软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(3):663−679 [doi: 10.13328/j.cnki.jos.004516]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

运行网络背景辐射的获取与分析*

缪丽华[1,2], 丁 伟[1,2], 杨 望[1,2]

[1](东南大学 计算机科学与工程学院,江苏 南京 211189)

[2](江苏省计算机网络技术重点实验室,江苏 南京 211189)

通讯作者: 缪丽华, E-mail: lhmiao@njnet.edu.cn

摘 要: 因特网背景辐射(Internet background radiation,简称 IBR)是一种无功流量,已被广泛用于网络安全和管理等领域的研究中.传统的 IBR 获取方式——暗网系统存在较难满足的布置条件和易被避开的弊端,因此,提出一种从运行网络中获取 IBR 的算法.该算法基于灰空间、单向流和行为学习这 3 个概念,能够较准确地获取运行网络的所有 IBR 流量.一方面,它同时获取了不活跃地址和活跃地址的 IBR 流量,比现有的基于不活跃地址的算法漏判率低;另一方面,该算法在单向流基础上增加了基于源点的行为学习.与现有的基于单向流的算法相比,虽然查全率有少许降低,但查准率从约 93%提升至 99%以上.通过将算法运用到一个拥有约 128 万个 IP 地址的运行网络,从多个角度对该运行网络中的 IBR 进行了分析.结果显示,近两年,样本数据中 70%以上的入流为 IBR 流,这一现象应引起相关研究的注意.最后,通过几个安全事件案例说明了运行网络 IBR 流量在网络安全和管理等领域中的重要作用.

关键词: 因特网背景辐射;灰空间;单向流;IBR 分类;网络威胁

中图法分类号: TP393

中文引用格式: 缪丽华,丁伟,杨望.运行网络背景辐射的获取与分析.软件学报,2015,26(3):663−679. http://www.jos.org.cn/1000-9825/4516.htm

英文引用格式: Miao LH, Ding W, Yang W. Extracting and analyzing Internet background radiation in live networks. Ruan Jian Xue Bao/Journal of Software, 2015,26(3):663−679 (in Chinese). http://www.jos.org.cn/1000-9825/4516.htm

Extracting and Analyzing Internet Background Radiation in Live Networks

MIAO Li-Hua[1,2], DING Wei[1,2], YANG Wang[1,2]

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

[2](Key Laboratory of Computer Network Technology in Jiangsu, Nanjing 211189, China)

Abstract: Internet background radiation (IBR) is a type of unproductive traffic which has been used for years in the network security and management fields. Traditionally, IBR can be obtained by darknets. Nevertheless, the deployment of darknets typically requires large dark address blocks which are hard to acquire and also potentially detectable and avoidable. To address the issue, this article proposes an algorithm to extract IBR from raw traffic in live networks. The algorithm is based on the notions of grey spaces, one-way flows and behavior learning and has a better performance than previous work. On one hand, the algorithm obtains IBR destined to both inactive addresses and active addresses, resulting a lower missing rate compared with algorithms based on inactive addresses. On the other hand, the algorithm employs a behavior learning mechanism. Although the metric “recall” decreases slightly, “precision” increases from about 93% to above 99% in contrast to algorithms based on one-way flows. After applying the algorithm to a live network consisting of about 1.28 million IP addresses, the study analyzes the extracted IBR from several aspects. Results show that more than 70% of the inbound flows are IBR flows in the past two years’ data samples and this should draw enough attention from related research. Finally, several cases suggest the important role the live networks’ IBR traffic plays in the network security and management fields.

Key words: Internet background radiation; grey space; one-way flow; IBR classification; Internet threat

* 基金项目: 国家重点基础研究发展计划(973)(2009CB320505); 国家科技攻关计划(2008BAH37B04)
收稿时间: 2013-05-03; 修改时间: 2013-07-30; 定稿时间: 2013-11-11

因特网背景辐射(Internet background radiation,简称 IBR)指未经请求的单向流量(unsolicited one-way traffic),它是一种无功流量(unproductive traffic),并客观存在于因特网中[1,2].蠕虫或黑客的扫描、拒绝服务攻击(denial of service,简称 DoS)的反向散射(backscatter)、网络设备的错误配置等均可能会导致 IBR 的产生[1,3].因此,IBR 是网络安全和网络管理等研究领域很有价值的分析数据源.

所有 IBR 相关研究均基于实测数据.传统的 IBR 流量获取方法是使用暗地址空间(dark address space),即未使用的 IP 地址空间[3].此类系统中较著名的案例有:CAIDA 的 UCSD Network Telescope[4]、美国密歇根大学的 Internet Motion Sensor(IMS)[5]、威斯康星大学麦迪逊分校的 Internet Sink 系统(即 iSink 系统)[6]和 Cymru 团队的暗网项目(darknet project)[7],这些系统均位于 IP 地址资源相对充足的美国.

基于 UCSD Network Telescope 捕获的 IBR 流量,Moore 等人检测到 2003 年 12 月针对 SCO 组织的一次 DDoS 攻击、2003 年 1 月爆发的 Slammer 蠕虫以及 2004 年 3 月的 Witty 蠕虫[8−10];Dainotti 等人检测到 2011 年由政治原因造成的埃及和利比亚网络中断事件,并分析了 2011 年新西兰以及日本地震对网络基础设施造成的影响[2,11,12].基于 IMS 系统获取的 IBR 流量,Bailey 等人发现了 Blaster 蠕虫、Bagle 后门扫描、SCO 分布式拒绝服务攻击等安全事件[5,13].以上工作均是基于暗网 IBR 流量完成的.

使用暗网获取 IBR 流量方法的最大优势在于:所获取流量均是完全符合定义的 IBR 流量,在数量方面也可以满足分析需要.但这个方法也存在明显的问题,主要在于:(1) 为提高检测网络威胁的能力,暗网的布置通常需要充分大的地址空间[4],上述暗网系统大多使用/8 大小的暗地址块,这在 IP 地址相对匮乏的地区是很难做到的;(2) 这些暗网地址是固定的,长时间运行后,随着这些地址信息逐步被外界知晓,在 IBR 流量中占相当比重的扫描和反向散射流量的发起者可以很容易地避开这些暗网地址.这意味着:随着时间的推移,暗网所获得的 IBR 流量与实际运行网络中的 IBR 流量的特性和组成会有越来越大的偏差.

运行网络(a live network[14] or production network[15])是实际使用中的网络.为了更好地描述这个问题,简单地定义网络 N 的入流量 $Traffic(N)$由两个不相交的子集 IBR 流量(符号:IBR)和非 IBR 流量(符号:NonIBR)构成,即:$Traffic(N)=IBR(N)\cup NonIBR(N)$,其中,$IBR(N)\cap NonIBR(N)=\varnothing$.如果 N 是暗网,则 $NonIBR(N)=\varnothing$;若 N 是运行网络,则 $NonIBR(N)\neq\varnothing$.

近年来,针对运行网络中 IBR 流量的研究开始出现[14,16−21].在这个相关领域,有待解决的主要问题有两个:一是如何更准确地获取运行网络中的所有 IBR 流量,另一个是如何使用这些 IBR 流量.前者的重要性所有研究工作的基础;对后者而言,由于这些流量是从运行网络中获得的,因此除了检测 DDoS 攻击和蠕虫爆发等网络威胁外,研究工作可以首先围绕网络 N 的 IBR 流量和非 IBR 流量的比对等统计分析展开.文献[14,20,22]的研究表明,相关网络的 IBR 流已经占到其网络总入流数的 50%以上.在这样的情况下,IBR 流量的识别,特别是基于流记录的 IBR 流量识别是一项有意义的工作.

本文的研究工作围绕运行网络中 IBR 流量的获取与分析展开,主要贡献在于:

(1) 提出了适用于报文和流数据的运行网络 IBR 流量的抽取算法.基于一组通过向暗网流量混入正常流量构造的基准,将该算法与文献[14]的算法进行了比较,分析表明,本算法拥有较高的查准率.而上述基准还可以提供给 C4.5 等基于机器学习的流量识别经典算法使用;

(2) 使用该算法对时间跨度超过 4 年在同一采集点获得的 5 条实测数据(2008 年~2012 年)进行了 IBR 流量的抽取;

(3) 对所获取的 IBR 流量进行了统计和分类分析,从中获得了面向流数、报文数和字节数等不同测度的 IBR 流量比例、IBR 流量中各类成分的比例、IBR 流量的污染空间特点等有价值的统计数据;

(4) 对 IBR 流量中的 2 个主要成分进行了成因分析,并分析了其中较为突出的安全事件.

本文第 1 节对相关领域研究现状的简单介绍.第 2 节是 IBR 流量的获取算法和有关的分析.第 3 节是对原始流量数据的介绍.第 4 节是将算法作用到原始流量数据后所获得的 IBR 流量的分析.

