A Baseline Modeling Algorithm for Internet Port Scanning Radiation Flows

Qiushi Qong
Research Base of International Cyberspace Governance
Southeast University
Nanjing, Jiangsu
qsgong@njnet.edu.cn

Chenwei Gu
School of Cyber Science and Engineering
Southeast University
Nanjing, Jiangsu
213181214@seu.edu.cn

Abstract—**Aiming at the port scanning traffic obtained from the background radiation flow on a specific Internet location, this paper proposes a baseline model construction method named Scanning_Baseline. The proposed algorithm maps the scanning numbers of each port to a set of coordinates and tracks the centroid to form the baseline model. With the Scanning_baseline model proposed in this paper, we can compare the morphological differences of scanning background traffic in different time windows or different network locations. The algorithm is applied on 420 groups of measured data obtained from 60 campus network boundaries in seven consecutive days according to two schemes respectively. The results show that Scanning_Baseline algorithm can construct stable baselines for different scenarios and find anomalies by comparing the shape and distance between baselines. This method can also be used to model Internet traffic in other contexts.**

Keywords—Internet scanning, Internet measurement， Traffic Modeling， Internet Background Radiation， Internet Security

I. Introduction

Scanning is a ubiquitous phenomenon on the Internet, caused by a variety of purposes. Most scanning behaviors are not malicious, such as specific or scientific research projects[1][2][3], service discovery operations of cyberspace search engine service providers[4][5], etc. However, scanning is also an important link in many malicious behaviors on the Internet, e.g. botnet propagation[6][7], exploitable attack object discovery[8][9][10], etc. When a security vulnerability is discovered[11][12]， the amount of scanning of relevant ports will surge instantly[13]. From the perspective of scanning coverage, network scanning traffic can be divided into global scanning of the Internet (for a specific port or for all ports), and local scanning of a specific network range. Corresponding researches focus on a variety of topics regarding port scanning including attacks, defense, detection methods etc [2][14][15][16][17][18][19][20][21][22]. This work also focuses on this kind of scanning flow.

Early works have proposed guidelines for scanning behavior, including creating web pages for all hosts that perform scanning operations and accept "whitelist" requests of exempting from being scanned. However, due to the insufficient support on legitimacy as well as the increasing cost on maintenance, associated with this advice, most scanners, such as Sodan[5][13] failed to follow the guidelines. With the emergence of efficient open-source scanning tools such as Zmap[23][24], scanning on the Internet has become more convenient and unpredictable[13][25], and the amount of scanning traffic has also been increasing. Except for Shadowserver[3] and some other institutions, the vast majority of scanners do not follow guidelines.

Without effective monitoring and regulating methods, scanning traffic has become a stable background traffic on today's Internet. According to the statistical analysis of the scanning background traffic in a specific network space[25], scanning traffic reveals a relative stability. The prototype of similar ideas has been mentioned in [26]. We agree with this view based on our long-term observations of network scanning[27]. We also find that in general, the pattern of background traffic may be inconsistent at different network locations, but in the same location, the pattern is relatively stationary.

In this work, we put forward a dynamic-weighting-oriented Internet port scanning method for the construction of the baseline model. The proposed method converts the number of scanning of each specific port into a set of two-dimensional coordinates. Based on this model, we can compare the morphological differences of scanning background traffic in different time windows or different network locations, quickly and accurately locating the abnormality in the mass scanning flow. It can further contribute to the network traffic anomaly detection which is the key component of scanning-situational-awareness. At the same time, this method can also be used to model Internet traffic in other contexts.

The rest of this paper is organized as follows: section 2 summarizes and analyzes the research emphasis and status quo of network scanning; the third section describes the basic research idea of this paper and the source of analysis data; the fourth section defines the specific behavior of scanning baseline model generation algorithm; section 5 gives the results based on the algorithm run on measured data; the work is summarized in section 6.

