojans. [9] combined fuzzy theory to extract network packets into sequences and separate heartbeat packets. [10] and [11] conducted similar research. Recent research generally concentrates on single heartbeat, rather than heartbeat association.

A heartbeat detection algorithm based on UDP application layer [12] has been deployed on the boundary of CERNET Nanjing master node network to provide real-time detected heartbeat. And there is also an IDS system in the same location of the network, which can provide the DDoS malicious host information synchronously. These are the basis of the research work of this paper.

3. HEARTBEAT NETWORK AND HEARTBEAT ASSOCIATED GRAPH
The research work of this paper is based on algorithm [12] which presents a heartbeat detection algorithm applied on the original

Fund project: Internet basic behavior measurement and analysis: Internet basic behavior indicator system and measurement method (2018YFB1800202).

heartbeat flow from the original traffic provided by the network probe. The relevant definitions are as follows:

Definition 1: Original heartbeat flow <srcIP, dstIP, dstPort, startTime, endTime> represents the heartbeat from the source address to the port of the destination address between the start and end time.

Definition 2: Combined heartbeat flow <srcIP, dstIP, dstPort> is the set of original heartbeat flow from the source address to the port of destination address during the observation period of length T. In other words, we separate the entire observation into equal T-length period and combine the original heartbeat flow within such time slots.

Definition 3: Heartbeat network <srcIP_set, dstIP, dstPort> is the set of combined heartbeat flows by further combining flows according to their destination addresses and ports.

Definition 4: Direct association between heartbeat networks:

For any two heartbeat networks *heart_net*$_A$ and *heart_net*$_B$, if

$$heart\_net_A.srcIP\_set \cap heart\_net_B.srcIP\_set \neq \emptyset$$

then we say that there is a direct association between heartbeat networks *heart_net*$_A$ and *heart_net*$_B$.

The heartbeat network generally relies on some network applications, and hence it is reasonable to speculate that the behaviors of the member hosts in the heartbeat network are related. Also, a heartbeat sender srcIP may exist in multiple heartbeat networks simultaneously. Therefore, the heartbeat network can be associated based on such srcIP.

Definition 5: Association between heartbeat networks exists if heartbeat networks *heart_net*$_A$ and *heart_net*$_B$ have direct association:

$$heart\_net_A.srcIP\_set \cap heart\_net_B.srcIP\_set \neq \emptyset$$

or indirect associations via intermediate heartbeat networks:

$$hear\_net_A.srcIP\_set \cap hear\_net_B.srcIP\_set = \emptyset$$

$$\&hear\_net_A.srcIP\_set \cap hear\_net_1.srcIP\_set \neq \emptyset$$

$$\&hear\_net_i.srcIP\_set \cap hear\_net_{i+1}.srcIP\_set \neq \emptyset$$

$$(\mathrm{i} = 1, 2 \cdots\cdots \mathrm{n} - 1)$$

$$\&hear\_net_n.srcIP\_set \cap hear\_net_B.srcIP\_set \neq \emptyset$$

i.e. heartbeat network *heart_net*$_A$ and *heart_net*$_B$ have no direct association but there exist heartbeat networks that have direct associations with both *heart_net*$_A$ and *heart_net*$_B$.

Definition 6: Heartbeat association network is a set of heartbeat networks where any pairs of heartbeat networks have association.

Definition 7: Heartbeat associated graph is a weighted undirected graph with:

1) Each vertex of the graph represents a srcIP of a heartbeat network;

2) If two vertices belong to the same heartbeat network, there is an edge between them;

3) The weight between the two vertices is the number of the heartbeat networks where they are both in.

The heartbeat associated graph is a visualization of a heartbeat association network. Since the graph needs at least two vertices for constituting edges and weights, heartbeat networks with only one srcIP would not affect the number and weight of each edge hence can be ignored in the graph.

