doi:10.3969/j.issn.1001-0505.2017.S1.005

基于 IBR 的 ShadowServer TCP 扫描行为分析

丁 伟 王 力 武秋韵 夏 震

(东南大学计算机科学与工程学院,南京 211189)

摘要:为了对恶意扫描与非恶意扫描进行过滤,提出了一种基于白名单过滤非恶意扫描流量的方法.该方法首先以著名的安全机构 ShadowServer Foundation 的扫描主机作为白名单基础,将从 Shodan 搜索引擎中找出的部分 ShadowServer 扫描主机作为初始白名单集合.然后基于初始白名单集合以及在 CERNET 南京主节点边界获取的 IBR 流量,过滤出属于初始白名单主机的 TCP 扫描流量.最后通过分析这些流量的扫描行为,设计了一种完整白名单获取算法,运行算法并找出所有的白名单主机.实验结果表明,找到的白名单主机共计 229 个,其 IP 地址主要分布在 4/26 个网段中,在其中的 3 个网段内为连续地址,另一个网段内也有一定规律.此外,根据实验过程中的流量数据,提供了对 30022 端口和 445 端口(勒索病毒)扫描的两个案例及分析.

关键词:互联网背景辐射;扫描;ShadowServer;勒索病毒

中图分类号:TP393.0 **文献标志码**:A **文章编号**:1001-0505(2017)S1-0025-05

Analysis on ShadowServer TCP scanning behavior based on IBR

Ding Wei Wang Li Wu Qiuyun Xia Zhen

(School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

Abstract: To distinguish between malicious scanning and non-malicious scanning, a method for filtering non-malicious scanning traffic based on white list is proposed. First, a well-known security agency ShadowServer Foundation's scanning hosts are used as white list and some of the ShadowServer scanning hosts from the Shodan search engine are regarded as the initial white list. Then, the TCP scanning traffic is filtered based on the initial white list and the IBR traffic acquired on the CERNET Nanjing master node boundary. Finally, by analyzing the scanning behavior of the scanning traffic, a complete white list acquisition algorithm is designed to find out all the white list hosts. The experimental results show that, a total of 229 white list hosts are found and their IP addresses are mainly distributed in 4/26 network segment, in which the three network segments have the continuous addresses and another network segment also has a certain law. In addition, based on the data obtained in the experiment, two cases and their analyses about the scanning for port 30022 and port 445 (extortion virus) are provided.

Key words: internet background radiation (IBR); scan; ShadowServer; extortion virus

互联网背景辐射(internet background radiation, IBR)流量是指未经请求的单向流量[1-3].根据相关文献[4],IBR 可以分为反向散射、扫描和其他三类,其中,扫描流量是 IBR 流量的主要组成部分[5],所以可以通过对 IBR 流量的分析进行扫描相关研究.

扫描是用于探查活跃主机上开放了哪些 TCP/UDP 端口的技术方法[6].攻击者利用扫描技术,可以识别出在线的计算机上启用的服务、系统信息、应用程序的版本,从而分析和发现漏洞,这是

收稿日期:2017-09-20. **作者简介**:丁伟(1962—),女,博士,教授,博士生导师,wding@njnet.edu.cn.

基金项目:国家自然科学基金资助项目(61602114).

引文格式:丁伟,王力,武秋韵,等.基于 IBR 的 ShadowServer TCP 扫描行为分析[J].东南大学学报:自然科学版,2017,47(S1):25-29.[doi:10.3969/j.issn.1001-0505.2017.S1.005]

所有非法入侵行为的第一个步骤[7]. 然而扫描并不完全是恶意的,它同样也是研究和测量互联网的工具[8]. 目前比较正规的进行非恶意扫描的组织有 ShadowServer Foundation[9],Shodan[10]和 Zoomeye[11]等.

IBR 中的 TCP 扫描报文具有明显的特征(TCPflags = SYN),可以非常方便地将它们从 IBR 流量中分离出来. 如果可以获取足够规模的 IBR 流量,是进行互联网 TCP 扫描分析的理想的数据源. IBR 流量的获取一般分为基于暗网和基于运行网络获取两大类. 暗网指的是配置了路由但未被使用的网络空间(IP 地址段),因此暗网收到的所有流量都是 IBR 流量[12]. 暗网大多使用/8 大小的地址块且均位于美国. 这样的暗网在像中国这样的 IP 地址相对匮乏的地区是很难部署的,IBR 流量只能基于运行网络获取. 运行网络是实际使用中的网络,它的地址空间可以划分为活跃地址、不活跃地址和未触碰地址三类,其中不活跃地址空间被视为运行网络中的"暗网". Harrop 等[13-14]定义有暗地址和活跃地址混合的网络为灰网络,并使用灰网络获取 IBR 流量,同时他们证明了灰网依然可以检测扫描等网络威胁.