1 相关工作

1.1 基于运行网络的IBR流量获取

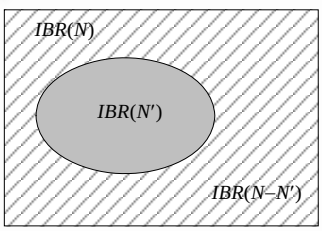
现有的运行网络 IBR 流量获取方法可以分为两类:

- 一类是基于运行网络中不活跃地址的方法,其基本原理与暗网相似.Harrop 等人[16,17]使用运行网络中的暗地址空间获取背景辐射流量,并证明依然可以检测扫描等网络威胁.Hoekzema[23]使用网内不活跃的(IP 地址,端口)二元组来获取背景辐射,并研究了活跃地址和不活跃地址收到的背景辐射的区别.这类方法的本质是:将网络 N 进一步划分成子网,使得其中的某个子网 N' 具有暗网的性质,即 N' 的入流量 $Traffic(N')$ 的 $NonIBR(N')$ 子集为空.用这样的方式获得的 IBR 流量并不是真正意义上的运行网络 IBR 流量,因为在网络 N 的其余部分 $N-N'$ 中还存在很多未分离出的 IBR 流量,如图 1(a)所示,其中,灰色部分表示该类方法获取的 IBR 流量,右斜纹部分表示漏判的 IBR 流量 $IBR(N-N')$;
- 另一类是基于单向流(one-way flow)的方法[14],这类方法只能工作在单宿主运行网络(single-homed live networks)的边界.根据 IBR 的定义,单向流仅是 IBR 的必要条件.在实际网络环境中,服务器的短时间关闭、设备对主动测量支持功能的关闭、组播应用、非公开的单向 UDP 应用都会产生单向流,但根据定义,它们均不是 IBR 流量.因此,该类方法会将一些非 IBR 流量的单向流误判为 IBR.这个问题可以通过将 $NonIBR(N)$ 划分成单向流量 $OneWayNonIBR(N)$ 和双向流量 $DoubleWayNonIBR(N)$ 两个不相交子集的方法进行描述,即 $NonIBR(N)=OneWayNonIBR(N)\cup DoubleWayNonIBR(N)$,此时,

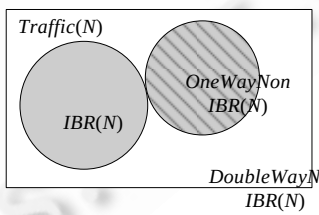
$$Traffic(N)=IBR(N)\cup OneWayNonIBR(N)\cup DoubleWayNonIBR(N).$$

文献[14]中算法将单向流全部认定为 IBR,则该算法所获得的 IBR 为 $IBR(N)\cup OneWayNonIBR(N)$,是存在误差的,误差的大小取决于 $OneWayNonIBR(N)$ 子集的情况,如图 1(b)所示,其中,灰色部分同样表示获取的 IBR 流量,左斜纹部分表示误判的 IBR 流量.

本文的目标是寻找一种可以较准确地获取集合 $IBR(N)$ 的算法.



(a) 基于不活跃地址的 IBR 获取算法示意图



(b) 基于单向流的 IBR 获取算法示意图

Fig.1 Existing algorithms of obtaining IBR in live networks

图 1 已有运行网络 IBR 获取算法

1.2 背景辐射分类

对 IBR 分类并分析其构成,是相关研究领域的一个重要研究内容,几乎所有文献均涉及这个问题.

Wustrow 等人[1]采用如下分类策略:TCP SYN 报文为扫描流量,TCP SYN+ACK,RST,RST+ACK,ACK 报文为反向散射流量,其余流量为错误配置流量.Dainotti 等人[2]认为,TCP SYN+ACK,TCP RST,ICMP echo reply 报文为反向散射流量,目的端口为 445、报文长度为 48 字节的 TCP SYN 报文为 Conficker 扫描.Glatz 等人[14]首先依据本地主机行为、远端主机行为、通信主机对的行为、流特征这 4 个方面,从一条流记录中提取出 17 个测度;其次,依据流记录的测度值和 13 条启发式规则将该 IBR 流分为以下 7 类:恶意扫描(malicious scanning)、反向散射(backscatter)、无法到达的服务(service unreachable)、良性 P2P 扫描(benign P2P scanning)、可能为良性的流量(suspected benign)、bogon 和其他(other).Brownlee[24]依据行为特征将源点分为 14 类、依据报文序列的到

达间隔将源点分为 10 组,从而将 IBR 报文集合划分为 140 个子集.这有助于从更细粒度上观测 IBR,并观测到总体 IBR 无法体现出来的变化.

本文在对这个问题进行讨论时,将综合采用文献[1,2]的分类方法.

2 运行网络 IBR 获取算法

本节要解决的问题是:在可以获取运行网络全部流量的条件下,过滤出运行网络收到的 IBR 流量.

2.1 问题描述

设 A 表示某个运行网络,将整个 IPv4 地址空间看作一个全集 U,则 U 可以划分为两个不相交的集合 A 和 B,即,$B=U-A$.A 与 B 之间可以存在一条或者多条通信路径.

设 T 为观测时间段,令 $Traffic(A,B,T)$表示 T 时间内 A 向 B 发送的所有流量集合.若为报文格式,则可写成 $Pkt(A,B,T)$;若为流记录格式,则可写成 $Flow(A,B,T)$,下同.同理,$Traffic(B,A,T)$表示 T 时间内 B 向 A 发送的所有流量集合,则 A 发送和收到的原始流量可以通过以下方法获得:

(1) 当 A 和 B 之间仅存在一条通信路径时,$Traffic(A,B,T)$和 $Traffic(B,A,T)$可以从该通信路径上采集获得;

(2) 当 A 和 B 之间存在 n 条通信路径时,$Traffic(A,B,T)$和 $Traffic(B,A,T)$必须从每条路径上采集并进行流量合并操作后获得.

在该前提满足的情况下,本文在 IBR 流量获取方面所要解决的问题可以描述为:基于 $Traffic(A,B,T)$和 $Traffic(B,A,T)$,设计算法 FIBR,从 $Traffic(B,A,T)$中抽取出运行网络 A 收到的 IBR 流量.其中,$B=U-A$,U 是整个 IPv4 地址空间.

2.2 IBR流量获取算法FIBR描述

首先,将所关注的运行网络 A 内的地址分为 3 种类型:

定义 1(活跃地址、灰地址、未触碰地址). 任意 IP 地址 $a\in A$,若 a 在 T 时间内既发出流量又收到流量,则称 a 为活跃地址;若 a 在 T 时间内不发出任何流量但收到流量,则称 a 为灰地址;若 a 在 T 时间内未收到任何流量,则称 a 为未触碰地址.

根据以上定义,A 可以被分为 3 个不相交的子集:

(1) 活跃空间 $Lit(A,T)$,包含所有活跃地址;

(2) 灰空间 $Grey(A,T)$,包含所有灰地址;

(3) 未触碰空间 $Untouched(A,T)$,包含 A 内所有未触碰地址.

根据 A 的划分结果,B 内的地址可以分为 4 类:

定义 2(正常地址、可疑地址、异常地址、不相关地址). 任意给出 $b\in B$,若 b 仅向 A 的活跃空间 $Lit(A,T)$发送流量,则称 b 为正常地址;若 b 既向 $Lit(A,T)$发送流量又向 A 的灰空间 $Grey(A,T)$发送流量,则称 b 为可疑地址;若 b 仅向 $Grey(A,T)$发送流量,则称 b 为异常地址;若 b 不向 A 发送任何流量,则称 b 为不相关地址.

根据以上定义,B 可以划分成 4 个不相交的子集:

(1) 正常空间 $Normal(B,T)$,包含所有正常地址;

(2) 可疑空间 $Suspicious(B,T)$,包含所有可疑地址;

(3) 异常空间 $Abnormal(B,T)$,包含所有异常地址;

(4) 不相关空间 $Unrelated(B,T)$,包含所有不相关地址.

将 $Flow(B,A,T)$中的流与 $Flow(A,B,T)$中的流进行匹配,存在匹配的流称为双向流,剩余的则是单向流.则 A 和 B 的划分以及与 $Flow(B,A,T)$之间的关系如图 2 所示,其中,B 的不相关空间未在图中画出.图中,粗箭头(1#, 3#)表示双向流,细箭头表示单向流.

事实上,灰地址和暗地址本质上类似,唯一区别是灰地址存在时间的约束.因此,灰地址不仅包含了未分配给主机的地址(即暗地址),而且包含了已分配给主机但是 T 时间内不活跃的地址,其覆盖范围比暗地址更广.类

似于暗地址空间,灰空间收到的流量均为背景辐射流量.此外,运行网络的灰空间是动态变化的,不易被攻击者避开.