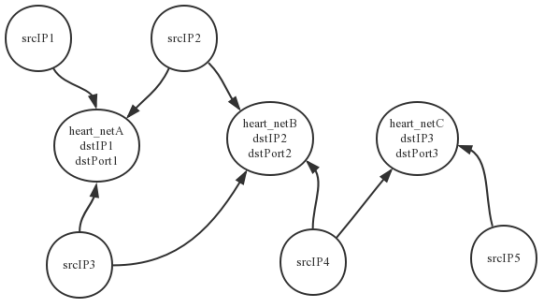

Figure 1. Example of Heartbeat network

There are 3 heartbeat networks in Figure 1. According to Definition 7, there are five vertices srcIP1, srcIP2, srcIP3, srcIP4 and srcIP5 in the heartbeat association graph constructed by Figure 1. Among them, srcIP1, srcIP2 and srcIP3 are connected to each other, srcIP2, srcIP3 and srcIP4 are connected to each other, and srcIP4 and srcIP5 are connected to each other. Since the number of edges between srcIP2 and srcIP3 is 2, the edge between srcIP2 and srcIP3 has a weight of 2, and the other edges have a weight of 1. The heartbeat association graph finally constructed by Figure 1 is shown in Figure 2.

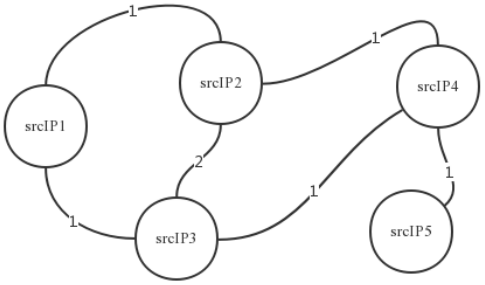

Figure 2. Example of Heartbeat associated graph

4. BOTNET HOST DETECTION BASED ON HEARTBEAT ASSOCIATION

The core of this paper is to obtain the relationship between the hosts in the association graph and the known botnet hosts based on the heartbeat associated graph and the known botnet host information, and to find those hosts that have been infected but not located yet. There are three main tasks that need to be accomplished further:

1) Quantify the malicious behavior characteristics of the known botnet hosts to obtain the initial malicious attribute; 2) Construct the attribute propagation algorithm based on the heartbeat associated graph and initial attribute; 3) Obtain the termination condition of a stable value for each host and result analysis.

For 1), the starting point of the attribute propagation algorithm is that the initial attribute value should come from the real network behavior of the hosts. The initial attribute value needs to be provided according to known botnet host behavior, such as DDoS attack, spam sending, etc., by other systems (such as IDS systems). In general, if the host in the heartbeat associated graph reveals botnet host behavior during the detection period T, its initial malicious value is set to 1, otherwise it is set to 0.

For 2), attribute propagation refers to the transfer of attribute values in the associated graph according to an attribute propagation algorithm, and finally results in a stable attribute value for each vertex. The core of the attribute propagation algorithm is to give a new attribute value to the target vertex based on the attribute values of other vertices connected to the

target vertex. PageRank [13] is an algorithm employed by search engines to page ranks based on hyperlinks between web pages. The core of the PageRank algorithm is that "PageRank of a page is calculated by PageRank of other pages associated with it." This is like the core of the attribute propagation algorithm "Based on the attribute values of other hosts associated with the target host to give the target host a new attribute value." Therefore, an improved PageRank algorithm is proposed as a malicious attribute propagation algorithm.

According to the PageRank algorithm, the malicious value of host A after round R of propagation is:

$$Evil_R(A) = \sum_{i=1}^{n} \frac{Evil_{R-1}(T_i)}{C(T_i)} \quad (1)$$

Evil (A) represents the malicious value of host A, T_i denotes the host associated with A (i = 1, 2 · · · n), and C (T_i) stands for the outbound of host T_i, which means the sum of the associated weights between T_i and other hosts.

