基于流记录的主干网活跃 IP 地址空间检测

张凌峰 丁伟 龚俭 缪丽华

（东南大学 计算机科学与工程学院， 南京 211189）

摘要 掌握 IP 地址的实际使用情况对于网络管理和网络安全等研究领域有着重要的意义。本文提出一种以抽样流记录为分析数据源的活跃地址检测算法，其核心思路是将存在双向通信流量作为地址活跃判定条件。算法基于被动测量技术，以流记录为分析数据源使得其可以在主干网边界运行。文章讨论了抽样、伪造地址等问题对算法的影响以及相应的应对策略，用 DPI 分析检验了算法的准确性和有效性。最后基于 NBOS 平台，将其部署在 CERNET 全部 38 个主节点，完成了全网活跃 IP 地址空间的检测。

关键词 IP 地址活跃性；流记录；抽样；双向通信流量

Backbone Active IP Address Space Detection Based on Flow Records

ZHANG Ling-Feng DING Wei GONG Jian MIAO Li-Hua

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

Abstract Understanding the utilization of IP addresses is much important for network administrators and network security research etc.. This paper proposes a methodology of detecting active IP addresses based on sampled flow records. The core idea of the methodology is that IP addresses with two-way communication traffic are active. The method is based on passive measurements and uses sampled flow records as data source, making it possible to be deployed at the boundary of backbones. Furthermore, we discuss the impacts of flows' sampling and spoofed traffic on the method. DPI technology is used to validate accuracy and efficiency of the method. Finally, the method is deployed at all 38 nodes of CERNET, detecting active IP address space in the whole CERENT network.

Keywords IP address activity; flow records; sampling; two-way traffic

1 引言

目前，虽然 IPv4 地址池已经被完全用尽，但已分配的地址并没有得到充分利用[1]。已经分配并正常使用的地址称为活跃地址。IP 地址的活跃性测量工作有利于 ISP 掌握当前 IP 地址的使用情况和管理方式。地址活跃性还可以用于网络安全和网络管理领域相关的研究工作。基于不活跃地址空间构成的灰网，可以获取 IBR 流量[3,4]，这是网络安全领域非常珍贵的数据源，可以用于网络威胁的检测，如蠕虫、扫描或 DDoS 攻击等，发现最新出现的导致网络安全事件的因素。另外，不活跃地址空间内没有合法的主机或设备，这些地址产生的流量都是伪造地址流量，在 SDN 等新技术的支持下，可以对其进行过滤，从而使得网络带宽可以得到更有效的利用。

IP 地址的活跃性可以用于地址管理和网络安全等领域，相关的研究也已经开展多年，测量方法主要有主动探测和被动测量两种。最早的研究是 ISI 的网络普查项目[5]，采用主动探测方法，周期性向每个 IPv4 地址（除私有地址和组播地址外）发送 ICMP echo request 报文来跟踪地址活跃性。经过一个月的探测，他们共发现将近 4.3M /24 活跃地址块。这类方法虽然简单易行，但存在以下缺点：①探测会产生额外的网络流量；②一些网络的边界防火墙

可能过滤或代为回复这些探测报文，因此获取的信息可能不准确。被动测量方法可以避免这些不足，但也需应对以下两个方面的挑战：①观测点视野的限制，②伪造源地址造成的误判问题。Alberto Dainotti 和 Karyn Benson 等人[1]采用一种启发式算法过滤 IBR 流量和未抽样的 NetFlow 流数据中的伪造源地址流量，统计了可见地址空间的地址活跃性。IBR 流量的获取有一定困难，使用未抽样流记录作为分析数据源也需要占用大量的网络资源，在目前阶段不适合采用这种方法。另外一些研究[6,7]使用基于 TTL 的推断来检测伪造源地址报文，尝试为不同流量类别建立 TTL 参考值，将 TTL 值偏离较大的报文判定为伪造源地址报文。基于 TTL 的方法需要完整的 IP 报文信息，因此也不可能在大规模范围内实时展开。