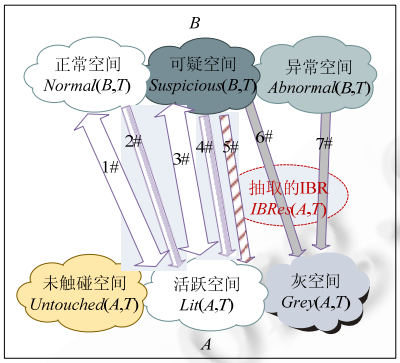


Fig.2 Schematic diagram of FIBR's main principles

图 2 算法 FIBR 的主要思想

根据上述分析,算法 FIBR 所要解决的主要问题就是找出活跃空间 *Lit*(*A*,*T*)收到的 IBR 流.文献[3,25]基于实测的背景辐射流量证明:一个源点发送的背景辐射流量通常是由一款或多款自动程序产生,因此,发往不同宿点的流量在行为上应具有相似性.本文的 IBR 流量获取算法 FIBR 便基于这个思路产生,其主要原理是:(1) IBR 均为单向流;(2) 图 2 中 6#和 7#箭头代表灰空间 *Grey*(*A*,*T*)收到的辐射流.通过对 6#箭头进行基于辐射源点的行为学习,标记 5#箭头为 *Suspicious*(*B*,*T*)向 *Lit*(*A*,*T*)发送的 IBR 流.因此,5#,6#和 7#箭头流之和即为本文算法 FIBR 抽取的背景辐射流估计集合 *IBRes*(*A*,*T*).

基于源点行为学习的原理是:对任意可疑 IP 地址,针对其向灰色空间 *Grey*(*A*,*T*)发送的流建立行为矩阵 *bvm*,并根据该行为矩阵,判定其向活跃空间发送的流是否为 IBR 流.行为矩阵是由若干行为向量构成的,而行为向量 *bv* 则由选定的若干流属性构成.设 $bvm=[bv_1,bv_2,\ldots,bv_n]$,则对任意 $bv_i,bv_j\in bvm,bv_i\neq bv_j$.本文算法使用到的两个行为矩阵相关操作为匹配操作和扩展操作,其中:匹配操作 *BVMM*(*bv*,*bvm*)的核心思想为,当 *bv* 与 *bvm* 的某列完全相等时匹配成功,返回 1,否则返回 0;扩展操作 *BVMI*(*bv*,*bvm*)的核心步骤为,执行 *BVMM*(*bv*,*bvm*),如果返回 0,则将 *bv* 作为新的一列插入 *bvm*.从行为矩阵的描述可看出:多条流记录有可能对应于同一条行为向量,行为向量是从流记录中进一步提取得到的更为简单的表述可疑 IP 地址行为规律的方式.

行为向量的流属性选择是算法 FIBR 的一个关键,选择的属性必须能够有效地区分 IBR 和非 IBR 流.因此,本文使用基于样本距离的特征选择算法——Fisher 记分法[26].当某个特征能使不同类样本之间具有最大距离而同类样本之间具有最小距离时,Fisher 记分法赋予该特征最高的 Fisher 分值.记一条运行网络的实测数据为 *Data*,其可疑地址集合为 *S*,对任意 $s\in S$,抽取出其属于 3#和 6#箭头(如图 2 所示)的流,并分别标记为非 IBR 流和 IBR 流,则所有可疑地址的流便构成了 Fisher 记分法的训练集.由于所有的 IBR 流均为单向流,根据文献[27],本文选定以下流属性为属性全集 *C*:{源 IP 地址,源端口,宿端口,协议,TCP flags,ICMP 类别,ICMP 代码,报文数,字节数,开始时间,结束时间,流持续时间,最大/最小/平均报文长度,最大/最小/平均报文到达间隔}.

对任意可疑地址 $s\in S$ 及任意属性 $c\in C$,记相应的 IBR 样本为 X_1,非 IBR 样本为 X_2.若 X_1 及 X_2 均不为空,则对可疑地址 *s*,*c* 的 *Fisher* 分为

$$F_s=S_b/S_w \tag{1}$$

其中,

- S_b 为类间离散度,描述两类样本间的距离,其计算公式为

$$S_b=(\bar{m}_1-\bar{m})^2+(\bar{m}_2-\bar{m})^2,$$

其中, $\bar{m}_1=\frac{1}{|X_1|}\sum_{x\in X_1}x, \bar{m}_2=\frac{1}{|X_2|}\sum_{x\in X_2}x, \bar{m}=\frac{1}{|X_1|+|X_2|}\left(\sum_{x\in X_1}x+\sum_{x'\in X_2}x'\right)$ 分别为 IBR 类样本、非 IBR 类样本和所有样本的均值;

- S_w 为类内离散度,描述类内样本间的距离,其计算公式为

$$S_w=S_1+S_2,$$

其中, $S_1=\frac{1}{|X_1|}\sum_{x\in X_1}(x-\bar{m}_1)^2, S_2=\frac{1}{|X_2|}\sum_{x\in X_2}(x-\bar{m}_2)^2$ 分别为 IBR 类样本、非 IBR 类样本的方差.

那么,对数据 *Data*,*c* 的综合 *Fisher* 值为

$$F_{Data}=\sum_{s\in S}\frac{|X_s|}{|X|}F_s \tag{2}$$

其中,*X* 指数据 *Data* 中属性 *c* 的所有样本,X_s 指可疑地址 *s* 发出流中属性 *c* 的所有样本.考虑到样本越多,方差和均值估计值的精度越高,*Fisher* 分值就越具代表性,因此,公式(2)中权值选取为每个可疑地址样本数在所有样本数中的比重.

对 CERNET(China Education and Research Network)江苏省网 2008 年~2012 年的 10 条 1 小时实测数据[28],按照上述方法计算每个属性的综合 *Fisher* 值.为了综合考虑这 10 条数据,本文采用公式(3)加权求和得到每个属性的最终 *Fisher* 值,从而确定本文的行为向量 *bv* 为(源 IP 地址,协议,报文数,字节数,TCP flags,ICMP 类别,ICMP 代码):

$$F_{final}=\sum_{i=1\sim10}\frac{|X_i|}{\sum_{i=1\sim10}|X_i|}F_{Data_i} \tag{3}$$

其中,X_i 表示数据 $Data_i$ 中属性 *c* 的所有样本.

设算法的输入为原始流量 *Traffic*(*A*,*B*,*T*)和 *Traffic*(*B*,*A*,*T*),抽取的背景辐射流量放在集合 *IBRes*(*A*,*T*)中(其初值为空),则背景辐射获取算法 FIBR 描述如下:

(1) 根据 *Traffic*(*A*,*B*,*T*)和 *Traffic*(*B*,*A*,*T*),定位 *T* 时间内 *A* 的灰空间 *Grey*(*A*,*T*)和活跃空间 *Lit*(*A*,*T*);

(2) 根据 *Traffic*(*B*,*A*,*T*),*Grey*(*A*,*T*)和 *Lit*(*A*,*T*),计算 *T* 时间内 *B* 的可疑空间 *Suspicious*(*B*,*T*);

(3) 根据 *Grey*(*A*,*T*),*Lit*(*A*,*T*),*Suspicious*(*B*,*T*)和 *Traffic*(*B*,*A*,*T*),获取 *Grey*(*A*,*T*)收到的背景辐射流量,并将其存入 *IBRes*(*A*,*T*);将 *Suspicious*(*B*,*T*)发送给 *Lit*(*A*,*T*)的流量抽取出来,暂时存入 *tmp_in*(*T*);

(4) 根据 *Lit*(*A*,*T*),*Suspicious*(*B*,*T*)和 *Traffic*(*A*,*B*,*T*),将 *Lit*(*A*,*T*)发送给 *Suspicious*(*B*,*T*)的流量抽取出来,暂时存入 *tmp_out*(*T*);

(5) 根据 *tmp_in*(*T*)和 *tmp_out*(*T*)获得 *Suspicious*(*B*,*T*)发送给 *Lit*(*A*,*T*)的单向流集合 *owf*(*T*).如果源数据为流数据,则仅需将出方向和入方向的流记录进行双向流匹配;如果源数据为报文数据,则首先需要将每个方向的报文组流,再进行双向流匹配.

(6) 根据 *IBRes*(*A*,*T*)和 *Suspicious*(*B*,*T*),获取 *Suspicious*(*B*,*T*)向 *Grey*(*A*,*T*)发送的 IBR 流;

(7) 初始化行为矩阵 *bvm*.对第 6 步获得的每条流,建立其行为向量 *bv*,通过矩阵扩展操作 *BVMI*(*bv*,*bvm*)将其插入 *bvm* 中;

(8) 对第 5 步获得的 *owf*(*T*)集合中的每条流,建立其行为向量 *bv*,并通过匹配操作 *BVMM*(*bv*,*bvm*)与 *bvm* 进行匹配.如果匹配成功,则将该条流加入 *IBRes*(*A*,*T*),即,将其标记为背景辐射流.