In Equation 1, because the host's own malicious value is not included, some hosts with high initial malicious value may, however, have very low malicious value after t multiple iterations of the attribute propagation algorithm. The initial malicious value is based on known botnet host behavior, and they should be the most important source of propagation. To this end, based on Equation 1, we add an attribute correction to the host in each round of propagation, which is the product of the correction coefficient d and the initial malicious value of the host.

$$Evil_R(A) = d * initEvil(A) + (1 - d) * \sum_{i=1}^{n} \frac{Evil_{R-1}(T_i)}{C(T_i)} \quad (2)$$

InitEvil (A) is the initial malicious value of host A.

For 3), there are two aspects. One is to solve the problem of finiteness. It is necessary to add a suspension condition to the improved PageRank algorithm. In this paper, the difference between the multi-dimensional variables after the iterations of the round R and round R+1 is used to measure the degree of change by the Euclidean distance method and recorded as D (R, R+1).

$$D(R, R+1) = \sqrt{(Evil_R(T_i) - Evil_{R+1}(T_i))^2 + \cdots + (Evil_R(T_n) - Evil_{R+1}(T_n))^2} \quad (3)$$

$Evil_R(T_i)$ is malicious value of host T_i after the round R, and T_i represents the host in the associated network.

When D (R, R+1) > D (R-1, R), the propagation result is bumped, and the attribute propagation algorithm is terminated.

The second aspect of 3) is that after obtaining the malicious value of all the hosts in the heartbeat associated graph based on the attribute propagation algorithm, it is also necessary to perform qualitative analysis according to the malicious value of the hosts, which means to set a malicious-value-judgment threshold G to determine the malicious value of the hosts after propagation. The hosts which exceed the threshold and are not the initial malicious hosts are suspected botnet hosts. In summary, the complete workflow of the botnet host detection method is:

1) Based on the original heartbeat flow during the detection period T, construct the heartbeat associated graph by using the heartbeat network; 2) Based on the DDoS malicious host information during the same period, perform initial malicious evaluation on the hosts in the heartbeat associated graph; 3) Use the attribute propagation algorithm to obtain the malicious value of all hosts in the heartbeat associated graph; 4) According to the malicious-value-judgment threshold and the DDoS malicious host information, obtain the suspected botnet hosts.

The input of the method is the original heartbeat flow during the detection period and the DDoS malicious host information in the same period. The output are the suspected botnet hosts. The parameters of the method are detection period T, malicious-value-judgment threshold G, and correction coefficient d.

5. TESTS AND RESULT

5.1 Test environment

The test is carried out at the boundary of CERNET Nanjing master node network. The network access bandwidth is 40Gbps, and the entire network address exceeds 1 million. The original heartbeat flow and DDoS malicious host information required for this paper can be obtained simultaneously at the network boundary.

5.2 Test plan

This paper selects T = 24 hours as a detection period. Based on the UDP heartbeat detection result during the detection period, the heartbeat network and heartbeat associated graph are constructed. Combined with the DDoS malicious host information provided by the same IDS, the attribute propagation algorithm in Section 4 is used to generate the suspected botnet hosts in the same period without the hosts which have DDoS malicious behavior at the current period or in the past 30 days. The suspected botnet hosts will be observed for comparisons of hosts which have DDoS malicious behavior in the next 14 days.

5.3 Test parameters

5.3.1 Malicious-value-judgment threshold G

We use dynamic G value for the test to avoid subjective bias introduced by fixed threshold. Let G change from low to high, and use the value of each G to classify all the hosts in the heartbeat associated graph according to the relationship between malicious value and G, and whether it is the initial malicious host:

1) The malicious value exceeds the threshold and host belongs to the initial malicious hosts; 2) The malicious value does not exceed the threshold but host belongs to the initial malicious hosts; 3) The malicious value exceeds the threshold but host does not belong to the initial malicious hosts; 4) The malicious value does not exceed the threshold and host does not belong to the initial malicious hosts.

Hosts of 3) are the suspected botnet hosts that this paper expects to find. They have high malicious value, but are not the initial malicious hosts. Hosts of 2) are the classification result that needs to be avoided as much as possible. Therefore, by using the dynamic acquisition method, the threshold that minimizes the total number of 2) and 3) is selected as the daily threshold.