本文将基于在 CERNET 南京主节点边界获取的 IBR 流量,尝试对其中的 TCP 扫描流量进行分析. 文中所有的工作均围绕 TCP SYN 扫描进行(以下所有的扫描均指 TCP 扫描),具体的工作将从两个思路展开. 首先是尝试对扫描流量进行非恶意过滤. 采用的思路是"白名单",选择的对象是 ShadowServer Foundation,即尝试找到该机构所有非恶意扫描主机的 IP 地址;另一个是选择了两个特别的扫描案例进行介绍和分析.

1 相关背景

1.1 ShadowServer 机构

ShadowServer 机构成立于 2004 年,由专业的从事于网络安全方面的志愿者组成. 他们的主要活动是收集、跟踪和报告恶意软件、僵尸网络活动和计算机欺诈行为. 对全球范围的主机进行扫描是他们宣布的项目之一. 扫描的目的是发现具有漏洞的端口并将找到的漏洞报告给相应的网络运营商. 他们根据 US-CERT 的报告和其他具有重大安全隐患的协议确定扫描目标.

ShadowServer 所有的扫描项目在他们的官网上都有相应介绍. 目前,它公布的 TCP 扫描端口一共有 18 个,分别是 23, 80, 137, 138, 139, 443, 1 900, 2 323, 3 389, 5 900, 6 379, 7 547, 9 200, 11 211, 27 017, 27 018, 27 019, 28 017. 为了陈述方便,将上述 18 个端口构成的集合称为 ShadowServer_Scaning_Port. 所有参与扫描的主机都开放了 80 端口,并放置了相关网页. 因为结果的准确性可以方便地验证,所以选择它作为首个白名单建立对象.

1.2 CERNET 南京主节点网络边界的 IBR 流量

从 2014 年开始,作者进行 IBR 相关的研究工作[3-5],并在 CERNET 南京主节点网络边界建立了一个 NJNET_IBR 系统对 IBR 流量进行实时采集. 该系统以 RIBRM(Real-time Internet Background Radiation Measurement)算法为核心工作. 获取的 IBR 流量主要来自于网内接入单位的已配置但未使用的地址空间. NJNET_IBR 系统可以发现这些没有使用的空间,并捕获来自网外发往这些非活跃空间的 IBR 流量. 实际上,这样的非活跃空间是"流动"的,因此具有不易"定位"的特点,其中 IBR 流量更加真实,从而更具分析价值. 表 1 给出了近期 NJNET_IBR 获取的 IBR 流量的一些统计数据.

表 1 2017 年 7 月 9 日 NJNET_IBR 基本情况

Space(IP)	Bytes	Pkts	SYN_Pkts	SYN_Pkts/ Pkts
1198088	1 346 G	3.1 G	1.43 G	46.17%

NJNET_IBR 系统并不保存 IBR 流量,为了进行本文相关的研究工作,作者从 2017 年 5 月 9 日开始保存其中所有的 TCP_SYN 相关的摘要数据.

2 基于 NJNET_IBR 的 ShadowServer 扫描行为观测

本文工作主要目标是建立 ShadowServer 完整的、面向 CERNET 南京主节点网络的扫描地址白名单,在此基础上获取其在 NJNET_IBR 流量中全部的扫描流量,并对其行为进行统计分析. 本节将详细介绍这项工作的思路、过程和结果.

2.1 白名单的建立思路

虽然现在互联网上 IP 地址归属的查询是一个常见的服务,但根据归属查询其所有的 IP 并不是一件容易的事情. 为了完成这个工作,首先考虑到有这样的一个事实存在:所有的属于该机构(ShadowServer Foundation)的扫描主机执行相同的程序,虽然在每次运行时它们各自会被赋予的参数不同,但在程序所体现出的行为上,应具有一定一致性. 为此,白名单的建立思路是:首先尝试找到一部分符合白名单条件的主机(称为初始白名单集合),通过对其在 NJNET_IBR 中的扫描行为的观

测,确定检测规则并以此为基础形成检测算法,然后通过用其对 NJNET_IBR 中的扫描流量进行分析,得到最终的目标白名单.

定义 1[白名单条件]:一个在 NJNET_IBR 中有扫描有行为的 IP 地址,如果其 80 端口存在 web 页面则称其满足白名单条件.