综上所述，用被动测量方法获取活跃 IP 空间所面临的问题在于数据源获取和实施规模的限制。被动测量需要大量的资源，如果数据源过于庞大，则无法在更大规模上展开。随着以 NetFlow 为代表的流技术逐步走向成熟，目前 Cisco 和华为等设备商提供的主干路由器均可以提供抽样的流记录数据，尝试用这样的数据源进行大规模接入网的地址活跃性分析，是本论文希望解决的问题。在对环境和定义进行描述和刻画的基础上，本文提出了一个基于抽样流记录的活跃地址检测算法，在讨论了抽样、伪造地址等问题对算法的影响后，基于实测数据验证了算法的有效性。

2 基于流记录的活跃地址检测算法

定义 1. 活跃地址

已分配并正常使用的地址称为活跃地址[1]。

定义 1 对企业网内 IP 地址的直接管理者而言是可操作的，但在大规模的接入网边界，这样的定义是不适用的。

互联网上的流量都是由应用程序产生的，从这个角度出发寻找这些正常使用的地址的流量行为特征，其中之一是都会有双向的通信流量。大部分应用都是基于 TCP 和 UDP 协议的，TCP 是面向连接的，其应用流量都是双向的；UDP 虽是无连接的，但除了组播应用外，目前已知的所有基于 UDP 的应用也都会产生双向流量，而产生单向流量的组播地址使用固定的地址段。所以，正常的活跃地址（除组播地址外）都会产生双向通信流量。基于该流量特征，本文把在一定时间长度 T 内存在双向有效通信流量作为地址活跃条件，T 称为活跃性检测窗口长度。

定义 2. 活跃地址判定条件

任意 IP 地址 a, 若 a 在 T 时间内有双向有效流量，则判定 a 是活跃地址。

定义 2 相对于定义 1 而言是可检测的。由于需要观测双向流量，那么观测位置应该在网络边界处。

2.1 影响算法准确性的因素

基于流记录的活跃地址检测算法的准确性受到两个因素的影响：一个是抽样，另一个是伪造源地址流量。

2.1.1 流记录抽样损失分析

实际测量环境中，路由器输出的流记录一般都是抽样的。抽样会造成流的截断甚至消失，会对地址活跃性造成误判。对于活跃地址，如果在检测时间内没有被抽到任何报文或者只抽到单一方向的报文，就会被误判为不活跃地址。对此，可以通过对时间和空间的扩展来弥补抽样的损失。在时间方面，可以用适当放宽活跃性检测窗口长度 T 的方法来提高活跃地址双向流量均被抽到的概率；在空间方面，由于临近地址的使用情况相似[8]，可以采用聚合地址空间的方法，将地址块作为活跃空间统计单位，记地址块前缀长度为 m。如果/m 地址块中有一个或多个地址检测为活跃，该地址块中的所有地址将会判定为活跃地址。这样的处理可能会将不活跃地址误判为活跃，从不活跃地址空间角度，可以更好地保障查准率。

2.1.2 伪造源地址流量对算法的影响

网络攻击者为了隐匿自身地址，通常会伪造源地址发起攻击，这样的流量也会影响地址活跃性检测。如果被伪造的是网内不活跃地址，且被攻击的目标产生回复，就可能会观测到被伪造地址的双向流量，简单地基于定义 2，就会将其误判为活跃地址。所以必须尽可能的鉴别出伪造源地址流量并清除。文献[1]对伪造源地址流量的过滤规则是要求检测时间内 TCP 连接的报文数和报文平均尺寸高于阈值，而对于大规模接入网络，组双向流并计算流量将耗费较多资源。经分析，需反馈的 TCP 伪造源地址攻击通常为 SYN Flood 攻击，因此可以根据流的 tcp flags 过滤流量。通过过滤 TCP SYN 流、SYN 置位的单包流，尽可能清除伪造源地址流量。另外，为了避免异常 FIN 及 RST 攻击，算法也过滤掉 FIN