由于本文算法基于灰空间和单向流两个概念,因此本文算法工作有两个前提:

(1) 获取了运行网络 *A* 在 *T* 时间段内的所有通信流量;

(2) 运行网络 *A* 中存在灰空间.

为了满足第 1 个前提,算法 FIBR 一般应运行在接入网边界,其运行商或管理机构有条件获取该网络的全部网络流量,而其余研究人员则可从目前已有的数据公布网站获取研究流量.本文使用的所有流量数据均来自 CERNET 江苏省网,流量经匿名化处理后已公布于文献[28]处.对于第 2 个前提,在校园网或企业网中,灰空间是

普遍存在的;对其他网络,NAT 等技术的使用以及合理选择的 T 也可使第 2 个前提得到满足.

2.3 基于源点的漏判分析

在图 2 中,1#和 3#箭头表示存在匹配的双向流,它们一定不是 IBR.发往灰空间的 6#和 7#箭头则一定是 IBR.2#和 4#箭头中存在漏判可能,图中这 2 条箭头中白色表示非 IBR 单向流,而灰色代表漏判的 IBR 流.2#中 IBR 流被漏判的原因是源点不向 $Grey(A,T)$发送任何流量,因而无法将其标记为可疑或异常地址;4#中 IBR 被漏判的原因是其行为特征没有体现在 6#中,所以无法通过行为学习将其定位.此外,5#箭头中存在误判可能.

本节讨论 2#箭头中的 IBR 流被漏判的概率与活跃空间及灰空间大小之间的关系.根据上述分析,这部分流量被漏判的原因是其源点没有向 $Grey(A,T)$发送任何流量.

令 $ms(A,T)$为仅向 $Lit(A,T)$发送背景辐射的源点集合,$ms_ibr(A,T)$为 $ms(A,T)$发出的背景辐射流集合,则 $ms_ibr(A,T)$即为 2#箭头中的灰色部分.设 A 在 T 内真正收到的辐射流集合为 $IBR(A,T)$,定义测度辐射报文漏判率,用于本节的漏判分析,其定义如下,其中,$ms_ibr(A,T)$和 $IBR(A,T)$的大小用报文数来衡量:

定义 3. 辐射报文漏判率 $mr(A,T)=|ms\_ibr(A,T)|/|IBR(A,T)|$.

对任意背景辐射源点 H_i,假设其以固定的速率 r 向 A 发送背景辐射,并且每发送一个报文均以独立而随机的方式选择其目的 IP 地址.令 $IPcount=|Lit(A,T)|+|Grey(A,T)|$表示 A 内能够收到流量的 IP 地址数目,n 表示 $Lit(A,T)$的大小.

定理 1. H_i 被漏判的概率为 $P=(n/IPcount)^{rT}$,其中,T 为监测时间段.

证明:在上述假设成立的条件下,一个背景辐射报文的目的地址属于 $Lit(A,T)$的概率为 $p_0=n/IPcount$,则其发送背景辐射报文的过程为伯努利过程.在 T 内,H_i 发送 IBR 报文的数量为 $m=r\times T$,则 H_i 被漏判的条件是这 m 个报文均落入 $Lit(A,T)$.因此,H_i 被漏判的概率为 $P=p_0^m=(n/IPcount)^{rT}$,H_i 被检测到的概率为 $1-P$. □

这个定理说明,$p_0=n/IPcount$ 和 $m=r\times T$ 是影响漏判概率的重要因素.由于 p_0 代表活跃空间大小与活跃空间和灰空间总大小的比例,这意味着活跃空间越小或者灰空间越大,H_i 被漏判的概率越低.而 m 代表 H_i 发送的 IBR 报文数,意味着发送的 IBR 报文数量越多,H_i 被漏判概率越低.

图 3 给出了不同 p_0 和 m(横轴)与 H_i 被检测到概率 $1-P$(纵轴)之间的关系,p_0 取值为{0.1,0.3,0.5,0.7}.相应地,在 m 取最小值{3,6,10,20}时,H_i 被检测概率 $1-P\geqslant 99.9\%$.

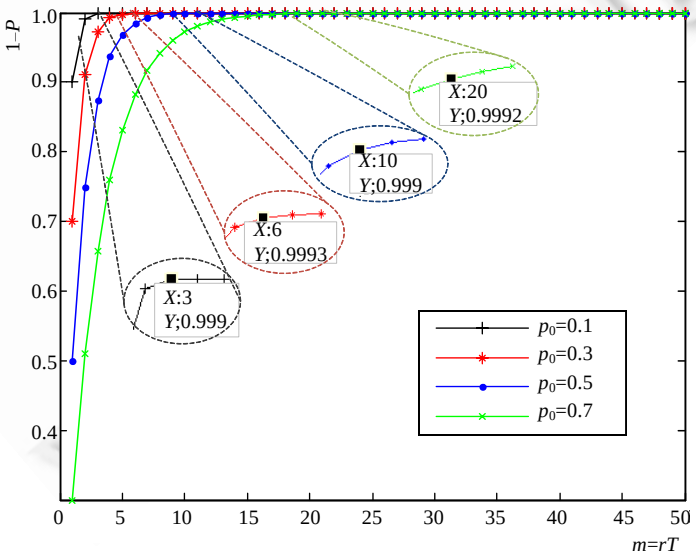


Fig.3 The relationship between IBR packet number m and 1–P

图 3 被检测概率 $1-P$ 与 IBR 报文数 m 之间的关系

由图 3 可知:$1-P$ 随着 m 的增大而增大(当 p_0 固定时),随着 p_0 的增大而减小(当 m 固定时).根据这个定理,

当 m=1 时,H_i 被漏判的概率最大(为 p_0),即,当该辐射源点只向 A 发送一个 IBR 报文时.

定理 2. 设向 A 发送 IBR 的实际辐射源点个数为 H.令随机变量ξ表示一个辐射源点发送的辐射报文数量,则其分布为 $P(\xi=i)=p_i,i=1\sim M$,其中,M 表示一个辐射源点可能发送的最大辐射报文数.因此,辐射报文漏判率为 $mr(A,T)=\sum_{i=1}^{M} ip_i p_0^i \Big/ \sum_{i=1}^{M} ip_i$.

证明:A 收到的所有辐射报文数为 $|IBR(A,T)|=E(\xi)\times H=\sum_{i=1}^{M} ip_i \times H$.由定理 1 可知:当一个辐射源点发送的所有辐射报文均发往 $Lit(A,T)$时,该辐射源点被漏判.因此,一个辐射源点被漏判的概率期望为 $\sum_{i=1}^{M} p_i p_0^i$,其被漏判的辐射报文数期望为 $\sum_{i=1}^{M} ip_i p_0^i$;又由于 A 的实际辐射源点数为 H,因此漏判的辐射源点数为 $|ms(A,T)|=\sum_{i=1}^{M} p_i p_0^i \times H$,漏判的辐射报文数为 $|ms\_ibr(A,T)|=\sum_{i=1}^{M} ip_i p_0^i \times H$.根据定义 3,辐射报文漏判率为

$$mr(A,T)=\sum_{i=1}^{M} ip_i p_0^i \Big/ \sum_{i=1}^{M} ip_i \tag{4}$$

根据上述分析,$mr(A,T)$与 H 无关,且当灰空间 $Grey(A,T)$不为空,即 p_0<1 时,$mr(A,T)<p_0$.假设$\delta=\xi-1$ 服从泊松分布,则 $P(\delta=i)=(\lambda^i e^{-\lambda})/i!,i=1\sim\infty$.此时,公式(4)的分子和分母分别可写为

$$\sum_{i=1}^{\infty} iP(\xi=i)p_0^i=\sum_{j=0}^{\infty} P(\delta=j)\times(j+1)\times p_0^{j+1}=p_0 e^{-\lambda}\sum_{j=0}^{\infty}\frac{(\lambda p_0)^j}{j!}(j+1)=p_0 e^{-\lambda}[\lambda p_0 e^{\lambda p_0}+e^{\lambda p_0}] \tag{5}$$

$$\sum_{i=1}^{\infty} ip_i=E(\xi)=E(\delta+1)=\lambda+1 \tag{6}$$

因此, $mr(A,T)=p_0 e^{-\lambda+\lambda p_0}(\lambda p_0+1)/(\lambda+1)$. □

λ和 p_0(横轴)以及辐射漏判率 $mr(A,T)$(纵轴)的关系如图 4 所示.图中λ有两个不同的取值λ={5,15},相应地,$E(\xi)$={6,16}.为保证 $mr(A,T)$<0.5%,相应的 p_0 最大值={0.32,0.69}.如图 4 所示:$mr(A,T)$随着 p_0 的增大而增大(当λ固定),随着λ的增大而减小(当 p_0 固定).