II. Related Work

Internet measurement institutions: Cooperative Association for Internet Data Analysis (CAIDA) released their Network Telescopes[28] project results in the form of technical report in 2004. The project collected unsolicited traffic on the Internet through a /8 address block, while the other two similar systems: Internet Sink[29] and Internet Motion Sensor - IMS [30] also published a similar job almost at the same time, and called the address blocks obtaining these flows ‘Darknet’. R.Pang et al. called the traffic obtained in this way Internet Background radiation-IBR traffic for the first time[31]， and divided the IBR traffic into three categories, namely backscatter, scanning and others, through statistical analysis. In the paper, the high-frequency scanning phenomenon of 445, 80, 135 and other ports were studied and analyzed. This was the first time that scanning traffic has been referred to as a type of Internet background traffic.

Works have been done on IBR traffic acquisition and analysis[13][26][32][33][34][35][36][37][38]. Some important conclusions from the analysis of scanning traffic so far include: 1) scanning as a kind of background radiation flow exists in the whole Internet space for a long time; 2) the scanning flow and behavior features obtained at the same observation position are relatively stable; 3) worms, software vulnerabilities, botnets and other network security events are highly correlated with scanning traffic.

Some work related to this work includes:

[26] studied two /16 network prefixes and a small amount of /24 network prefixes as observation objects in California, Berkeley's Lawrence Berkeley national laboratory (LBNL) (in CA, USA), analyzing their network traffic logs and connection record abstracts for up to 14 years (from June 1, 1994 to December 23, 2006), and carrying on a comprehensive discussion and analysis on the scanning information from many different angles. The main conclusions are: 1) the scanning traffic increased rapidly since 1998; 2) there was an explosive increase in the number of scanning hosts in 2001, which was closely related to the Code Red and Nimda worm outbreaks; 3) since 2001, the detection activities of scanning have become normal, and this was considered as the beginning of "background radiation" on the Internet. Similarly, the work in [32] studied the Internet background radiation flow in 2010 and found that the Conficker in 2009 led to an increase in the scanning activity of port 445.

Before that, two open-source high-speed scanning tools Zmap[24] and Masscan[39] were launched in 2013, followed by the work[13] published in 2014. By some scanning operations and scanning traffic from a Darknet’s which received traffic from January 1, 2013 to May 1, 2014, consisting of 5.5 million addresses, the literature analyzed and summarized the detected scanning methods from the Internet scanning phenomenon, and concluded that: 1) large horizontal scanning was very common, most of which was not malicious; 2) user networks generally lack scanning-oriented security defense.

From the above analysis, it can be seen that the current scanning monitoring is still mainly focused on the observation of ‘Darknet’ traffic data. The analysis of scanning behavior still remains at intuitive observation (such as using various simple forms of statistical analysis and visualization processing on observation results), and lacks accurate and normative theoretical description methods. The research on the baseline model of network scanning behavior in this paper is an attempt to establish a more detailed theoretical description model for scanning behavior to support the intention recognition and anomaly detection of scanning behavior.

III. Problem Formulation

I Background of the Problem

Scanning traffic in the operating network is composed of both the Internet-wide scanning radiation baseline and the local scanning[25]. In general, the scanning traffic received in the local network space within 24 hours is relatively stable, and can be called a scanning radiation baseline for the local network. Shifts in scanning radiation baseline are often associated with some malicious network activities, such as worm propagation and botnet. This suggests that anomalies in the network can be found in real time by observing the fluctuation of the local network scanning radiation.

This paper proposes a method to construct the scanning radiation baseline model based on the original scanning flow data of the Internet. This model can support 1) by analyzing the changes of the baseline in the same network space during different cycles, to find the possible scanning related anomalies in the traffic; 2) to find differences in composition of scanning background traffic in different network locations.

Since 2011, we have measured the IBR traffic (NJNET-IBR) for an ISP Network (Cernet Jiangsu Regional Network) covering more than 100 campus networks[21][27][40]. That is in possession of more than 1.2 million IP addresses among which 800,000 - 900,000 are inactive.