或 RST 置位的单包流。基于 UDP 的伪造源地址双向流量，很难总结出规则来鉴定和过滤。定义 3 是本算法对源伪造流量的过滤规则，未被过滤掉的流称为有效流。

定义 3. 过滤源伪造流量的启发式规则

(1) 过滤 TCP 和 UDP 协议以外的流。

(2) 过滤 TCP SYN 流，TCP SYN 、FIN、RST 置位的单包流。

2.2 算法描述

基于以上的分析讨论，本小节给出一个基于抽样流记录活跃 IP 的检测算法，该算法可以在任何规模的接入网边界上，检测该接入网内地址的活跃情况。

输入：流记录，接入网的地址空间 *adrSet*，地址块前缀长度 *m*，检测周期长度 *T*。

输出：活跃地址空间 *activeSet*（长度为/*m* 地址块）

(1) 每个检测周期开始时，重置 *adrSet* 中所有地址的流量信息（接收和发送的流量情况等)；

(2) 在检测周期内，依次读取未处理的单方向流记录，按照定义 3 过滤，对于有效流：

① 若为入方向，则更新宿地址收到的流量信息。如果该地址首次出现，则将其类型设置为“待判定地址”；否则，如果该地址第一次变为活跃地址（第一次满足定义 2），则设置为“活跃地址”类型并更新 *activeSet*。

② 若为出方向，则更新源地址发送的流量信息。处理方式与入方向类似。

(3) 在检测周期结束时，按照/*m* 地址块输出 *activeSet*。

3 基于流记录活跃地址检测算法实现

基于流记录的活跃地址检测算法是要部署到大规模网络中进行实际运行的，因此以下几个问题需要讨论：算法的实现平台，检测窗口滑动的解决方案，算法重要参数的确定以及算法的正确性验证。

3.1 算法实现平台

本算法将基于 NBOS 系统平台[13]实现。NBOS(Network Behavior Observation System)是由 CERNET 华东(北)地区网络中心在国家支撑计划支持下研发的网络行为观测系统。它基于主干网和地区接入网互联节点的流记录数据实现对网络流量行为的观测与精细化管理。目前 NBOS 已在 CERNET 全部 38 个主节点部署，并在实际网络环境中运行超过 1 年。NBOS 系统的预处理平台根据流终止时间，以 5 分钟为时间粒度对来自接入路由器的流记录进行整理后输出，整理工作包括合并 5 元组相同的流、处理乱序情况等。时间粒度长度记为 $T_0 = 300$ 秒。为充分借助 NBOS 平台，我们将检测窗口长度 *T* 设置为 T_0 的整数倍，即 $T = n * T_0$。地址活跃性是根据当前时刻之前 *T* 时间内的流量来判定的，那么，面对时间粒度的活跃地址定义为最近连续 *n* 个时间粒度内存在双向通信流量的地址。

3.2 检测窗口滑动的解决方案

由于检测窗口包含 *n* 个时间粒度，对于任意 IP 地址，我们分别用 *n* 个布尔变量来表示地址的出、入方向流量情况。如图 1 所示，记为 $In_0 \sim In_{n-1}$ 和 $Out_0 \sim Out_{n-1}$，初始置为 false。

In_0	In_1	...	In_{n-1}
Out_0	Out_1	...	Out_{n-1}

图 1　地址在检测窗口内的流量表示

记时间粒度为 G_j，粒度数 *j* 从 0 记起。初始检测窗口包含时间粒度 $G_0 \sim G_{n-1}$。在第 *i* 个时间粒度内($i \in [0 \sim n-1]$)，依次读取未处理的单方向流记录，按照定义 3 的启发式规则进行过滤，过滤后的有效流做如下处理：