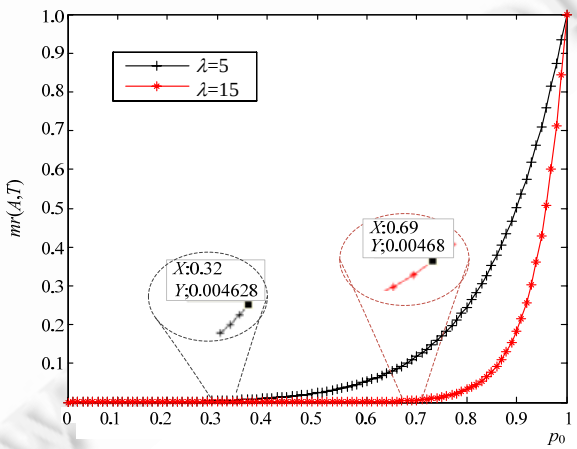


Fig.4 The relationship between p_0 and the missing rate $mr(A,T)$

图 4 辐射漏判率 $mr(A,T)$与 p_0 的关系

2.4 IBR行为学习有效性分析

本节将 $Suspicious(B,T)$发往 $Lit(A,T)$的背景辐射作为研究对象,比较本文算法 FIBR 和文献[14]中单向流方

法的查准率和查全率,从而证明算法 FIBR 中行为学习的有效性.FIBR 主要通过学习图 2 中 6#号箭头流量的行为来标记 *Suspicious*(*B*,*T*)发送给 *Lit*(*A*,*T*)的背景辐射(5#箭头),而 Glatz 等人[14]直接将 4#和 5#箭头的流量均认定为背景辐射.由于单向流是背景辐射的必要条件,Glatz 等人的方法可以获得 100%的查全率率,但查准率会偏低.

比较两种算法的方法是制作统一的基准(benchmark),基准制作算法的基本思想是向暗网流量中混入正常流量:

- 首先,对暗网 *D*,挑选一个与其规模相似的运行网络 *E*;
- 其次,将 *E* 的正常空间 *Normal*(*U-E*,*T*)与活跃空间 *Lit*(*E*,*T*)之间的交互流量抽取出来,再将 *Lit*(*E*,*T*)映射到 *D* 的某个随机选取的子集上,从而构造出 *D* 的虚拟活跃空间 *Lit*(*D*,*T*)和虚拟可疑空间 *Suspicious*(*U-D*,*T*);
- 最后,将 *Normal*(*U-E*,*T*)的某个随机选取的子集映射到 *Suspicious*(*U-D*,*T*)上,便可获得基准数据.

这个算法的具体描述如下:

1. 令 *IBR*(*D*,*T*)表示 *D* 在 *T* 内收到的流,则这些流全为背景辐射流,将其标记为 IBR;
2. 令 *flow_bi* 表示 *E* 内活跃空间 *Lit*(*E*,*T*)和正常空间 *Normal*(*U-E*,*T*)之间的交互流.将这些流标记为非 IBR;
3. 如果|*Lit*(*E*,*T*)|<|*D*|,将 *Lit*(*E*,*T*)与 *D* 内任意子集建立一一映射关系.令该子集为 *D* 的虚拟活跃空间 *Lit*(*D*,*T*),*D* 内剩余地址为虚拟灰空间 *Grey*(*D*,*T*),而 *Untouched*(*D*,*T*)为空;若|*Lit*(*E*,*T*)|≥|*D*|,则退出.
4. 根据步骤 3 的结果以及 *IBR*(*D*,*T*),可获得 *D* 的虚拟可疑空间 *Suspicious*(*U-D*,*T*);
5. 令 *X*=*Suspicious*(*U-D*,*T*),*Y*=*Normal*(*U-E*,*T*):
   a) 如果|*X*|≤|*Y*|,则在 *X* 和 *Y* 的任意子集之间建立一一映射关系.将 *flow_bi* 中的 IP 地址替换成其在 *D* 和 *Suspicious*(*U-D*,*T*)中的映射.将 *flow_bi* 和 *IBR*(*D*,*T*)混合,便可得到人造的基准;
   b) 如果|*X*|>|*Y*|,设 $K=\lceil |X|/|Y| \rceil$,将 *X* 划分成 *K* 个子集 X_1~X_K,其中前 *K*–1 个大小为|*Y*|,最后一个大小为|*X*|–|*Y*|×(*K*–1).将 X_1~X_{K-1} 分别与 *Y* 建立一一映射关系,将 X_K 与 *Y* 的任意子集建立一一映射关系.将 *flow_bi* 复制 *K* 份 *flow_bi*(1)~*flow_bi*(*K*),分别将其中的地址替换成其在 *D* 和 X_1~X_K 中的映射.将 *IBR*(*D*,*T*)和 *flow_bi*(1)~*flow_bi*(*K*)混合,从而得到人造基准.

第 3 步确保了 *D* 的虚拟灰空间 *Grey*(*D*,*T*)不为空,从而满足算法 FIBR 的工作前提.第 5 步主要是为了保证 *X* 中的所有地址均被混入非 IBR 流量.根据上节的分析,*flow_bi* 中有可能会混入少部分 IBR 流量,但当 *E* 的活跃空间占 *E* 整个空间的比例较小时,可以降低这个影响.

基于上述方法,我们从 CERNET 江苏省网内选择了一个包含 256 个地址的暗网 *D* 以及 3 个相似规模的运行网络 E_1,E_2 和 E_3.基于 2011-11-17 的 24 小时实测流量,我们制作了一组基准{BM_1,BM_2,BM_3}.3 个运行网络均为该省网的接入子网,其 IP 地址总数分别为{256,1024,256},在观测期内,活跃的 IP 地址数分别为{8,48, 154}.基于这些基准,算法 FIBR 和文献[14]中 E. Glatz 给出方法的查准率和查全率的对比见表 1,结果均用流数计算.其中,*TP* 表示真阳的个数,*FP* 表示假阳的个数,*TN* 表示真阴的个数,*FN* 表示假阴的个数.查准率、查全率和 F-Measure 的计算公式为

$$precision=\frac{TP}{TP+FP},recall=\frac{TP}{TP+FN},F\text{-}Measure=2\times\frac{precision\times recall}{precision+recall}.$$

从表 1 可以看出:虽然文献[14]的方法从原理上保证了 100%的查全率,但最低为 92.09%的查准率偏低;FIBR 在查全率上最低也可以达到 98.26%,超过 99%的查准率保证了所获得 IBR 的分析价值.这说明:简单地将单向流全部当作 IBR 流量,会在一定程度上造成误判.同时,表 1 还说明了本文所选行为向量的有效性.

Table 1 The Comparison between FIBR and E. Glatz's method

表 1 算法 FIBR 与 E. Glatz 方法的对比

		TP	*FP*	*TN*	*FN*	*Precision* (%)	*Recall* (%)	*F-Measure* (%)
BM_1	FIBR	4 056	0	2 788	72	100	98.26	99.12
	Ref.[14]	4 128	287	2 501	0	93.50	100	96.64
BM_2	FIBR	24 006	215	9 325	403	99.11	98.35	98.73
	Ref. [14]	24 409	2 097	7 443	0	92.09	100	95.88
BM_3	FIBR	75 547	529	48 029	1 300	99.30	98.31	98.80
	Ref. [14]	76 847	5 339	43 219	0	93.50	100	96.64

3 基于实测数据的 IBR 流量获取

3.1 分析数据

分析数据集采集的运行网络是 CERNET 的江苏省网,该网络是单宿主网络,采集的流量符合算法 FIBR 的条件.该运行网络接入的校园网数量超过 100 个,覆盖的 IP 地址总量接近 5 000 个 C 类地址.管理该接入网的 ISP 定期以 1/4 流抽样的方式在接入点采集报文头部(即,仅采集网内 1/4 地址空间的流量),并将这些数据经匿名化处理后发布[28].本文选取的分析数据集见表 2,其中,*IPcount* 和 p_0 的含义见第 2.3 节中的有关定义,入方向指进入省网的方向(下同).此外,表中所有流量及地址空间均为 1/4 流抽样后的结果(下同).p_0 偏小的原因是 NAT 技术在该运行网络中被普遍使用.根据第 2 节中的分析,较低的 p_0 能使算法 FIBR 的漏报率有效降低.