2.2　初始白名单集合和其扫描行为特征

本文用 ShadowServer 和 80 端口等关键词,基于搜索引擎 Shodan[10] 获得了 39 个 IP 地址,通过浏览器连接后发现其全部存在 web 页面,随后在 NJNET_IBR 中的扫描流量中找到了这 39 个地址的扫描流量. 这样,初始白名单集合由这 39 个地址构成.

选择从 2017 年 6 月 3 日起连续 14 d 的数据作为分析源,对初始白名单中的每个 IP 从扫描端口和扫描次数进行了以 d 为单位的统计. 结果发现它们的扫描行为类似,规律如下:

1) 所有白名单主机扫描的端口均属于 ShadowServer_Scaning_Port.

2) 14 d 的观测期内,只有个别 IP 在个别天内没有扫描行为,将这种情况排除后呈现的规律如表 2 所示. 进一步分析,初始白名单中所有主机在 14 d 总体的扫描端口数量都为 11,且扫描次数接近,这表明它们的扫描行为是高度一致的.

表 2　初始白名单主机每日扫描行为

日期	最短扫描时间	最长扫描时间	最小扫描端口数	最大扫描端口数
6/3	6 543	25 334	4	10
6/4	5 192	31 405	4	10
6/5	3 453	18 656	3	11
6/6	2 971	21 538	4	10
6/7	4 721	17 864	3	10
6/8	4 476	23 913	4	11
6/9	2 634	20 052	4	11
6/10	1 893	16 110	4	11
6/11	1 647	13 648	2	10
6/12	3 896	29 394	5	10
6/13	7 913	45 167	3	10
6/14	3 793	41 862	2	11
6/15	2 521	14 561	3	11
6/16	2 428	18 134	2	10

2.3　完整白名单获取算法

根据初始白名单的扫描特征,本文设计了如下算法来寻找所有符合定义 1 的白名单主机.

算法由两个阶段构成. 首先根据初始白名单主机在 14 d 观测期内每天的扫描行为,从当天的扫描流量中过滤出候选名单,称之为灰名单-checkedip. 这样的灰名单一共有 14 个(每天一个). 第一阶段的算法描述如下:

输入:1 d 内所有 TCP 扫描报文(SIP DIP PORT)

输出:灰名单-checkedip,每台初始白名单主机当天的扫描端口数和扫描次数

操作:

① 根据当天全部的输入数据,统计每个源地址扫描的端口和对应的扫描次数,构造链表 map ipport;同时统计所有 39 个初始白名单主机当天的扫描端口数和扫描次数,获取其中有扫描行为的主机中的所有扫描的端口 ShadowServer_DailyScaning_Port、最少扫描端口数 p_0 和最少扫描次数 s_0.

(2) 删除 map ipport 中有对 ShadowServer_DailyScaning_Port 之外的端口有扫描行为的 IP;

③ 对剩余的 map ipport,删除其中所有扫描端口数少于 p_0 与扫描次数少于 s_0 的 IP.

④ 用剩余 map ipport 生成当天的灰名单-checkedip.

算法的第二阶段将根据 14 d 观测期单个初始白名单主机的持续行为,对 checkedip 中的地址进行二次过滤.

输入:14 d 的 checkedip

输出:目标白名单 whitelist(符合定义 1 的 IP 集合)

操作:

① 合并 14 d 的 checkedip,同时统计每个白名单 IP 在 14 d 内的扫描端口数和扫描次数,然后获取其中的最少扫描端口数 p_1 和最少扫描次数 s_1;

② 对合并后的 checkedip 中的每个 ip 地址,若其扫描的端口总数小于 $p_1 - n(1 <= n <= 7)$,或其扫描总数小于 $s_1 m(70\% <= m <= 90\%)$,则将其删除;

③ 用剩余的 checkedip 构成 whitelist.

2.4　白名单获取结果

灰名单中一共有 21 061 个 IP,对这些 IP 进行二次过滤,结果如表 3 所示. 对这 191 个 IP 分别进行浏览器连接,其中 189 个 IP 都出现了 web 页面,根据定义 1 将它们都放入白名单中. 之后,对其余两个 IP 地址 74. *. 47. 6 和 23. *. 141. 142 分别在 Shodan 和 AbuseIPDB 上查询它们的归属,前者归属于 Hurricane Electric,后者归属于 Columbus Networks Puerto Rico. 因为 74. *. 47. 63 的 IP 归属和白名单主机相同且扫描行为也与白名单主机一致,所以认为它也属于白名单. 最终,白名单由初始白名单中的 39 个 IP 和算法定位的 190 个 IP 共计 229 个 IP 地址构成,它由 4 个/26 网段组成,其中 3 个网段都是连续的,还有一个也是有规律的.