5.3.2 Correction coefficient d

The value of the correction coefficient also uses a dynamic scheme. The scheme used in the experiment is to change from 0.01 to 0.49 at the granularity of 0.01, and test 49 possible values which maximizes $\frac{\text{the DDoS malicious hosts in the next 4 days}}{\text{suspected botnet hosts} - \text{the DDoS malicious hosts in the past 30 days}} \times 100\%$ will be selected as the correction coefficient of that detection period.

5.4 Test result

The final test was from September 18 to September 20 for 3 detection periods according to the above scheme. And the test compared the DDoS malicious host information provided by IDS from September 19 to October 4 for 16 days.

5.4.1 The constructions of the heartbeat associated graph

The constructions of the heartbeat network and the heartbeat associated graph are shown in Table 1 for 3 detection period:

Table 1. Result of heartbeat networks and heartbeat associated graphs for 3 detection periods

Detection periods	Hosts with heartbeat(srcIP)	Heartbeat networks	Hosts in heartbeat network(srcIP_set) more than 3	Heartbeat associated graphs
1	14672	14568	2346	12
2	14980	14791	2046	15
3	14445	14372	1883	17

5.4.2 Malicious-value-judgment threshold G, Correction coefficient d

According to the method of the malicious-value-judgment threshold G and the correction coefficient d, the value for 3 detection periods can be obtained, as shown in Table 2:

Table 2. The value of the malicious-value-judgment threshold G d correction coefficient d

Detection periods	Threshold	Correction coefficient
1	0.250292	0.23
2	0.256676	0.24
3	0.287038	0.27

5.4.3 Test result and analysis

The initial malicious hosts based on the DDoS malicious host information provided by IDS in the same detection period and in the heartbeat associated graph is shown in Table 3:

Table 3. The initial malicious hosts for 3 detection periods

Detection periods	DDoS hosts with heartbeat
1	828
2	786
3	709

Table 4. The list of hosts in 3 detection periods

Detection periods	All_hosts	Previous_hosts	Suspected_hosts
1	635	358	277
2	568	308	260
3	381	234	147

All_hosts refer to the hosts whose malicious value exceeds the threshold G after the attribute propagation algorithm in the heartbeat association graph. Previous_hosts refer to malicious hosts that have been reported by the IDS at the same detection period or in the previous 30 detection periods. Suspected_hosts refer to the remaining hosts after the all_hosts have removed previous_hosts.

Table 5-7 show cases about the suspected hosts detected in the three detection periods which will have malicious behavior in the next 14 days. Maliciousness is the number of the suspected hosts which have malicious behavior in the next 14 days. Ratio is the number of malicious hosts located to the number of suspected hosts.

Table 5. Botnet hosts test result and comparison at the detection period 1

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Maliciousness	8	3	3	2	7	6	4	9	3	2	4	3	2	3
Ratio	2.89	1.08	1.08	0.72	2.53	2.17	1.44	3.25	1.08	0.72	1.44	1.08	0.72	1.08
Total	8	11	14	16	23	29	33	42	45	47	51	54	56	59
Total ratio	2.89	3.97	5.05	5.78	8.30	10.47	11.91	15.16	16.25	16.97	18.41	19.49	20.22	21.30

Table 6. Botnet hosts test result and comparison at the detection period 2

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Maliciousness	8	2	1	5	7	8	6	1	2	3	4	4	1	3
Ratio	3.08	0.77	0.38	1.92	2.69	3.08	2.31	0.38	0.77	1.15	1.54	1.54	0.38	1.15
Total	8	10	11	16	23	31	37	38	40	43	47	51	52	55
Total ratio	3.08	3.84	4.23	6.15	8.16	11.92	14.23	14.62	15.38	16.54	18.08	19.62	20.00	21.15

Table 7. Botnet hosts test result and comparison at the detection period 3

Day	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Maliciousness	1	2	2	3	4	6	0	0	1	1	2	3	2	0
Ratio	0.68	1.36	1.36	2.04	2.72	4.08	0	0	0.68	0.68	1.36	2.04	1.36	0
Total	1	3	5	8	12	18	18	18	19	20	22	25	27	27
Total ratio	0.68	2.04	3.40	5.44	8.16	12.24	12.24	12.24	12.93	13.61	14.97	17.01	18.37	18.37

The result given in the three tables above indicate that about 20% of the suspected hosts that are located will have malicious behavior in the next 14 days.