Table 2 The Dataset

表 2 数据集

日期		*T*	*IPcount*	p_0 (%)	入方向				出方向		
					#flows (10^8)	fps (10^3)	pps (10^3)	bps (10^6)	fps (10^3)	pps (10^3)	bps (10^6)
2008	Dec. 14	[14:00~15:00)	175 851	6.15	0.13	3.48	114.52	709.38	5.48	116.48	490.62
2009	Nov. 14	[00:00~24:00)	223 475	9.52	3.52	4.07	89.88	512.21	4.34	83.88	341.67
2010	Nov. 14	[00:00~24:00)	268 381	8.74	7.69	8.90	97.96	549.46	3.67	82.37	300.99
2011	Nov. 17	[00:00~24:00)	295 793	13.60	15.23	17.62	110.09	675.42	5.47	96.33	358.57
2012	Dec. 22	[00:00~24:00)	317 569	7.72	14.28	16.53	134.76	1009.92	5.94	71.69	370.26

3.2 IBR流量的基本情况

将 FIBR 算法分别作用于上述数据,组流基于五元组(源地址,宿地址,源端口,宿端口,协议),不活跃超时设置为 64s,其余组流和双向流匹配算法细节见文献[20].获得的 IBR 流量的基本情况见表 3,其中,IBR 平均速率由获取的 IBR 总量除以相应的观测时长(单位为 s)得到,IBR 平均大小则表示每个小时内每个 IP 地址平均收到的 IBR 流量大小,IBR 比例表示背景辐射占总入流量的比例.

Table 3 The Overview of the extracted IBR

表 3 运行网络 IBR 流量的基本情况

	IBR 平均速率			IBR 平均大小			IBR 比例		
	fps (10^3)	pps (10^3)	bps (10^6)	#flows	#packets	#bits (10^3)	#flows (%)	#packets (%)	#bits (%)
2008	1.01	1.41	0.64	20.60	28.83	13.12	28.89	1.23	0.09
2009	1.36	1.71	1.45	21.89	27.51	23.31	33.40	1.90	0.28
2010	5.33	5.59	2.14	71.53	75.01	28.69	59.89	5.71	0.39
2011	13.38	13.81	5.16	162.88	168.13	62.81	75.93	12.55	0.76
2012	11.74	12.58	5.99	133.05	142.55	67.91	71.01	9.33	0.59

表中的数据表明:IBR 流量有逐年增加的趋势,这与文献[1,20]中的研究结果一致.特别值得注意的是,在 2011 年和 2012 年的两条数据样本中,IBR 流量占用带宽不到 1%,但其流数比例超过 70%,这会对以流为基本处理单位的网络中间设备产生影响,造成网络资源的浪费;其次,文献[14]也反映了类似的现象,这说明高的 IBR 流

比例可能普遍存在,该现象值得引起相关研究的关注,如基于流的流量识别研究.

根据第 2.4 节的分析结果可知,本文算法的查准率较高,这意味着实验中被误判的非 IBR 流较少.而此类误判的产生原因是非 IBR 流的源点为可疑 IP 地址,且该非 IBR 流符合其源点发往灰空间流量的行为规律.

另一方面,本文算法存在一定的漏判,主要原因有两个:

(1) IBR 流的源点未向灰空间发送流量,则该 IBR 流被漏判.由第 2.3 节的分析可知,p_0 越小该类漏判越少,结合表 2 的 p_0 列可推断,实验中该类漏判的 IBR 流较少;

(2) IBR 流的源点为可疑 IP 地址,但该 IBR 流不符合其源点的行为矩阵,则它将被漏判.

4 运行网络中的背景辐射分析

本节从不同角度对实验获取的 IBR 流量进行分析,本节所有的讨论均以流数为单位进行.

4.1 协议比例分析

表 4 给出了 IBR 流量使用传输层协议 TCP,UDP 以及网络层协议 ICMP 的情况,从中可以看出,TCP 协议在背景辐射中仍然占据主导地位,这与整体流量在该测度上的表现存在差别.有关研究表明,近年来,国内网络整体流量中的 UDP 与 TCP 流量份额已不相上下[29].注意到,2009 年和 2012 年两条数据中 UDP 协议的 IBR 流量较多,我们将在第 4.4 节中具体分析该现象.

Table 4 IBR protocol breakdown

表 4 IBR 协议比例分析

	TCP (%)	UDP (%)	ICMP (%)
2008	89.63	3.76	6.61
2009	56.29	36.87	6.84
2010	96.30	2.08	1.62
2011	96.08	3.33	0.58
2012	86.66	11.63	1.71

4.2 分类分析

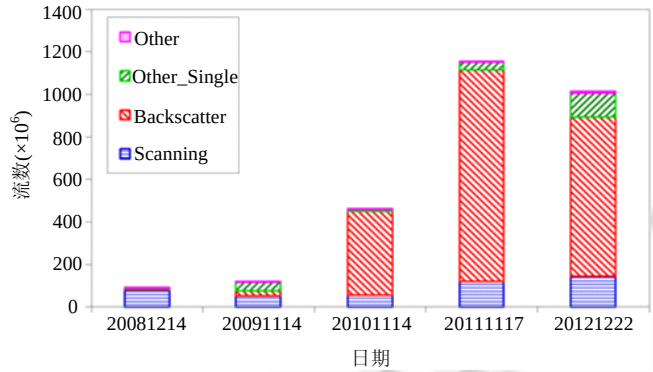
分类分析是 IBR 流量研究的一项重要内容,因为通过这项工作可以在很大程度上获得 IBR 的成因和特征,前者可用于发现网络中的异常现象,后者有助于在过滤规则的帮助下大幅提高从运行网络中获取 IBR 流量的效率.本节参考文献[1,2]中的分类方法,设计了一组基于流特征的易操作的分类规则,具体见表 5.

Table 5 IBR classification rules

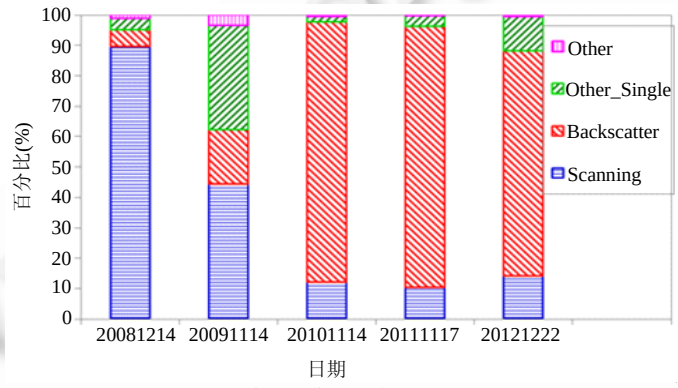
表 5 IBR 分类规则

类名称	描述
Backscatter	1. TCP SYN+ACK 单包流 2. TCP RST 单包流 3. TCP RST+ACK 单包流 4. TCP ACK 单包流 5. ICMP echo reply 单包/多包流
Scanning	6. TCP SYN 单包流 7. ICMP echo request 单包/多包流
Other_Single	8. UDP 单包流 9. 其余不属于 backscatter 和 scanning 类的 ICMP 单包流 10. 其余不属于 backscatter 和 scanning 类的 TCP 单包流
Other	11. 其余不属于以上 3 类的流

图 5 是根据上述分类完成的 IBR 各组成的堆积柱形图以及百分比堆积柱形图(图 5(a)中,2008 年数据所有成分的量均扩大为 24 倍,因为这条数据只有 1 小时的长度).从图中可以看出,other 成分所占比例很小(2008 年 1.07%,2009 年 3.54%,2010 年 0.47%,2011 年 0.36%,2012 年 0.71%).



(a) 各成分流数



(b) 各成分占 IBR 的比例

Fig.5 IBR classification

图 5 IBR 的分类分析

根据表 5,扫描流量可进一步分为 TCP SYN 扫描和 ICMP 扫描.图 6 是扫描流量各组成的堆积柱形图(类似于图 5(a),图 6 中,2008 年数据所有成分的量均扩大为 24 倍).从图中可以看出,TCP SYN 扫描为扫描流量的主要成分.同理,反向散射流量可进一步划分为 5 个子类.

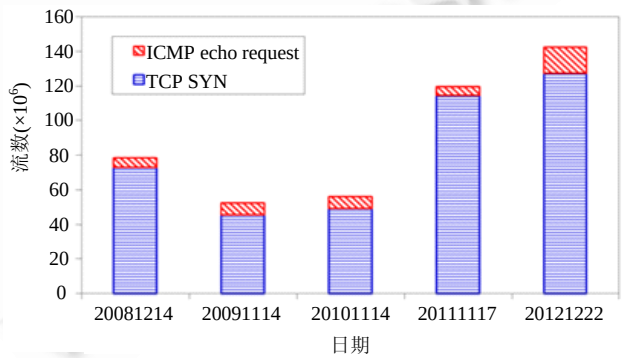


Fig.6 The composition of scanning

图 6 扫描的子类分析

图 7 是反向散射流量各组成的堆积柱形图(同样,图 7 中,2008 年数据所有成分的量均扩大为 24 倍).从图中可以看出:2008 年和 2009 年两条数据中的反向散射流量较少,而 2010 年~2012 年,这 3 条数据样本中的反向散

射流量相对较多,其中的主要成分为 TCP SYN+ACK 反向散射流量.