(1) 若为入方向流，关注宿地址的 In_i 变量，若值为 false，置为 true。

(2) 若为出方向流，关注源地址的 Out_i 变量，若值为 false，置为 true。

当表达式(1)为真时，地址判定为活跃：

$$(In_0 \| ... \| In_{n-1}) \&\& (Out_0 \| ... \| Out_{n-1}) \quad (1)$$

地址的活跃性需要长期监测，时间粒度将会不断向后推移，检测窗口也会随之向后滑动。由于算法基于 NBOS 平台实现，所以设检测窗口每次以 T_0 长度不断向后滑动。如图 2 所示，检测窗口初始时包括时间粒度 $G_0 \sim G_{n-1}$，相应流量存储在 $In_0 \sim In_{n-1}$ 和 $Out_0 \sim Out_{n-1}$ 变量中。当时间粒度推移到 G_n 时，窗口位置滑动到 $G_1 \sim G_n$。对此，将新包含的 G_n 粒度流量存储在 In_0 和 Out_0 变量中，保证始终以一组变量保存地址当前检测窗口的流量，以此类推。图 2 中以入方向变量为例说明变量与粒度的对应关系。

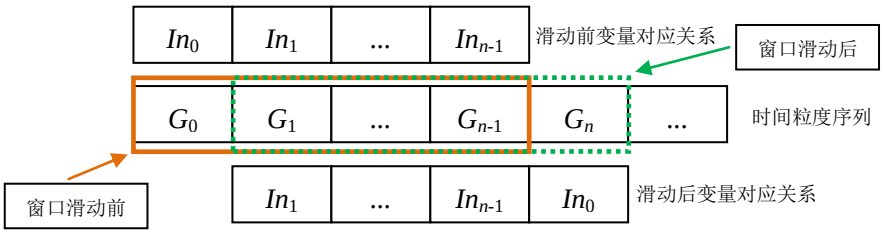

图 2 检测窗口滑动示意图

综上所述，设当前时间粒度数为 p，该粒度对应的流量变量为 In_i 和 Out_i，那么 p 与 i 的对应关系如公式(2)所示：

$$i = p\%n \qquad (2)$$

在这样的方法下，设待判定地址空间共有 c 个 IP 地址，本算法只需要 $2*c$ 个布尔型变量即可保存和计算地址活跃性。同时，利用时间粒度的周期性，每个时间粒度的流记录可根据公式 2 直接定位到相应地址的某个流量变量中。设一个时间粒度的流记录数为 f，那么本算法在该粒度内的时间复杂度仅为 O(f)。

3.3 算法重要参数的确定

算法实现中有两个重要参数 n 和 m，n 是 T 时间内包含的时间粒度数，m 是地址块前缀长度，这两个参数在实践中必须设定取值。为尽量减少占用的资源，在获取相同活跃地址空间的情况下，n 取值越小越好。对于参数 m，若 m 值过小，即地址块粒度过大，会增加活跃地址的误判可能性；相反若 m 过大，地址块粒度过小，会增加活跃地址的漏判率。为使得误差降到最低，我们选择与相邻较大地址块粒度的结果相差最小的粒度作为缺省值。两个参数的合适取值将通过具体实验确定。实验的具体方案如下：

Step1：将观测点设在 CERNET 江苏省网边界处，该网大约有 6000 个/24 地址段。选取某 985 高校一段/16 地址块作为实验对象。

Step2：分别选取 n = { 6,12,18 }，分别对应检测窗口长度 T = 0.5h、1h、1.5h，同时运行算法，基于/32 记录详细地址活跃性。

结束后，对参数的每个取值获得的活跃地址空间，分别统计面向/24、/28 和/32 三个粒度的活跃地址块数。观察不同地址块粒度下的结果，当活跃地址数随着 n 的增长不再变化或变化较小时，趋于稳定的最小值即为 n 的缺省值。

Step3：在确定参数 n 后，将算法获得的面向 IP 级别的活跃地址空间按 m = { 24 ~ 32 }粒度分别统计。计算各/m 粒度下的活跃地址数 $num(m)$，并统计各粒度与相邻较大粒度的活跃地址数差异 $diff(m)$，将 $diff$ 值最小的 m 作为缺省值。$diff(m)$的计算见公式(3)：

