CERNET 江苏省网边界辐射流量的获取与分析

龚皓，缪丽华，丁伟

东南大学计算机科学与工程学院，南京 211189

摘 要：本文首先介绍了 IBR 流量的基本原理以及产生 IBR 流量的主要原因，并根据 IBR 流量的特点，设计和实现了一种从接入网边界的原始流量中分离出 IBR 流量的算法。通过将该算法应用于从 2008 年到 2011 年四年的实测流量中，本文展现了 IBR 流量在时间和空间上的特性。时间上，IBR 流量的总量与在总流量中所占比例皆呈逐年增长的趋势，且主要成分从 TCP SYN 单包流转变为 TCP SYN+ACK 单包流。空间上，IBR 流量则呈现了普遍性和不均匀性。

关键词：网络背景辐射，网络测量，扫描，反向散射，灰空间

中图分类号：TP393.0　　**文献标识码：**A

0．引言

背景辐射流量(IBR: **Internet Background Radiation)**[1]是网络中的一种非正常的流量，多是由 TCP SYN+ACK 单包流、TCP SYN 单包流等特殊的网络流量构成的，并且这些单包流在网络中有很高的出现频次。IBR 流量产生的原因是不一而同、复杂的，例如黑客的扫描行为、DDOS 攻击的反射流量、错误的路由配置等[2]。IBR 流量的捕获对于发现和追踪网络威胁[3]、解决安全隐患具有重要的价值。

网络中的暗地址，是指全球可路由但未分配给活跃主机的地址[2]。发往暗地址的流量被认为是不正常的流量，属于 IBR 流量。监测暗地址是获取 IBR 流量并进行后续分析的有效方法。但是，布置暗网需要具有一定规模且连续的 IP 地址块，对于普通的接入网这并不是容易满足的条件。本文从普通接入网的角度，提出了一种从原始流量中分离 IBR 流量的方法。实验结果表明，该方法是可行且有效的。

1．IBR 流量的获取

在 TCP/IP 协议中，IP 层在转发报文时仅依赖目的地址，而不检查源地址。通信对端则总是假设 IP 报文的源地址是真实的，这使得其在回复报文时，有可能使用的是错误的目的地址，这是产生网络 IBR 流量的原因之一。伪造源地址的活动，会使得网络流量具有不对称性（如图 1 所示）。基于这种不对称性，我们就可以从原始流量中分离出 IBR 流量。

要分离出 IBR 流量，首先就要定义出灰地址空间。设 g 为接入网内部的一个 IP 地址，若其在时间段 T 中，不发送流量、仅接收流量，则称 g 为灰地址。此外，在实际网络中存在一部分地址并没有实际的出流量，而是由接入网的边界防火墙代替做出应答的情况，则实际上该类地址并不存在实际活跃的行为。即，设 r 为接入网内部的一个 IP 地址，若其在时间段 T 内，仅发送 TCP RST+ACK 报文，将此类地址也纳入灰空间的范围。

如图 1 所示：

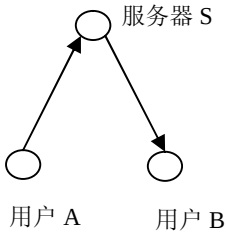

图 1 用户 A 使用伪源地址情况

由于灰空间中并没有活跃用户，因此称向灰空间发送报文的 IP 为可疑 IP 地址。称所有可疑 IP 地址组成的空间为可疑空间。可疑 IP 地址发送的流量并不全都是辐射流量，其中存在着部分正常交互的流量，

收稿日期：2012-8-30

作者简介：龚皓（1988-），男，广东兴宁人，硕士研究生，主要从事网络测量与网络行为学的研究.(TEL)15850607069

如图 2 示。称可疑空间发出的单向流为 IBR 流量。即，判定可疑 IP 发送的单向流为 IBR 流量，双向流则为正常流量。

称使用被动测量方法从 CERNET 江苏省网边界捕获的原始流量为 Trace[in]和 Trace[out]，则从 Trace 中分离出 IBR 流量

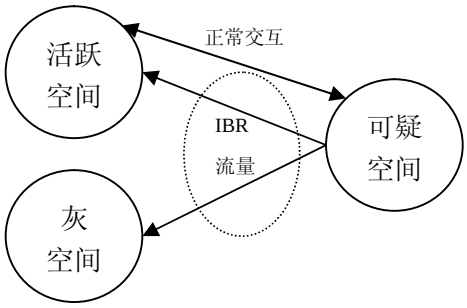

图 2　可疑空间发送流量的组成

的算法 F 可归结为以下 4 步。其中，Trace[in]指流入接入网的报文头集合，Trace[out]指流出接入网的报文头集合。由于并未拥有接入 CERNET 江苏省网的所有 IP 地址表，本算法使用近似的方法，统计该 IP 地址表所拥有的地址空间。