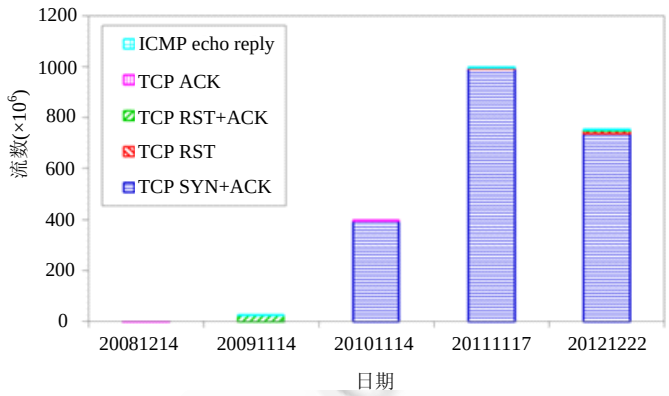


Fig.7 The composition of backscatter

图 7 反向散射的子类分析

4.3 空间特性分析

本节以/24 的地址空间为基本单位,对运行网络的不同地址空间所承受的 IBR 流量强度进行分析.图 8 和图 9 是 2009 年和 2012 年两条数据地址空间分析结果的 CDF 图.

为了方便比对,本文首先取 2009 年和 2012 年数据地址块的并集,并将两年数据中的地址空间均扩展为上述并集.图中横轴上的每个数字代表一个/24 地址块,是按实际地址顺序从小到大排列的.

两幅图显示:

(1) 扫描流量在整个地址空间上比较均匀;

(2) 反向散射和 other_single 类流量则集中在相同的一段地址空间上,2012 年数据的这个特征更加明显.IBR 流量也因此呈现出这种趋势;

(3) 2009 年和 2012 年数据中高 IBR 流量的 IP 地址段基本相同.

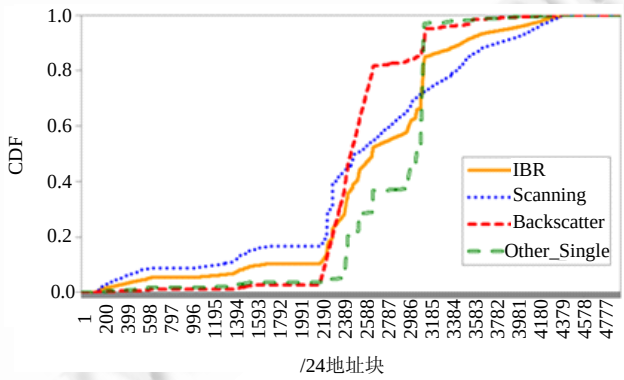


Fig.8 The distribution of destination /24 blocks in 2009

图 8 2009 年数据地址空间的累积分布图

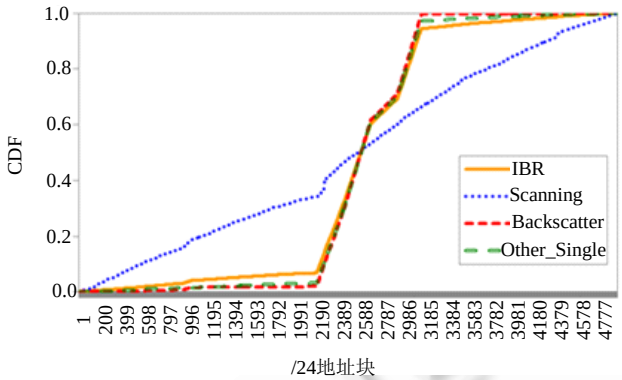


Fig.9 The distribution of destination /24 blocks in 2012

图 9 2012 年数据地址空间的累积分布图

4.4 other_single类流量成因分析

由于2009年和2012年数据IBR流量中的other_single类流量较多,本节基于这两年的数据分析other_single流量的可能成因.面向源端口的聚类分析显示:2009年数据的other_single类流中有90.69%的流是来自53端口,而其中99.99%由同一个主机$Victim_1$发出.报文层的分析显示:这些UDP报文均为DNS查询的标准回复报文,被查询的域名均为$Name_1$.显然,该现象是伪装源地址的DNS查询报文导致,这是一起针对域名服务器$Victim_1$的伪装源地址攻击.在时间方面,这起攻击贯穿整个24小时的观测期;在强度方面,根据本运行网络的观测,$Victim_1$收到的攻击报文速率至少为2000pps.2012年的数据显示:93.14%的other_single流来自53端口,其中99.84%由7个主机{$Victim_2$,$Victim_3$,$Victim_4$,$Victim_5$,$Victim_6$,$Victim_7$,$Victim_8$}发出,且同样均为DNS查询的标准回复报文,被查询的域名均为$Name_2$.这显然是一起针对$Name_2$相应域名服务器的伪装源地址攻击,攻击时间同样贯穿整个24小时观测期,强度超过5000pps.两次观测中,DNS反向散射的目的IP地址范围(即,运行网络内收到散射的地址范围)均属于上节空间分析中的高辐射地址段.

由于上述分析结果显示出了较高的攻击强度,我们尝试从同一运行网络的Cisco边界路由器提供的2 048抽样条件下(根据Cisco提供的文档,该抽样为报文层系统抽样,抽样比为1/2048)的Netflow数据中观测这个现象,结果也多次发现了类似的攻击.图10是根据Netflow流数据检测到的2013年5月1日2:00~17:15发生的DNS反向散射事件.将一天划分为288个5分钟时间片,图中横轴表示这288个时间片,纵轴表示每个时间片内根据抽样样本估计后的入方向源端口为53/UDP以及出方向宿端口为53/UDP的每秒平均报文数.

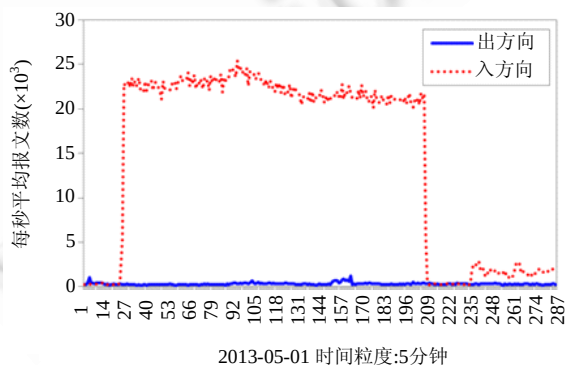


Fig.10 DNS backscatter observed on 2013-05-01

图 10 2013 年 5 月 1 日收到的 DNS 反向散射

由于 Netflow 数据的实时性,ISP 可以很快定位被攻击服务器和网内收到散射的 IP 地址段.此外,因为攻击的持续时间较长,所以在更大范围内,ISP 之间的多方位合作成为可能,这样的合作可进一步获取参与攻击的僵尸网络的情况.

4.5 TCP SYN泛洪DDoS攻击检测

图5和图7显示:2010年~2012年的数据中,TCP SYN+ACK反向散射流量是IBR的主要成分,与other_single类的情况类似,这是由伪造源地址的TCP SYN泛洪DDoS攻击引起的.通过使用如下启发式规则,可以从这部分流量中分析出因特网中发生的DDoS攻击:

(1) 从IBR中过滤出TCP SYN+ACK单包流;

(2) 将过滤得到的流量依据(源IP地址、源端口号)聚类;

(3) 对数量超过一定阈值的类,考虑相应的(源IP地址、源端口号)受到了伪装源地址的TCP SYN泛洪DDoS攻击.

基于以上规则,我们从2012年数据的反向散射流量中获取了TOP1二元组($Victim_9$,80),其流数占反向散射流的46.46%.在观测期的前12个小时内,它向网内4个不连续的B类地址块发送反向散射,强度约为27315pps.据此可判断:服务器$Victim_9$在观测期内受到强度较大的TCP SYN泛洪DDoS攻击,并且攻击报文伪装的源地址并非随机生成而是有选择性地倾向于伪装成若干个*B*类地址块内的IP地址.此外,基于类似的原理和双向的Netflow流数据,我们还可以发现运行网络内主机被攻击或参与这类攻击的事件.