The above test result shows that the proposed method in this paper can screen hundreds of malicious behavior suspected hosts based on specific DDoS malicious behavior in a large-scale network with a size of millions of addresses. The probability that these suspected hosts will have malicious behavior in the next 14 days is about 20%. It is valuable and feasible to focus on the monitoring of hundreds of suspected hosts.

6. CONCLUSION

This paper designs heartbeat network and heartbeat associated graph by using detected heartbeat and designs a botnet host detection method based on heartbeat associated graphs, attribute propagation algorithm and malicious attribute information of hosts. We use the DDoS malicious host information provided by the IDS system deployed on the boundary of CERNET Nanjing master node network and the UDP heartbeat detection result in the same location as the data source to test the method. The effectiveness of the method is verified in the real environment.

In addition, the method is a general-purpose method that is not only suitable for botnet host detection. In the future, based on the heartbeat associated graphs, the attribute propagation algorithm, and other relevant attribute information of the host, it is also able to perform more accurate detection on the hosts.

7. REFERENCES

[1] Osagie, M. S. U., Enagbonma, O., and Inyang, A. I. 2019. *The historical perspective of botnet tools*.

[2] Gouda, M. G., and Mcguire, T. M. 1998. *Accelerated heartbeat protocols*. International Conference on Distributed Computing Systems.

[3] Eslahi, M., Salleh, R., and Anuar, N. B. 2012. *Bots and botnets: An overview of characteristics, detection and challenges*. Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on. IEEE.

[4] Rajab, M. A., Zarfoss, J., Monrose, F., and Terzis, A. 2006. *A multifaceted approach to understanding the botnet phenomenon*. Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement 2006, Rio de Janeriro, Brazil, October 25-27, 2006. ACM.

[5] Perdisci, R., Perdisci, R., Zhang, J., and Lee, W. 2008. *BotMiner: clustering analysis of network traffic for protocol- and structure-independent botnet detection*. Conference on Security Symposium. USENIX Association.

[6] Zhou, C., and Huang, Z. 2018. *Deep learning detection based on botnet traffic characteristics*. Information Technology.

[7] Lei, M., Sheng-Li, L., Long, L., Jia-Yong, C., and Hai-Tao, S. 2012. *Trojan rapid detection method based on heartbeat behavior analysis*. Computer Engineering.

[8] Bai, H., Pang, J., Dai, C., and Yue, F. 2016. *Trojan heartbeat behavior detection method based on wavelet transform*. Computer Science, 43(4), 150-154.

[9] He, T., and Zhong, H. 2010. *Network heartbeat packets recognition based on DTW and HC-FCM algorithm*. Sixth International Conference on Natural Computation. IEEE.

[10] Zhao, T., Zhou, D., Wang, K., and Zhang, B. 2011. *A rebounding Trojan detection method based on network behavior analysis*. National Computer Security Academic Exchange Conference.

[11] Yi, J., Chen, L., & Sun, J. 2011. *Data stream clustering detection method for network heartbeat packet sequence*. Computer Engineering, 37(24), 61-63.

[12] Ding, W., Jiang, H., and Li, P. 2019. *Measurement and Analysis of Internet Heartbeat Behavior*. Internet Conference of China.

[13] Brin, S., and Page, L. 2012. *Reprint of: the anatomy of a large-scale hypertextual web search engine*. Computer Networks, 56(18), 3825-3833.