II Ports' Scanning Number Analysis

The algorithm proposed in this work is based on the assumption that each port being scanned should not be treated with equal significance. And this assumption is derived from following reasons:

First, the independence among all ports has not been shown. In our observation, we also find the existence of a certain proportion of completely random scanning phenomenon, in which the scanners send scanning packets to a target IP and port that is selected completely at random. This means that the scenario of the baseline model should consider all scanning traffic as a whole and not simply divide it up by port.

Secondly, the distribution of scanning traffic for 65,536 ports in each cycle is significantly uneven. In general, the order of ports in terms of received scanning packets is relatively stable over adjacent observation periods. This phenomenon is an important factor for the baseline. Since the background scanning traffic of the popular ports is obvious, the fluctuation caused by accidental factors will not have an impact on its stability, and vice versa for other ports. This observation shows that all ports cannot be treated equally in the construction of the baseline model.

IV. IBR Baseline Construction Algorithm

Consider dividing a circle into S segments of equal length while the weight of each segment is different. The centroid of a circle is shown in Fig.1. We associate this circle with the problem described in the previous section by putting the number of scanning on each individual arc segment, i.e. the weight of each arc segment is the number of scanning of the corresponding port.

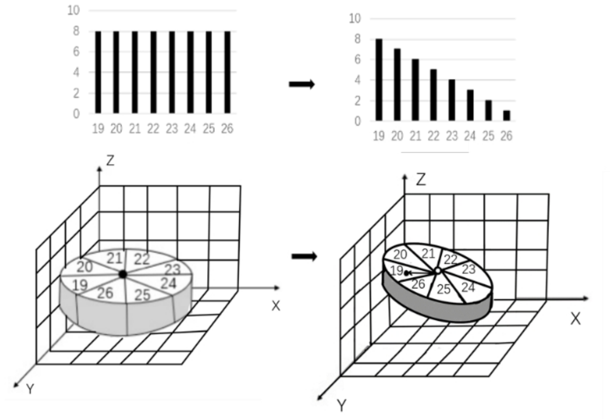

Fig.1 Diagram of the centroid of a ring

For simplicity, we directly provide the coordinates of the centroid of the circle in (1), in which S stands for the number of segments and m_i is the weight of the i-th segment.

$$x=\frac{S\sin\frac{\pi}{S}}{\sum_{i=0}^{S-1} m_i}\cdot\sum_{i=0}^{S-1} m_i\cos\frac{2\pi i}{S},\quad y=\frac{S\sin\frac{\pi}{S}}{\sum_{i=0}^{S-1} m_i}\cdot\sum_{i=0}^{S-1} m_i\sin\frac{2\pi i}{S} \qquad (1)$$

We create the baseline by tracking the coordinates of the centroid while removing the lightest "segment" on the circle one at a time. The algorithm is denoted as the Scanning_Baseline and we also use ToP_N to represent the sequence of removed ports. For ports not included in ToP_N, their traffic is combined and treated as a single port. The details are described as follows:

First, put the quantity of scanning of each port into the array Top_n_amount[0...N] accordingly. Specifically, ports in ToP_N[1...N] correspond to Top_n_amount[1...N] accordingly and the sum of remaining scanning numbers are stored in Top_n_amount[0], i.e：

$$\text{Top\_n\_amount}[i]=\text{Port\_Scanning\_Daily}_{(T,P)}[\text{ToP\_N}_{[i-1]}],\ i\in 1,\ldots,N \quad (2)$$

$$\text{Top\_n\_amount}[0]=\sum \text{Port\_Scanning\_Daily}_{(T,P)}[k],\ k\notin \text{ToP\_N} \quad (3).$$

Next, calculate the coordinates. In order to remove ports according to ToP_N[N…1], a new array Top_s_amount [0...S-1] is needed to record the first S scanning amounts in Top_n_amount(S decreases from N+1 to 1)， and the array Top_s_amount is required to be in reverse order, i.e：

$$\text{Top\_s\_amount}_{[i]}=\text{Top\_n\_amount}_{[S-i-1]},\quad i=0,\ldots,S-1 \quad (4).$$

Calculate the coordinates according to (1) by using Top_s_amount each time. There are N+1 coordinate points.