表 3 第二阶段运行结果

IP 数	$m=90\%$	$m=80\%$	$m=70\%$
$n=1$	186	190	190
$n=2$	186	190	190
$n=3$	186	190	190
$n=4$	186	190	190
$n=5$	186	190	190
$n=6$	186	190	190
$n=7$	187	191	191

2.5 ShadowServer 扫描主机行为观测

根据上小节获取的 ShadowServer 的扫描白名单,本文对其在 5 月 9 日 ~6 月 19 日连续 42 d 的扫描数据进行了统计分析,出现的异常情况有:

1) 在 5 月 10 日和 12 日,白名单主机出现了一定规模的对 30022 端口的扫描,这个端口不在其公布的 ShadowServer_Scaning_Port 内. 在下一节中,将对这个问题进行进一步的讨论. 除了这个事件外,所有的白名单主机的扫描行为均符合其公布的规范;

2) 主机 74.*.47.63 的 80 端口主页缺失. 除了上述 1)描述的异常情况数据,其余扫描数据统计结果如表 4 所示. 从表中可以看出,ShadowServer 公布的扫描端口一共有 18 个,但是真正进行扫描的只有 11 个,余下的 7 个中 27 018,27 019,28 017这三个端口与 27 017 属于同一类,137,138,139 和 1 900 这四个端口同时有 tcp 和 udp 扫描,他们进行了 udp 扫描没有进行 tcp 扫描.

表 4 白名单主机扫描数据统计结果

最大单日扫描次数	平均单日扫描次数	扫描端口数	扫描次数最多的端口
4 953 694	2 726 261	11	7 547(CWMP)

3 IBR 扫描流量中的两个案例

本小节介绍两个观测期出现在 NJNET_IBR 扫描流量中的值得关注的案例:一个是对 30 022 端口的扫描,另一个发生在勒索病毒相关的 445 端口.

3.1 30022 端口的扫描

对 ShadowServer 扫描主机的观测表明,该机构所有的主机的扫描范围均基本在其公布的端口范围内,唯一的例外是在 5 月 10 日和 12 日,出现了一定规模的白名单主机扫描 30022 端口的情况,如表 5 所示. 30022 端口不是 ShadowServer 公布的扫描端口,也不是有特定 Internet 服务和漏洞的端口,因此这个事件引起了人们的关注. 为此本文进一步过滤出相邻时间段内 NJNET_IBR 中面向 30022 端口的全部扫描流量,又发现了 6 个 IP 具有类似行为,其大规模扫描时间与白名单中的主机完全一致. 这 6 台主机全部是服务器,具体情况见表 6.

表 5 白名单主机 5 月 10 日和 12 日扫描 30022 端口情况

IP 数	IP 数占比	SYN 报文数	SYN 报文数占比
161	70.30%	1 348 000	28.73%

表 6 扫描 30022 端口的非白名单主机信息

IP	IP 归属	510 次数	512 次数	开放端口(Shodan)	开放端口(ZoomEye)
58.*.113.34	中国电信	31 846	32 292	80, 123, 8080, 8443	23, 3389, 443, 843, 8080, 9000
58.*.11.21	中国电信	11 812	12 281	80	80
58.*.113.34	中国电信	11 738	12 239	81,902,8090, 9080	81, 443,5445,8090,9080
112.*.4.98	中国移动	5 227	4 982	23	23, 443
211.*.35.126	中国移动	2 531	2 073	443, 3306, 4500	443, 3306
218.*.113.74	中国电信	5 098	5 193	23	23, 443

IP 数是扫描了 30022 端口的主机数,IP 数占比为 IP 数与主机总数之比,SYN 报文数是这两天扫描 30022 端口的报文总数,SYN 报文数占比为 SYN 报文数与这两天的扫描报文总数之比.

3.2 445 端口在 Onion 勒索病毒爆发期间的扫描流量

由于从 2017 年 5 月 9 日开始保留 NJNET_IBR 中的扫描流量,在随后的 5 月 12 日爆发了著名的 Onion 勒索病毒,故对全部 NJNET_IBR 中的扫描流量进行了面向该端口的以天为时间单位的统计,如图 1、表 7 所示. 从中可以看出 445 端口的 TCP 扫描报文数在 5 月 12 日当天有所增长但增长幅度不大,在 5 月 14 日大幅下降,随后一直没有提升.