5 结束语

IBR是因特网中客观存在的无功流量,用暗网捕获的IBR流量可以用于网络威胁检测等实际应用中.然而,暗网需要较难满足的布置条件,并存在被恶意攻击回避从而失去使用价值的可能.因此,本文提出了一种从运行网络的原始流量中抽取出IBR流量的方法,该算法既适用于报文数据,又能够用于流数据.与现有的基于运行网络中不活跃地址的IBR流量获取算法相比,本文算法能同时获取活跃地址以及不活跃地址收到的IBR流量,因而具有更低的漏判率.基于单向流的方法虽拥有100%的查全率,但却是以降低查准率为代价.基于基准数据的验证显示,通过增加行为学习这一环节,虽然本文算法查全率有所降低,查准率却从约93%提高到99%以上.总的来说,本文算法较准确地获取运行网络的IBR流量.通过将算法应用到一个拥有约128万个IP地址的运行网络中,并获取了该运行网络的实测IBR数据.通过对这些获取的IBR流量进行分析,我们获得了该运行网络中IBR流量的有关特征,其中,关于空间特性的讨论是相关类型论文中首次提出.分析显示:近两年,数据中的IBR流比例高达70%以上,该发现与相关文献的结论相一致.在如此高的IBR流比例下,基于流的流量识别工作应该也必须考虑到这类流量(IBR)的存在.通过对流量的进一步分析,从中发现了一些网络安全事件.基于这些安全事件呈现出的特征,我们已经可以从较高抽样比的Netflow数据中实时检测到大量的类似事件.

References:

[1] Wustrow E, Karir M, Bailey M, Jahanian F, Huston G. Internet background radiation revisited. In: Proc. of the 10th ACM Conf. on Internet Measurement (IMC 2010). New York: ACM Press, 2010. 62−74. [doi: 10.1145/1879141.1879149]

[2] Dainotti A, Amman R, Aben E, Claffy K. Extracting benefit from harm: Using malware pollution to analyze the impact of political and geophysical events on the Internet. ACM SIGCOMM Computer Communication Review (CCR), 2012,42(1):31−39. [doi: 10.1145/2096149.2096154]

[3] Pang R, Yegneswaran V, Barford P, Paxson V, Peterson L. Characteristics of Internet background radiation. In: Proc. of the 4th ACM SIGCOMM Conf. on Internet Measurement (IMC 2004). New York: ACM Press, 2004. 27−40. [doi: 10.1145/1028788.1028794]

[4] Moore D, Shannon C, Voelker GM, Savage S. Network telescopes: Technical Report. In: Proc. of the Cooperative Association for Internet Data Analysis. 2004.

[5] Bailey M, Cooke E, Jahanian F, Nazario J, Watson D. The Internet motion sensor: A distributed blackhole monitoring system. In: Proc. of the 12th Annual Network and Distributed System Security Symp. (NDSS 2005). 2005.

[6] Yegneswaran V, Barford P, Plonka D. On the design and use of Internet sinks for network abuse monitoring. In: Proc. of the Symp. on Recent Advances in Intrusion Detection (RAID 2004). Berlin, Heidelberg: Springer-Verlag, 2004. 146−165. [doi: 10.1007/978-3-540-30143-1_8]

[7] Team cymru darknet project. http://www.team-cymru.org/Services/darknets.html

[8] Moore D, Voelker GM, Savage S. Inferring Internet denial-of-service activity. In: Proc. of the 10th Conf. on USENIX Security Symp. (SSYM 2001). USENIX Association Berkeley, 2001.

[9] Moore D, Paxson V, Savage S, Shannon C, Staniford S, Weaver N. Inside the slammer worm. IEEE Security & Privacy, 2003,1(4): 33−39. [doi: 10.1109/MSECP.2003.1219056]

[10] CAIDA. http://www.caida.org/research/security/

[11] Dainotti A, Squarcella C, Aben E, Claffy KC, Chiesa M, Russo M, Pescap A. Analysis of country-wide Internet outages caused by censorship. In: Proc. of the 11th ACM Conf. on Internet Measurement (IMC 2011). New York: ACM Press, 2011. 1−18. [doi: 10.1145/2068816.2068818]

[12] Benson K, Dainotti A, Claffy KC, Aben E. Gaining insight into AS-level outages through analysis of Internet background radiation. In: Proc. of the 2012 ACM Conf. on CoNEXT Student Workshop, New York: ACM Press, 2012. 63−64. [doi: 10.1145/2413247.2413285]

[13] Bailey M, Cooke E, Watson D, Jahanian F, Nazario J. The blaster worm: Then and now. IEEE Security & Privacy, 2005,3(4):26−31. [doi: 10.1109/MSP.2005.106]

[14] Glatz E, Dimitropoulos X. Classifying Internet one-way traffic. In: Proc. of the 12th ACM Conf. on Internet Measurement (IMC 2012). New York: ACM Press, 2012. 37−50. [doi: 10.1145/2398776.2398781]

[15] Sherwood R, Gibb G, Yap KK, Appenzeller G, Casado M, Mckeown N, Parulrar G. Can the production network be the testbed? In: Proc. of the 9th USENIX Conf. on Operating Systems Design and Implementation. USENIX Association Berkeley, 2010. 1−6.

[16] Harrop W, Armitage G. Greynets: A definition and evaluation of sparsely populated darknets. In: Proc. of the 2005 ACM SIGCOMM Workshop on Mining Network Data. New York: ACM Press, 2005. 171−172. [doi: 10.1145/1080173.1080177]

[17] Harrop W, Armitage G. Defining and evaluating greynets (sparse darknets). In: Proc. of the IEEE Conf. on Local Computer Networks 30th Anniversary (LCN 2005). 2005. 344−350. [doi: 10.1109/LCN.2005.46]

[18] Jin Y, Simon G, Xu K, Zhang ZL, Kumar V. Gray's anatomy: Dissecting scanning activities using IP grey space analysis. In: Proc. of the 2nd Workshop on Tackling Computer Systems Problems with Machine Learning Techniques (SysML 2007). 2007.

[19] Jin Y, Zhang ZL, Xu K, Cao F, Sahu S. Identifying and tracking suspicious activities through IP gray space analysis. In: Proc. of the 3rd Annual ACM Workshop on Mining Network Data. 2007. [doi: 10.1145/1269880.1269883]

[20] Miao LH, Ding W, Zhu HT. Extracting Internet background radiation from raw traffic using greynet. In: Proc. of the 18th IEEE Int'l Conf. on Networks (ICON). 2012. 370−375. [doi: 10.1109/ICON.2012.6506586]

[21] DUST 2012. http://www.caida.org/workshops/dust/1205/

[22] Gong H, Miao LH, Ding W. Statistics and analysis of IBR flows in Jiangsu CERNET. In: Proc. of the 2012 Annual Conf. of CERNET. 2012. 59−62 (in Chinese with English abstract).

[23] Hoekzema K. Background radiation on today's Internet with a focus on used IP addresses. In: Proc. of the 8th Twente Student Conf. on IT. 2008.

[24] Brownlee N. One-Way traffic monitoring with iatmon. In: Proc. of the 13th Int'l Conf. on Passive and Active Measurement. 2012. 179−188. [doi: 10.1007/978-3-642-28537-0_18]

[25] Bailey M, Cooke E, Watson D, Jahanian F, Provos N. Practical darknet measurement. In: Proc. of the 40th Annual Conf. on Information Sciences and Systems (CISS). 2006. 1496−1501. [doi: 10.1109/CISS.2006.286376]

[26] Zhang XQ, Gu CH. A method to extract network intrusion detection feature. Journal of South China University of Technology (Natural Science Edition), 2010,(1):81−86 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-565X.2010.01.016]

[27] Kim HC, Claffy KC, Fomenkov M. Internet traffic classification demystified: Myths, caveats, and the best practices. In: Proc. of the 2008 ACM CoNEXT Conf. New York: ACM Press, 2008. [doi: 10.1145/1544012.1544023]

[28] IP trace distribution system. http://iptas.edu.cn/src/system.php

[29] Zhang YB, Zhang ZB, Zhao Y, Guo L. Comparative analysis on TCP and UDP network traffic. Application Research of Computers, 2010,27(6):2192−2197 (in Chinese with English abstract). [doi: 10.3969/j.issn.1001-3695.2010.06.056]

附中文参考文献:

[22] 龚皓,缪丽华,丁伟.CERNET 江苏省网边界辐射流量的获取与分析.见:CERNET2012 年会.2012.59−62.

[26] 张雪芹,顾春华.一种网络入侵检测特征提取方法.华南理工大学学报(自然科学版),2010,(1):81−86. [doi: 10.3969/j.issn.1000-565X.2010.01.016]

[29] 张艺濒,张志斌,赵咏,郭莉.TCP 与 UDP 网络流量对比分析研究.计算机应用研究,2010,27(6):2192−2197. [doi: 10.3969/j.issn.1001-3695.2010.06.056]

缪丽华(1987－),女,江苏南通人,博士生,主要研究领域为网络测量,网络行为学,网络安全.

丁伟(1962－),女,博士,教授,博士生导师,主要研究领域为网络测量,网络行为学.

杨望(1979－),男,博士,讲师,CCF 会员,主要研究领域为网络安全.