TABLE I. ALGORITHM 1

```
Centroid Baseline Algorithm Scaning_Baseline
Algorithm parameters： 1) Port_Scanning_Daily /* Raw data used for analysis*/; 2) ToP_N /* Removal order of ports*/
For i=1 to N
    calculate Top_n_amount[i]
While k ∉ Top_n
    calculate Top_n_amount[0]
For S=N+1 to 1
    For i=0 to S-1
        calculate Top_s_amount [i]
    calculate coordinates of x and y
```

V. Numerical Results

In this section, we will use the proposed algorithm to build the scanning radiation baseline for the campus network based on the measured data provided by NJNET-IBR introduced in 3.1.

I Experimental Data and Configurations

We select the background radiation traffic data provided by NJNET-IBR from December 21 to 27, 2020, and filter out the port scanning traffic for TCP, and further divide the traffic according to different campus networks, and select 60 campus network scanning traffic, i.e Port_Scanning_Daily(T,P)[0..65535]. This array represents the number of scanning on all 65,536 ports at time cycle T and position P. Specifically, we use T=1…7(20201221-20201227), P=1...60 (representing 60 campus networks respectively). There are a total of 420 sets of original data as shown in Table 2:

TABLE II. Original experimental data

T P	1221	. . .	1227
1	Port_Scanning_Daily(1221,1)[0..65535]		Port_Scanning_Daily(1227,1)[0..65535]
2	Port_Scanning_Daily(1221,2)[0..65535]		Port_Scanning_Daily(1227,2)[0..65535]
3	Port_Scanning_Daily(1221,3)[0..65535]		Port_Scanning_Daily(1227,3)[0..65535]
4	Port_Scanning_Daily(1221,4)[0..65535]		Port_Scanning_Daily(1227,4)[0..65535]
5	Port_Scanning_Daily(1221,5)[0..65535]		Port_Scanning_Daily(1227,5)[0..65535]
. . .			
57	Port_Scanning_Daily(1221,57)[0..65535]		Port_Scanning_Daily(1227,57)[0..65535]
58	Port_Scanning_Daily(1221,58)[0..65535]		Port_Scanning_Daily(1227,58)[0..65535]
59	Port_Scanning_Daily(1221,59)[0..65535]		Port_Scanning_Daily(1227,59)[0..65535]
60	Port_Scanning_Daily(1221,60)[0..65535]		Port_Scanning_Daily(1227,60)[0..65535]

II Experimental Scheme and Results

The algorithm is tested on two scenarios with different settings accordingly.

Scenario 1: Scanning baseline in 7-time cycles for each campus network with identical ToP_N parameter. This scenario can support stability-dependent analysis by distance between baselines.

We use the following scheme to construct the ToP_N parameter for this scene：

1) 7 original flow vectors Day(1)-Day(7) are generated by merging the original traffic of 60 campus networks, with subscripts [0..65535], corresponding to the sum of scanning traffic of each port, i.e.

$$\text{Day}(T)[\text{port}]=\sum_{P=1}^{60} \text{Port\_Scanning\_Daily}(T,P)[\text{port}],\text{port}\in[0,65535] \quad (5).$$

2) The 70 ports with the largest traffic from Day(1) to Day(7) are selected respectively, and there are a total of 123 ports excluding reduplication.

3) Furthermore, the traffic in Day(1)-Day(7) is merged into a total traffic vector All_amount[0..65535] by ports, and ToP_N is generated based on the size of 123 ports in All_amount[0..65535], i.e

$$\text{All\_amount[port]} = \sum_{T=1}^{7} \text{Day(T)[port]} \qquad (6).$$

Based on this method, we obtain the 7-day scanning baselines of 60 campus networks and two typical results are depicted in Fig.2. The baseline of Campus Network A for 7 days is very stable, while Campus Network B shows significant anomalies on two of these days.