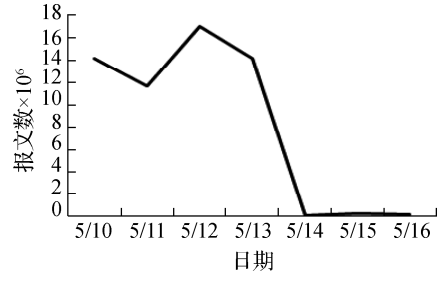

图 1 5 月 10 日 ~5 月 16 日 445 端口的 TCP 扫描报文数

表 7　5 月 14 日以后 445 端口的 TCP 扫描报文数

日期	5/14	5/15	5/16	6/17	6/18	6/19
报文数	104 794	197 563	185 221	60 752	66 415	54 004

4　结语

本文基于 NJNET_IBR 系统提供的的 IBR 流量,分析其中 TCP 扫描流量并进行了面向 ShadowServer 主机的扫描行为分析. 文中设计并实现了一种完整白名单获取算法,在实际网络环境中运行后得到白名单列表. 对其中白名单主机的扫描行为的观测发现,可以用一段时间内某主机扫描的端口和相应扫描次数判断其是否应该添加到白名单中. 此外,对实现过程中的两个案例进行了分析,案例 1 中 5 月 10 日和 12 日两天内,ShadowServer 对 30022 端口进行大规模扫描,但由于 30022 端口上并没有发现特殊的服务或漏洞,所以尚未分析出原因;案例 2 对 445 端口在 Onion 勒索病毒爆发期间的扫描流量分析,发现 445 端口在 5 月 12 日被扫描的次数略有上升,在 5 月 14 日大幅下降,之后没有回升.

参考文献(References)

[1] Wustrow E, Karir M, Bailey M, et al. Internet background radiation revisited[C]//*Proceedings of the* 10*th ACM SIGCOMM Conference on Internet Measurement*. ACM, Melbourne, Australia, 2010: 62 - 74.

[2] Dainotti A, Ammann R, Aben E, et al. Extracting benefit from harm: using malware pollution to analyze the impact of political and geophysical events on the Internet [J]. *ACM SIGCOMM Computer Communication Review*, 2012, **42**(1): 31 - 39.

[3] 张凌峰, 丁伟, 龚俭,等. 基于流记录的主干网活跃 IP 地址空间检测[J]. 软件学报, 2016, 27(S2): 43 - 49.
Zhang Lingfeng, Ding Wei, Gong Jian, et al. Active IP address space detection based on stream record[J]. *Journal of Software*, 2016, **27**(S2): 43 - 49. (in Chinese)

[4] 缪丽华. 互联网背景辐射流量的测量与研究[D]. 南京:东南大学, 2015.

[5] 杨扬, 丁伟. 互联网背景辐射流量的获取与统计分析[D]. 南京: 东南大学, 2016.

[6] 诸葛建伟. 网络攻防技术与实践[M]. 北京: 电子工业出版社, 2011:86 - 90.

[7] 李刚, 丁伟. 基于流记录的扫描和反射攻击行为主机检测[D]. 南京: 东南大学, 2016.

[8] Durumeric Z, Bailey M, Halderman J A. An Internet-wide view of Internet-wide scanning[C]//23*rd USENIX Security Symposium* (*USENIX Security* 14). San Diego, USA, 2014: 65 - 78.

[9] Shadowserver Foundation. Shadowserver [EB/OL]. [2017-09-24]. https://www. shadowserver. org/wiki/pmwiki. php/Main/HomePage.

[10] Wikipedia. Shodan [EB/OL]. [2017-09-24]. https://en. wikipedia. org/wiki/Shodan_(website).

[11] Baidu Encyclopedia. Zoomeye [EB/OL]. [2017-09-24]. http://baike. baidu. com/item/ZoomEye

[12] Pang R, Yegneswaran V, Barford P, et al. Characteristics of internet background radiation[C]//*ACM SIGCOMM Conference on Internet Measurement*. ACM, Sicily, Italy, 2004:27 - 40.

[13] Harrop W, Armitage G. Greynets: a definition and evaluation of sparsely populated darknets[C]//*ACM SIGCOMM Workshop on Mining Network Data*. ACM, Pennsylvania, USA, 2005:171 - 172.

[14] Harrop W, Armitage G. Defining and evaluating Greynets (Sparse Darknets)[C]//*The IEEE Conference on Local Computer Networks*, 2005. Anniversary. Sydney, Australia, 2005:344 - 350.