$$diff\,(m) = \frac{num(m-1) - num(m)}{num(m-1)} \qquad (3)$$

实验运行持续时间为 2014-12-21~2015-01-17，为期 28 天，共 8064 个时间粒度。参数 n 的实验结果如表 1 所示。

表 1　n 取值对活跃地址空间的影响

n	6	12	18
/24 地址块数	168	168	168
/28 地址块数	2037	2038	2038
/32 地址数	18678	18748	18768

从表 1 中可以看出，对于/24 来说，由于粒度较大，n = 6 即可满足要求；对于/28 的中等粒度，n = 12 能达到稳定；而对于/32，由于粒度太细，活跃空间在现有几个取值下都未达到稳定，但 n = 18 的活跃地址数较 n = 12 增长率较小。综合考虑，可以将参数 n 缺省取值为 12。

对于参数 m，将 n = 12 的算法获得的活跃地址空间按 Step3 统计分析，结果如表 2 所示。

表 2　*m* 取值对活跃地址空间的影响

/m	/24	/25	/26	/27	/28	/29	/30	/31	/32
num(m)	43008	39936	36672	34240	32608	31016	28636	24584	18748
diff(m)	--	7.14%	8.17%	6.63%	4.77%	4.88%	7.67%	14.15%	23.74%

表 2 中可以发现，$m = 28$ 时，*diff* 值最小，为 4.77%。所以可将参数 *m* 缺省取值为 28。通过以上实验和分析，我们为算法提供了两个参数的缺省值，参数 $n = 12$，对应一个小时的检测窗口长度；$m = 28$，活跃地址空间按/28 地址块统计。

3.4 算法的准确性验证

算法的准确性将基于 DPI 分析进行。具体方法如下：

Step1：选取一段较大地址空间，用基于 NBOS 平台实现的上述活跃性检测算法进行活跃性分析，运行充分长的时间，初筛出不活跃地址空间，记为 *inactiveSet*。

Step2：将该不活跃空间交给网络边界采集器进行一段时间的全报文采集和处理，检验 *inactiveSet* 中是否有活跃地址存在，验证算法对不活跃地址空间的准确性。

Step3：由于 DPI 采集与初筛时间不一致，地址的活跃性可能发生变化，由不活跃变成活跃，所以在采集用于 DPI 分析的报文的同时，也对 *inactiveSet* 同步用本文算法再次作基于流的活跃性分析。最后，比对 DPI 分析和基于流分析的统计结果，去除因观测时间不同导致活跃性变化的影响。

DPI 分析中，与本文算法处理方法相似，对每个 IP 地址维护出、入方向流量信息。对采集的每个报文，先按照定义 4 的规则进行过滤，过滤后的有效报文，根据流量方向标记相应地址的出或入方向有流量。出入方向都存在有效流量的地址判定为活跃地址。

定义 4. 报文过滤规则

(1) 过滤除 TCP 和 UDP 协议以外的报文。

(2) 过滤 TCP SYN、FIN、RST 置位的报文。

(3) 过滤无实际负载的报文。

为了保证工作的延续性，我们同样选取了 3.3 小节的实验结果进行验证。算法的两个参数使用的是 3.3 节中的缺省值，初筛时间是 2014-12-21~2015-01-17。初筛得到的不活跃地址空间有 2058 个/28 地址段。验证实验的持续时间为 2015-03-30~2015-04-04，为期 6 天。实验结果中，DPI 分析发现不活跃地址空间中有 16 个/28 地址段被重新检测为活跃地址段；在同时运行的本文算法的检测结果中，这些地址段也都被认定为活跃，说明是由于观测时间不同而导致的活跃性不一致。该实验从工程角度验证了本算法获取的不活跃地址空间的完全查准率。

4 算法在 CERNET 全网的应用

基于 NBOS 平台，将本算法部署在 CERNET 全部 38 个主节点，对全网活跃 IP 地址空间进行检测。所覆盖的 IP 地址空间接近 2 千万个 IP，检测时间为 2015-06-01~2015-06-10，为期 10 天。分别按 CIDR/24 ~ /32 地址块粒度统计所有主节点的活跃地址数，计算 CERNET 全网的活跃地址空间占用情况，结果如图 3 所示。