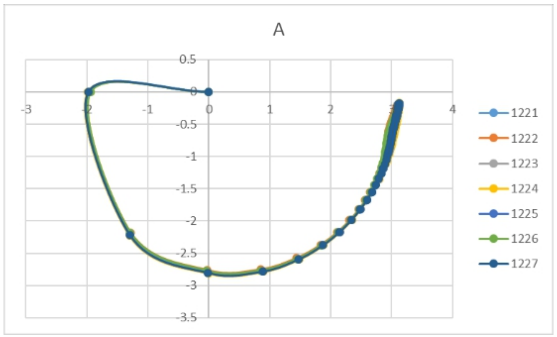

Fig.2(1) 7-day campus network A baseline

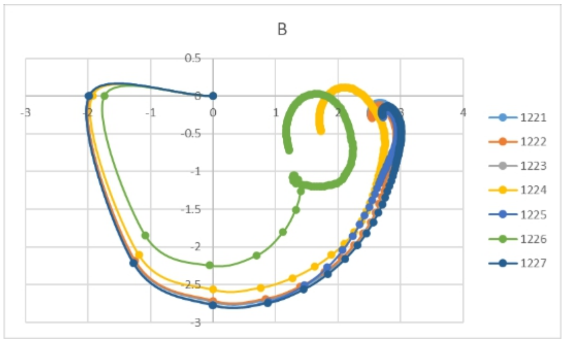

Fig.2(2) 7-day campus network B baseline

Scenario 2: In this part we focus on a single time cycle and observe the difference among 60 campus networks.

In this scenario, we construct a set of ToP_N of each individual day based on the original flow vectors Day(1)-Day(7), and set N to 120.

The baseline comparison of the maximum and minimum distance between the two campus networks A and B is shown in Fig.3. As can be seen, December 26, 2020 is the date with the maximum distance between the two campus networks, and December 22, 2020 is the date with the minimum distance between the two campus networks.

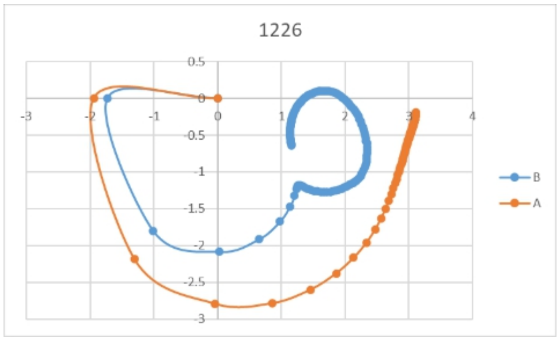

Fig.3(1) 1226 campus network A&B baseline

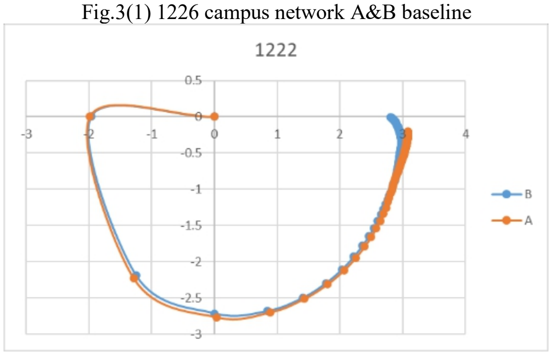

Fig.3(2) 1222 campus network A&B baseline

The experiments of the above two scenarios show that the Scaning_Baseline algorithm can profile background scanning traffic. Based on this profile, we can find anomaly in the original scanning traffic. On this basis, the cause of the anomaly can be located through further analysis.