(1) 对于报文记录集合 Trace[in]和 Trace[out]，根据通用流特征，组成五元组流（64 秒超时），得到 Flow[i]和 Flow[out]。

(2) 统计在 Flow[out]的源地址中未出现，而在 Flow[i]的宿地址中出现的地址，以此获得灰空间。

(3) 将 Flow[in]中向灰空间发送流量的源 IP 地址判为可疑地址，归入可疑 IP 地址空间中。

(4) 过滤出 Flow[in]中所有以可疑空间为源点的单向流，从 Flow[out]中过滤出所有以可疑空间为宿点的单向流，匹配两个方向的流记录。若成功匹配，则生成双向流。若不能匹配，统计其中可疑空间发送的单向流，则其很有可能即为 IBR 流量。

即：IBR=F(Trace[in,out])

算法输入：报文记录集合 Trace[in]和 Trace[out]

算法输出：背景辐射流量 IBR。

2．实验结果

本文使用的接入网流量来自 CERNET 江苏省网边界[4]。该省网的 IP 地址总量接近 5000 个/24 地址。该接入网的管理者，每年不定期以 1/4 流抽样的方式从接入点采集报文头，并以 IP Trace 的方式储存。2008 年到 2011 年四年，每年挑选一个小时的 Trace 作为本文的实验数据，见表 1。

表 1　实验数据集

Trace ID	IP count	Date	Timejavascript:void(0);
A	171226	2008-12-20	14:00:00 ~ 15:00:00
B	210915	2009-12-20	14:00:00 ~ 15:00:00
C	246549	2010-11-14	14:00:00 ~ 15:00:00
D	295407	2011-11-17	14:00:00 ~ 15:00:00

如表 1 所示，IP 地址的数量是动态变化的。如果仅从总量上分析，并不能实际反映变化情况。令 IP_i={第 i 年实际出现的地址}。实际中我们只统计 4 年共同出现的地址，即 $IP_{com}=IP_{2008}\cap IP_{2009}\cap IP_{2010}\cap IP_{2011}$ 作为实验对象进行统计和分析。经实验统计得，共同出现的地址数为：$|IP_{com}|=166021$。

以流数为单位统计的 IP_{com} 在这 4 年中的背景辐射情况如图 3 所示（本文以下的所有 IBR 流量均指流数）。我们可以发现在 2008 年到 2009 年，IBR 流量基本无变化，但从 2010 年开始，IBR 流量出现了剧烈的增长，相对于 2009 年，增长了约 4 倍，并在 2011 年继续增长了约 3.4 倍。这反映出近年来，IBR 流量出现迅速增长的趋势。并按流数计算，IBR 流量于总流量所占比例，从 20%左右，上升到 50%以上。

图 3　IBR 总流量变化趋势

3．实验结果分析

3.1 地址块间的 IBR 分析

由于 IP_i 地址过多，我们将 IP_i 内的地址做如下处理 $ip_{new}=ip_{old}\&0xfffff000$，即将前 20 位相同的 IP 地址归入同一个统计单元。我们计算一个地址块内，每个存在的 IP 地址平均所受到的辐射流量。使用平均辐射流量可以避免 $ip_{new}[i]$ 与 $ip_{new}[j]$ 在 IP_{com} 中可能存在的地址数不相同的情况对分析产生的影响。同时，为保证隐私，我们使用一个映射函数 f_{map} 将 ip_{new} 在不更改顺序的情况下，映射为新的数字：

$Num=f_{map}(IP)$

实验结果如图 4 所示。我们可以发现，不存在未收到 IBR 流量的地址块，而对于不同的 ip_{new} 地址块，所受的 IBR 流量差别是很大的。平均受到 IBR 流量最多的是 x.x.176.0/20 地址块，每个 IP 地址平均共受到 2998 个背景辐射流。而最少的为 x.x.0.0/20 地址块，平均只有 3 个背景辐射流。

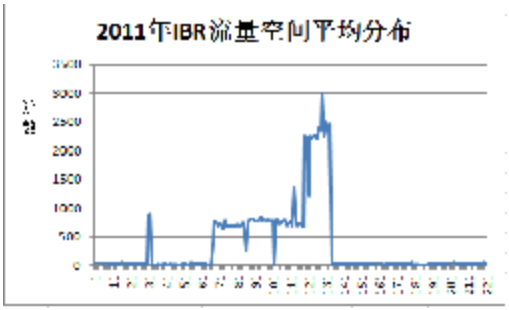

图 4　2011 年 IBR 流量空间平均分布

其中，Y 轴表示所受到的平均背景辐射流数。图 4 X 轴表示，/20 地址经过 f_{map} 变换后得到的分布。

图 5 是 2008 年到 2011 年四年间共同出现的 IP 地址块的中每个 IP 地址平均所受到的背景辐射流数与该年数据中，所受到平均辐射最大的地址块的比值，即将每个地址块的平均所受辐射数按照最大地址块所受辐射值进行归一化。由图 5 可以看出，IBR 流量的空间偏好是存在变化的，任意/20 地址块所受平均辐射流量比例在这四年间是不相同的。同时，IBR 流量也存在一定的聚集性，从 2010 和 2011 年的数据明显看出，从 $f_{map}(IP)$=65-135 这些/20 地址块，接受到较高的 IBR 流量。