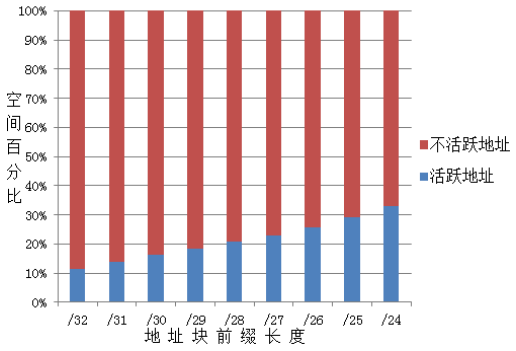

图 3 CERNET 全网地址活跃性分析

从图 3 中可以看出，当地址块前缀长度为 24，活跃地址数最多时，CERNET 全网有 33.02%的地址是活跃地址；当地址块前缀长度为 32 时，即针对 IP 地址进行活跃性分析时，只有 11.32%的地址是活跃的，由此可以认为 CERNET 全网地址的使用率低于 33.02%。

由于 NBOS 对各主节点 IP 范围的信息可能存在偏差，我们进一步给出该信息掌握准确的南京主节点所覆盖的江苏省网的数据。南京节点覆盖的地址范围接近 136 万个，活跃地址空间占用情况如图 4 所示。

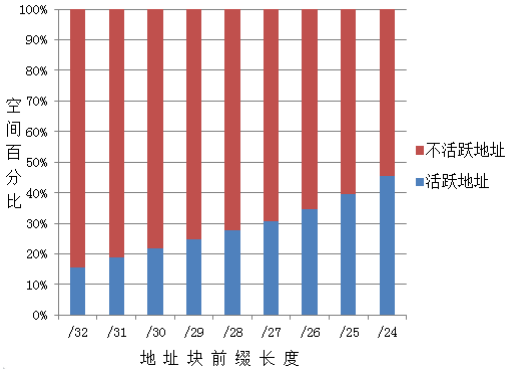

图 4 南京主节点活跃地址分析

图 4 中，当按照/24 的地址块前缀统计时，南京主节点有 45.38%的地址是活跃的；当按照/32 统计时，只有 18.93%的地址是活跃的。因此可以得出的结论是，南京主节点的活跃地址比例在[18.93%, 45.38%]范围内。

5 总结与展望

本文提出了一个工作在大规模接入网边界的基于抽样流记录的活跃地址空间检测算法，并基于 NBOS 系统实现。通过在 CERNET 南京主节点选取某 985 高校的一段/16 地址块进行活跃地址空间检测实验，探讨了算法重要参数的设置，并用 DPI 分析验证了算法的准确性和有效性。最后，基于 NBOS 平台，将本算法部署在 CERNET 全部 38 个主节点，完成全网活跃 IP 地址空间的检测和分析。值得提出的是，本算法一旦判定某个地址活跃，该地址会永久活跃，活跃地址空间会随着观测时间持续扩大。而实际情况下，地址的活跃性是会周期性变化的，因此，后继的工作将从两个角度展开：一是考虑对地址活跃性变化的监督，通过制定活跃地址的老化机制等更加实时、准确地掌握地址活跃性；二是基于概率和统计分析理论对现有算法进行分析和优化，包括抽样对算法的影响、更加准确地过滤伪造源地址流量以提高算法对活跃地址空间的查准率等。

参 考 文 献

[1] A.Dainotti., K.Benson.,A.king.,kcclaffy., M.Kallit-sis., E.Glatz., X.Dimitropoulos. Estimating Internet Address Space Usage Through Passive Measurements[Z]. SIGCOMM CCR, Issue 1, January, 2014, Vol(44): 42-49.

[2] Dainotti A, Benson K, King A, et al. Lost in Space: Improving Inference of IPv4 Address Space Utilization[J]. arXiv preprint arXiv:1410.6858, 2014.

[3] 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.

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

[5] Heidemann J, Pradkin Y, Govindan R, et al. Census and survey of the visible internet[C] //Proceedings of the 8th ACM SIGCOMM conference on Internet measurement.ACM, 2008: 169-182.