VI. Conclusion

In this work, we try to construct a baseline generation algorithm for the port scanning traffic in the Internet background traffic based on the centroid point calculation of the non-uniformly distributed ring, and test it based on the measured traffic. The results show that the algorithm can effectively express the port scanning traffic facing different network spaces. Based on this expression, the baseline can be quantitatively analyzed from different perspectives such as time and space, which is helpful to locate the anomalies in the massive scanning traffic quickly and accurately. This method can also be used to model Internet traffic in other contexts.

Acknowledgment

The authors of this paper would like to thank Ms. Qiuyun Wu for providing the original data. This work is in part supported by Internet basic behavior measurement and analysis project: Internet basic behavior indicator system and measurement method under grant 2018YFB1800202.

References

[1] S.Bano,P.Richter, M.Javed, S.Sundaresan, Z.Durumeric, S.Murdoch,R.Mortier, and V.Paxson. Scanning the Internet for Liveness. ACM CCR,48(2),2018.

[2] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and A. Halderman. A Search Engine Backed by Internet-Wide Scanning. In ACM CCS,2015.

[3] ShadowServer.[EB/OL] https://www.shadowserver.org/, 2018-05-30.

[4] ZoomEye - Cyberspace Search Engine.[EB/OL]https://www.zoomeye.org/, 2018-05-30.

[5] Matherly J. Shodan[EB/OL]. https://www.shodan.io/, 2018.

[6] M. Antonakakis, T.April, M.Bailey, M.Bernhard, E.Bursztein, J.Cochran, Z.Durumeric, J.A.Halderman, L.Invernizzi, M.Kallitsis, D.Kumar, C.Lever, Z.Ma, J.Mason, D.Menscher, C.Seaman, N.Sullivan, K.Thomas, and Y.Zhou. Understanding the MiraiBotnet. InUSENIX Security Symposium,2017.

[7] A.Dainotti, A.King, K.Claffy, F.Papale,andA.Pescapé. Analysis of a "/0" stealth scan from a botnet. IEEE/ACM Trans. Netw.,23(2):341–354,Apr2015.

[8] Technical Details Behind a 400Gbs NTP Amplification DDos Attack, https://blog.cloudflare.com/technical-details-behind-a-400gbps-ntp-amplification-ddos-attack/.

[9] Paxson V. An analysis of using reflectors for distributed denial-of-service attacks [J]. ACM SIGCOMM Computer Communication Review, 2001, 31(3): 38-47.

[10] C. Rossow. Amplification Hell: Revisiting Network Protocols for DDoS Abuse. In Proceedings of the 2014 Network and Distributed Systems Security Symposium (NDSS 2014), number February, pages 23– 26, San Diego, 2014. Internet Society.

[11] Durumeric Z, Kasten J, Adrian D, et al. The matter of heartbleed[C]//Proceedings of the 2014 Conference on Internet Measurement Conference. ACM, 2014: 475-488.

[12] Why would a Windows machine scan for port 137? https://superuser.com/ questions/1306406/why-would-a-windows-machine-scan-for-port-137. (IMC 2019-5).

[13] Z.Durumeric,M.Bailey,andA.Halderman.AnInternet-Wide View of Internet Wide Scanning. In USENIX Security Symposium,2014.

[14] Gadge J, Patil A A. Port scan detection[C]// Networks, 2008. ICON 2008. 16th IEEE International Conference on. IEEE, 2008: 1-6.

[15] Bhuyan M H, Bhattacharyya D K, Kalita J K. Surveying port scans and their detection methodologies[J]. The Computer Journal, 2011.

[16] Myers D, Foo E, Radke K. Internet-wide scanning taxonomy and framework[C]//Proceedings of Australasian Information Security Conference (ACSW-AISC), 27-30 January 2015. Australian Computer Society, Inc, 2015.

[17] Kao C N, Chang Y C, Huang N F, et al. A predictive zero-day network defense using long-term port-scan recording[C]//Communications and Network Security (CNS), 2015 IEEE Conference on. IEEE, 2015: 695-696.