图 5 中，Y 轴表示所受到平均辐射与最大平均辐射的比值。X 轴表示，/20 地址经过 f_{map} 变换后得到的分布。

图 5　2008-2011 年 IBR 流量空间分布比例

3.2 IBR 成分

本实验所有统计分析皆基于流。将 IBR 流量根据协议、TCP 标记字段分为若干类，如 TCP SYN 单包流，TCP SYN+ACK 单包流，UDP 单包流，ICMP 流等。则总量排在前三位的成分如表 2 所示。

表 2　IBR 流量组成成分比例

Trace	2008	2009	2010	2011
SYN+ACK (%)	3.46	6.43	77.84	84.07
SYN (%)	78.61	75.26	15.97	6.65
UDP 单包(%)	4.5	5.88	1.99	8.41
SUM（%）	86.57	87.57	95.8	99.13

从表 2 可得，TCP SYN 单包流和 TCP SYN+ACK 单包流是构成 IBR 流量的主要组成部分。而显然，TCP SYN+ACK 单包流和 TCP SYN 单包流不是正常的 TCP 流量。在 2008 年和 2009 年，IBR 流量的主要组成部分为 SYN 单包流，而在 2010 年和 2011 年，主要组成部分转变为 SYN+ACK 单包流。我们猜测，在 2008 年和 2009 年，背景辐射的主要成分为扫描流量；从 2010 年开始，背景辐射的主要成分转变为反向散射流量。这可能意味着从 2010 年开始，DDoS 攻击开始盛行。

对 2011 年数据中单包流进行分析，发现发送 SYN+ACK 量最大的一个 IP 地址为 x.x.237.32。这说明在观测时间 T 内，该网站受到了一定规模的 DDoS 攻击，并且此网站已被某著名搜索引擎列入黑名单

中，说明该网站确实存在异常行为。本文算法有效滤出上述流量是该攻击的反向散射流量，属于是典型的 IBR 流量。产生的主要原因是因为在攻击服务器时，攻击源使用的是假源地址的 TCP SYN 报文，以此来消耗服务器资源。服务器在收到 TCP SYN 报文时，向假源地址回复 TCP SYN+ACK 报文。

4.总结

本文提出了一种从接入网边界的原始流量中分离 IBR 流量的方法，并将该方法应用于 CERNET 某省网边界的原始流量。通过对 2008 年到 2011 年四年抽样的 Trace 进行分析可知，IBR 流量从 2010 年开始迅速增加，且主要成分由扫描流量转变为反向散射流量。通过对省网内地址的统计分析可知，IBR 流量具有普遍性，即每个/20 地址块均收到辐射流量；同时又具有不均匀性，即一部分地址块所受辐射明显多于其他地址块。

同时，通过本文的分析，可知单包流等非正常流量在网络中占有很高比例，这对于基于流的路由策略也有很大的不良影响。

参考文献：

[1] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson. Characteristics of Internet Background Radiation. In Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement(IMC'04), Oct 2004.

[2] E.Wustrow, M. Karir, M. Bailey, F. Jahanian, and G. Huston. Internet Background Radiation Revisited. In Proceedings of the 10th annual Conference on Internet Measurement(IMC'10), ACM, 2010.

[3] E. Cooke, M. Bailey, Z.M.Mao, D.Watson and F. Jahanian. Toward Understanding Distributed Blackhole Placement. In Proceedings of ACM CCS Workshop on Rapid Malcode, pp. 54-64. ACM Press, October 2004.

[4] IP Trace Distribution System. http://iptas.edu.cn/src/system.php

Statistics and Analysis of IBR Flows in Jiangsu CERNET

Gong Hao, Miao Lihua, Ding Wei

Department of Computer Science and Engineering, Southeast University, Nanjing 210018

Abstract: This paper describes the basic principles of Internet Background Radiation(IBR) and the main reasons that generate IBR . Based on the characteristic of IBR , the paper introduces how to extract IBR out of raw traffic captured from the boundaries of access networks. By applying the method to traffic captured from 2008 to 2011, this paper shows the temporal and spatial characteristics of IBR. From the temporal perspective, it shows a growing trend in IBR's volume, and IBR's composition also changes from TCP SYN single packet flow to TCP SYN+ACK single packet flow. Spatially, IBR is ubiquitous and uneven.

Keywords: IBR, Network Measurement, Scan, Backscatter, Gray Space