[6] Zander S, Andrew L L H, Armitage G, et al. Estimating IPv4 address space usage with capture-recapture [C] // Local Computer Networks Workshops (LCN Workshops), 2013 IEEE 38th Conference on. IEEE, 2013: 1010-1017.

[7] Templeton S J, Levitt K E. Detecting spoofed packets [C]//DARPA Information Survivability Conference and Exposition, 2003. Proceedings. IEEE, 2003, 1: 164-175.

[8] Cai X, Heidemann J. Understanding block-level address usage in the visible Internet[J]. ACM SIGCOMM Computer Communication Review, 2011, 41(4):99-110.

[9] Durumeric Z, Wustrow E, Halderman J A. ZMap: Fast Internet-wide Scanning and Its Security Applications[C]//Usenix Security. 2013: 605-620.

[10] Chien E. Downadup: attempts at smart network scanning[J]. Symantec Security Blogs, Jan, 2009, 23.

[11] Langley A. Probing the viability of TCP extensions[J].URL http: //www. Imperialviolet. org/binary/ecntest. pdf, 2008.

[12] Beverly IV R E. Statistical learning in network architecture[D]. Massachusetts Institute of Technology, 2008.

[13] 张维维,龚俭,丁伟等. NBOS:一个基于流技术的精细化网管系统[A].CERNET2012 年会[C]. 太原: 太原理工大学出版社, 2012.

ZHANG Ling-Feng, born in 1992, master candidate. Her main research interests include network measurement and network behavior.

DING Wei, born in 1962, Ph.D., professor, Ph.D. supervisor. Her main research interests include computer integrated manufacturing, general search engine, PKI certificate system, remote education under network environment and network behavior.

GONG Jian, born in 1957，Ph.D. , professor, Ph.D. supervisor. His main research interests include interconnection theory of open systems and its application, open distributed processing theory and its application, network management and network security.

MIAO Li-Hua, born in 1987, Ph.D. candidate. Her main research interests include network measurement and network behavior.

Background

Although IPv4 addresses have declared their exhaustion to the world, allocated addresses are often heavily under-utilized. Allocated and actual-utilized addresses are called active addresses. Understanding the utilization of IP addresses is important for network administrators and network security research etc. While the big challenge in managing IP address space is lacking reliable mechanisms to monitor the activity of addresses.

Much related work has been done till now. To our knowledge, ISI's Internet Census project [5] is the first to publish census of IPv4 address utilization, which periodically sends ICMP echo requests to each IPv4 address to track the activity except private and multicast addresses. They found approximately 4.3M /24 active address blocks through a month's scanning. While active probing is simple to practice, it has two limitations: significant probing network traffic overhead and inaccuracies due to some networks may either filter out echo requests or reply on behalf of target hosts. Passive traffic measurements avoid all these flaws with active probing. Nevertheless, it also has two challenges, one is the limited visibility of observation points, the other is the presence of spoofed traffic that can distort results by implying the faked addresses are active. Alberto Dainotti et al.[1] developed a heuristic methodology for removing spoofed traffic from both IBR traffic and operational network's unsampled flow data, and found the resulting filtered data effectively supporting census-like analyses of observed address space utilization. However, there is much difficulty in capturing IBR traffic from the Internet. What's more, unsampled flow dataset used in their approach requires a large quantity of resources, making it inappropriate for a larger scale of address space measurements. Two other studies have used TTL-based inference with active and passive measurements to detect spoofed packets, they tried to establish reference values of TTL for different traffic classes, inferring that packets with diverging values are spoofed. Complete IP packets are required for this TTL-based inference, which makes it impossible to extend the method on a large scale. Overall, the difficulties for passive measurements lie in dataset selection and scale limitation.

With the gradually maturing of flow technology, represented by NetFlow, all the mainstream backbone routers support providing flow records. This paper proposes a methodology of detecting active IP addresses based on sampled flow records captured from border routers. The method is based on passive measurements and could be deployed at the boundary of backbones, which is the greatest advantage.