[18] Katterjohn K. Port scanning techniques[EB/OL]. http://www.scribd.com/doc/3950444/Port-Scanning-Techniques, 2018-04-28.

[19] Graham R D. Masscan: Mass ip port scanner[J/OL]. https://github.com/robertdavidgraham/masscan, 2018-05-13.

[20] Lyon G. The art of port scanning[J]. Phrack Magazine, 1997, 7(52).

[21] Staniford S , Hoagland J A , Mcalerney J M . Practical Automated Detection of Stealthy Portscans[J]. Journal of Computer Security, 2002, 10(1-2):105-136.

[22] Leckie C , Kotagiri R . A probabilistic approach to detecting network scans[C]// Network Operations & Management Symposium. IEEE, 2002.

[23] D.Adrian, Z.Durumeric, G.Singh, and A.Halderman. Zippier Zmap: Wnternetwide Scanning at 10Gbps. InUSENIX WOOT,2014.

[24] Durumeric Z, Wustrow E, Halderman J A. ZMap: Fast Internet-wide Scanning and Its Security Applications[C]//USENIX Security Symposium. 2013, 8: 47-53.

[25] Philipp Richter and Arthur Berger. 2019. Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope. In Internet Measurement Conference (IMC ’19), October 21–23, 2019, Amsterdam, Netherlands.

[26] M.Allman, V.Paxson, and J.Terrell. A Brief History of Scanning. In ACM IMC, 2007.

[27] Wu Qiu-yun，Ding Wei. Analysis of Internet scanning behavior based on dynamic dark network. Journal of Zhejiang University (Engineering Science), 2020, Vol54(8),1550-1556.

[28] Moore D, Shannon C, Voelker G M, et al. Network Telescopes [R]. CAIDA, Technical Report, 2004.

[29] Yegneswaran V, Barford P, Plonka D. On the Design and Use of Internet Sinks for NetworkAbuse Monitoring [C]. In: Proceedings of the Symposium on Recent Advances in Intrusion Detection, Springer. 2004: 146-165.

[30] Cooke E, Bailey M, Mao Z M, et al. Toward Understanding Distributed Blackhole Placement [C]. In: Proceedings of 2004 ACM Workshop on Rapid Malcode, ACM, 2004: 54-64.

[31] Pang R, Yegneswaran V, Barford P, et al. Characteristics of internet background radiation [C]. In: Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, ACM, 2004: 27-40.

[32] Wustrow E, Karir M, Bailey M, et al. Internet Background Radiation Revisited [C]. In: Proceedings of the 10th ACM Conference on Internet Measurement (IMC'10), 2010: 62-74.

[33] Eduard Glatz, Xenofontas Dimitropoulos. Classifying Internet One-way Traffic [C]. In: Proceedings of the 12th ACM Conference on Internet Measurement(IMC'12), 2012.

[34] Eduard Glatz, Xenofontas Dimitropoulos. Classifying Internet One-way Traffic [C]. In: Proceedings of the 12th ACM Conference on Internet Measurement(IMC'12), 2012. [miao25].

[35] Team Cymru Darknet Project [EB/OL].

[36] Bailey M, Cooke E, Watson D, et al. Practical Darknet Measurement [C]. In: Proceedings of the 40th Annual Conference on Information Sciences and Systems, IEEE, 2006: 1496-1501.

[37] Bailey M, Cooke E, Jahanian F, et al. Data Reduction for the Scalable Automated Analysis of Distributed Darknet Traffic [C]. In: Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement, USENIX Association, 2005: 21-21.

[38] Sherwood R, Gibb G, Yap K K, et al. Can the production network be the testbed? [C] In: Proceedings of the 9th USENIX conference on Operating systems design and implementation, USENIX Association, 2010, 10: 1-6.

[39] R. Graham. MASSCAN: Mass IP port scanner. https://github.com/ robertdavidgraham/masscan.

[40] Miao L, Ding W, Yang W. Extracting and Analyzing Internet Background Radiation in Live Networks 2015 Journal of Software.